# Understanding Admissions Entrance Assessment to Inform the Design of Effective and Equitable Human-AI Collaborative Assessment Workflows in Higher Education

ANONYMOUS AUTHOR(S)

Effective entrance assessments of incoming university students can help align educational resources with the universities', students', and societal goals. Universities are increasingly considering using artificial intelligence (AI) to design and administer entrance assessments to assess applicants' preparedness for interdisciplinary applied graduate programs. Although generative AI can support personalized and open-ended assessment workflows that offer high-fidelity evaluation of student skills at scale, its use in admissions decisions can conflict with institutions' legal, fairness, and equity norms. To explore AI's role in equitable entrance assessments, we examined current assessment practices and interviewed admissions personnel ($N = 4$), instructors ($N = 5$), and students ($N = 5$) in a large interdisciplinary graduate program at a public US research university. We identify the limitations of using fixed knowledge tests, prior degrees and essays in evaluating diverse applicants' technical and professional preparedness. We examine the stakeholders' considerations when using AI for design, administration, and evaluation of alternative workflows, including performance-based assessments for high-fidelity and efficient evaluation of student skills. By studying processes for fairness, integrity, and legal trade-offs in existing assessment workflows, we highlight standardization of question difficulty, human oversight and use of rubrics in automated grading of AI-supported assessments as key to balancing efficiency with equitable outcomes. Our findings inform opportunities for the responsible use of AI in admissions entrance assessments to further the goals of all stakeholders.

Additional Key Words and Phrases: Human-AI Collaborative Assessment; Performance-Based Admissions; Equity and Fairness in Admissions Decisions; Sociotechnical Governance; Transparent Evaluation Workflows

**ACM Reference Format:**
Anonymous Author(s). 2025. Understanding Admissions Entrance Assessment to Inform the Design of Effective and Equitable Human-AI Collaborative Assessment Workflows in Higher Education. 1, 1 (March 2025), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Effective entrance assessments of incoming students can enable universities to provide the right curriculum, identify resources for equitable educational outcomes, and prepare students for changing labor markets [50, 58]. Adequate assessments within admissions can support these outcomes by accurately estimating applicants' academic and professional competencies through their GPA, prior degrees, performance on knowledge tests, and professional experience [22]. However, assessments impose applicant burden and increasing educational and professional diversity of applicants for graduate programs makes it challenging for admissions and instructors to accurately assess applicants' technical and professional preparedness using these currently available measures [8, 31].

Universities are increasingly considering the potential of artificial intelligence (AI) in designing and administering assessments at scale, such as triaging student applications using ML-based predictions of student success or providing personalized course recommendations [13, 35, 37, 51, 64]. However, admissions assessment determines who gets access to education and informs curriculum design and resources needed to support equitable education. Such assessments demand accurate knowledge assessment, fairness, minimizing design and administration burden, and interoperability in addition to efficiency [66]. Naively using out-of-the-box AI models to make critical decisions in high-stakes contexts or training on past decisions could overlook the complex trade-offs and value-laden judgments of decision-makers

2025. Manuscript submitted to ACM

essential to make fair, efficient, and effective decisions [19, 36]. Robust evaluation processes are necessary to ensure that AI deployment aligns with organizational values of equity and fairness [17, 48].

In this work, we investigate the challenges in prerequisite skill assessments within graduate admissions and examine how generative AI, such as Large Language Models (LLMs), can responsibly enhance well-established assessment workflows to improve equity while reducing student burden [55]. We answer the following research questions (RQs) about the considerations for the design, administration, and evaluation of AI-supported prerequisite skill entrance assessment for domain-specific preparedness criteria for interdisciplinary applied graduate programs:

- **RQ1**: How can AI-supported skill assessment formats be incorporated into the existing processes to address the challenges of increasing educational and professional diversity of applicants for graduate programs?
- **RQ2**: How do decision-makers ensure equity, integrity, and legal compliance in the design, and implementation of prerequisite skill assessments for an interdisciplinary graduate program, and how do these considerations impact the use of AI-supported skill assessment formats?
- **RQ3**: What evaluation processes are necessary to ensure a fair and effective AI-assisted assessment of prerequisite skills for an interdisciplinary graduate program?

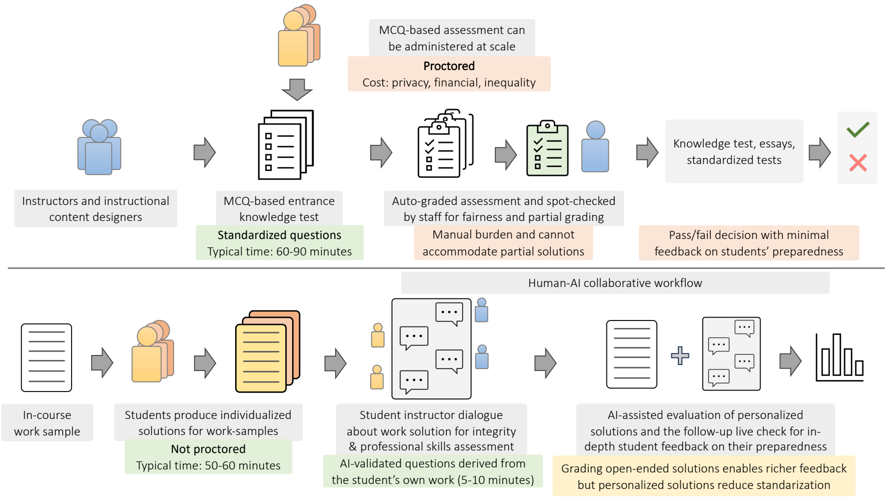

MCQ-based assessment can be administered at scale
Proctored
Cost: privacy, financial, inequality
Instructors and instructional content designers
MCQ-based entrance knowledge test
Standardized questions
Typical time: 60-90 minutes
Auto-graded assessment and spot-checked by staff for fairness and partial grading
Manual burden and cannot accommodate partial solutions
Knowledge test, essays, standardized tests
Pass/fail decision with minimal feedback on students' preparedness
Human-AI collaborative workflow
In-course work sample
Students produce individualized solutions for work-samples
Not proctored
Typical time: 50-60 minutes
Student instructor dialogue about work solution for integrity & professional skills assessment
AI-validated questions derived from the student's own work (5-10 minutes)
AI-assisted evaluation of personalized solutions and the follow-up live check for in-depth student feedback on their preparedness
Grading open-ended solutions enables richer feedback but personalized solutions reduce standarization

Fig. 1. Comparison of multiple choice question (MCQ)-based knowledge test and AI-assisted performance-based assessment workflows. The traditional MCQ knowledge test (top row) evaluates domain-specific knowledge at scale, using Autograding with human oversight to correct errors, and combined with essays and standardized tests determines applicant readiness for the program. The proposed AI-assisted performance-based workflow (bottom row) involves students completing open-ended assignments and discussing their personalized solutions with an instructor (for integrity). AI aids standardization, efficiency, and equity in personalized grading while maintaining human oversight. However, personalization brings fairness and standardization considerations. We study current assessments and the feasibility of AI-supported performance-based assessments.

To address our research questions, we reviewed the standard admissions processes at six U.S. masters programs in applied data science. We then conducted semi-structured interviews with admissions staff ($N = 4$), instructors ($N = 5$), and students ($N = 5$) from one of these six programs to explore how applicant preparedness is currently evaluated. In particular, we focused on fairness, integrity, legal requirements, and practical considerations for using AI-assisted evaluation of prerequisite skills (Figure 1). These programs offer rigorous, flexible curricula with online lectures, staff support, and community-building opportunities, making them appealing to diverse applicants seeking in-demand credentials. By studying the assessment workflow for a large, varied applicant pool, we aim to inform the design of AI-supported assessments that can enable programs to increase educational opportunities and identify resources for supporting diverse student cohorts [23, 78].

We find that standardized test scores, multiple choice question (MCQ) knowledge tests, and prior degrees offer limited insight into students' applied skills and learning attitudes while placing additional burdens on applicants. However, generative AI can reduce costs and improve evaluation fidelity by supporting performance-based assessments (i.e., problem-solving tasks) that measure both theoretical and practical knowledge [61]. From our review of existing assessment workflows, we identify three critical dimensions to ensure fairness and legal compliance in AI-assisted assessment workflows: (1) standardized question difficulty levels, (2) using rubric-based evaluations, and (3) conducting efficient human audits.

## 2 Related Work

We review prior work on graduate admissions assessments, AI in admissions, and stakeholder perspectives to contextualize our study and explore how AI can improve assessment workflows. We review the use of all types of AI in education, but our use of the term "AI" in the remainder of the paper corresponds to generative AI (e.g., LLMs).

### 2.1 Entrance assessments for graduate admissions

Universities strive for holistic admissions to assess applicants' backgrounds, experiences, and motivations [68], while shaping cohorts aligned with institutional goals [59]. However, standardized test scores, prior degrees, and GPA often fail to fully capture students' practical and professional readiness for graduate programs [22], disproportionately impacting certain socioeconomic groups [18, 49]. To address these gaps, admissions rely on additional criteria such as recommendation letters, essays, and prior experiences [49]. Yet, rising application volumes and increasing program diversity challenge admissions officers' ability to mitigate these limitations effectively.

Due to the limitations of prior degrees and grades in assessing diverse applicant backgrounds and evolving skill requirements [31], universities rely on standardized tests for general skills (e.g., English proficiency) and program-specific knowledge tests. However, the limited efficacy of standardized tests in predicting graduate success [6, 7, 42, 80] has led some programs to reconsider their use due to equity concerns, costs, and the burden on students [16, 29, 30, 52, 67]. Applied tests like IBM's programmer aptitude test assess programming skills but focus narrowly on lower levels of Bloom's taxonomy (e.g., understanding concepts rather than synthesizing solutions)[24, 39, 75], favoring test-taking expertise. Additionally, no scalable tests exist to assess essential soft skills like communication and teamwork, critical for graduate program success[21, 22, 27].

Alternative interactive assessment formats, such as work-sample solutions [72] and situational judgment tests (SJT) [53, 54], can assess practical and professional preparedness through open-ended problem solving [11, 43, 44, 46]. However, high personalization in such assessment formats makes scalable evaluations and standardization challenging, which are essential for graduate entrance assessments [40, 61, 73]. Some graduate admissions have already started experimenting with work-sample evaluations (also called **performance-based admissions** in education) [1–3]. While the efficacy of performance-based assessments for assessing higher-order skills is well-studied [70, 72], designing and administering them *efficiently* for entrance admissions assessment while *upholding principles of equity remains unclear.*

### 2.2 Understanding stakeholder values for algorithm design in educational contexts

Recent advancements in AI, showing high performance in lab settings, have led to its deployment in high-stakes areas like admissions, child welfare, and bail decisions [5, 28, 36]. However, field studies reveal limitations and unintended harms due to a lack of understanding of stakeholder values in algorithm design [60, 69]. While fairness in machine learning has been extensively studied [15, 20], concepts like statistical parity, equalized opportunity and tradeoffs between fairness and accuracy [47] [26] are often evaluated on researcher defined intuitions and may miss the complexities of real-world decision-making by domain experts [60, 69].

Research on ensuring fairness in automated decision systems (ADS) [28] suggests the need to go beyond model-specific factors (e.g., performance) [9, 38]. Factors such as decision-maker's interpretation of model outputs [25], processes established for logging decisions for auditing [17], and algorithmic impact assessments [48] show that stakeholder inclusive evaluation measures provide better evaluations of algorithmic fairness in socio-technical systems.

Due to the inherent incompleteness of data and the modeling of contextual social factors in statistical models, automated decision systems often struggle to make decisions that align with broader principles of fairness and equity. This can lead to oversimplified judgments in cases requiring deeper consideration, ultimately resulting in unfair outcomes [4, 41, 79]. Studies documenting stakeholders' decision processes for assessments [14, 34, 56, 71, 76] can offer valuable insights in evaluating students from diverse backgrounds, providing high-fidelity assessments of skills and ensuring fairness and equity but higher education lacks such studies [45].

## 3 Overview of Applied Graduate Programs

Graduate programs in applied fields like data science face challenges in assessing student readiness due to applicants' diverse educational backgrounds, ranging from prestigious universities to resource-limited community colleges [10, 57]. This diversity complicates admissions decisions and resource allocation, as admitting underprepared students strains resources, while rejecting prepared students limits access.

To address resource and selection limitations, many programs across the US are adopting flexible online instruction formats. These programs can offer affordable and adaptable pathways for a large number of individuals from diverse backgrounds to acquire high-demand professional skills while maintaining structured environments with instructor support, office hours, and student communities. The diversity and large-scale enrollment in these programs present a unique opportunity to study and refine the processes for assessing student preparedness. Our analysis of how these programs evaluate a diverse applicant pool reveals key considerations necessary to correctly implement AI-supported assessments in admission workflows to promote equitable educational opportunities, such as expanding class sizes, adapting learning formats, or developing new degree pathways.

### 3.1 Studying data science program admissions in the US

Given the diversity and high enrollment in data science programs, we reviewed admissions and curricula of leading online Master's in Data Science programs in the US, focusing on six prominent programs in universities (A-F). Following Fester [23]'s review of 62 programs, which highlighted the need for – 1) **Theoretical knowledge** ("foundational mathematical skills"), 2) **Technical knowledge** ("ability to select appropriate tools, implement and evaluate the results"), and 3) **Human Oriented Professional Skills** (HOPS), i.e., "teamwork, communication, ethical thinking skills, legal/privacy knowledge.", we selected programs emphasizing these aspects, and accessible to diverse academic backgrounds.

Table 1. Comparison of six online U.S. Masters in Applied Data Science programs. University D, selected for the stakeholder study, is highlighted in yellow. Filled bullets indicate applicable categories; empty bullets indicate otherwise.

| Dimension | Category | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|
| **Program Focus** | Data Management | ● | ○ | ○ | ● | ○ | ○ |
| | Ethics | ○ | ● | ○ | ● | ○ | ○ |
| | AI/ML | ● | ● | ● | ● | ○ | ● |
| | Business Analytics | ● | ○ | ○ | ○ | ● | ○ |
| | Visualization | ○ | ● | ○ | ● | ○ | ○ |
| | Cloud Computing | ○ | ○ | ● | ● | ○ | ○ |
| | Operations Research | ○ | ○ | ○ | ○ | ● | ○ |
| **Prerequisites** | Basic DS/Business | ● | ○ | ○ | ○ | ○ | ○ |
| | DS background | ○ | ● | ● | ○ | ○ | ○ |
| | Programming experience | ○ | ○ | ○ | ● | ○ | ○ |
| | Quantitative background | ○ | ○ | ○ | ○ | ● | ○ |
| | Related field degree | ○ | ○ | ○ | ○ | ○ | ● |
| **Format** | Online live classes | ○ | ● | ○ | ● | ○ | ○ |
| | Self-paced | ○ | ○ | ● | ○ | ○ | ○ |
| | Asynchronous | ○ | ○ | ○ | ● | ● | ● |
| | Synchronous Option | ○ | ○ | ○ | ○ | ○ | ○ |
| **Admission Process** | Rolling | ● | ● | ○ | ● | ○ | ○ |
| | Holistic review | ● | ● | ○ | ● | ○ | ● |
| | Technical focus | ○ | ○ | ● | ○ | ○ | ○ |
| | Performance-based test | ○ | ○ | ○ | ○ | ● | ○ |
| | Asynchronous knowledge test | ○ | ● | ● | ● | ● | ○ |
| **Prospective Students** | Practical experience | ● | ○ | ○ | ● | ● | ○ |
| | Holistic education | ○ | ● | ○ | ○ | ○ | ○ |
| | Technical expertise | ○ | ○ | ● | ● | ○ | ○ |
| | Budget-friendly | ○ | ○ | ● | ○ | ● | ○ |
| | Theory-Practice balance | ○ | ○ | ○ | ● | ○ | ● |

In these university programs A-F, we analyzed program descriptions, prerequisites, application processes, and degree structures, observing variations in admissions strategies, such as holistic reviews (A, B, D), technical skill focus (C), and performance-based admissions (E). Table 1 summarizes our findings, and Figure 3 in Appendix outlines a common admissions workflow with program-specific variations.

For an in-depth study, we selected University D as a convenience sample due to the proximity of its location to study members. The university's program places a balanced emphasis on ethics, AI/ML skills, liberal formats, and a combination of holistic reviews and knowledge tests covering all major aspects of the assessment dimensions across universities. This program's exploration of performance-based admissions provided a unique lens to study AI-supported assessments. Interviews with stakeholders focused on assessing theoretical, technical, and HOPS knowledge, identifying challenges, and opportunities for AI in robust assessments.

### 3.2 Specifications of the Applied Graduate Program in Data Science at University D

The Masters in Applied Data Science program at University D offers flexible online instruction, allowing students to attend from any location. Lectures are delivered live or pre-recorded via platforms like Coursera, enabling asynchronous learning within a structured timeline. The program supports students with office hours, Slack feedback, and community-building activities, including occasional in-person events. These online programs are increasingly popular for their lower costs, flexible schedules, and accredited, standardized curricula [33].

*3.2.1 Application and admissions process.* The program employs a holistic admissions process, requiring transcripts, three recommendation letters, and two essays—one on data science experience and the other on motivation and goals. International applicants must provide English proficiency scores if applicable. Figure 3 illustrates the admissions workflow. With three annual cycles, the program enrolls hundreds of students per cycle, including recent graduates and working professionals worldwide. Positioned for diverse applicants, the program trains all-around data scientists through technical and business-oriented courses.

*3.2.2 Python and statistics knowledge entrance test and waiver.* The program attracts students from diverse fields (e.g., engineering, medicine, finance) and thus requires all applicants to take a Python and statistics entrance knowledge test to assess programming and basic statistics proficiency (step (d), Figure 3). The 30-40 minute test is administered online via Coursera and **proctored** by a third-party service. The test includes Python MCQs and short coding tasks (see Figure 4a in Appendix for an illustration). The statistics test follows a similar format.

Knowledge test results are graded automatically using an **Autograder** tool [32] (step (e), Figure 3). Autograders run student code on instructor-provided test cases and flag a successful submission if the student's code output matches the reference answer in value and format(see Figure 4b). The tool does not *assess the process of problem-solving*.

Table 2. Assessment measures currently used in the program's entrance assessment for student preparedness.

| Measure | Evaluation measure |
|---|---|
| Prior degrees | Proxy measure to evaluate prerequisite programming and statistics knowledge, and basic undergraduate training. |
| MOOC certificates | Proxy measure to evaluate prerequisite knowledge of programming. |
| Knowledge test | Evaluate prerequisite Python and statistics knowledge. |
| Essay | Motivation for enrolling in the program and professional skills for conduct. |

Administering the knowledge test is resource-intensive for students and staff. To reduce the burden, the program grants waivers to STEM students as they are expected to have some programming and statistics skills from their degree. Students without any programming experience can complete an approved Python MOOC on Coursera and submit a completion certificate to waive the Python knowledge test. Table 2 summarizes admissions assessment aspects. In Appendix Section A, we provide an illustration of the knowledge test and more details on the admissions process.

## 4 Method for Understanding Entrance Admissions Assessment Process from Stakeholder's Perspectives

Here, we describe our method to understand admissions processes from stakeholders' perspectives to identify opportunities and considerations for safe, effective, and compliant use of AI to improve admissions assessments.

### 4.1 Study protocol

We conducted a qualitative study using semi-structured interviews with admissions personnel, instructors, and students in an online applied data science graduate program. Two researchers attended each interview; one led the discussion, obtained consent, and explained the process, while the other took notes, ensuring detailed insights were captured.

We developed tailored interview scripts for each stakeholder group to explore their backgrounds, roles, and perspectives on the admissions process and knowledge assessment. Admissions personnel discussed their tasks and challenges in designing and administering entrance assessments. Instructors shared insights on assessing students' prerequisite knowledge and addressing knowledge gaps. Students reflected on their backgrounds, motivations, admissions experiences, and prerequisite knowledge challenges. Around the time of the interviews, admissions considered shifting to performance-based admissions, and we also inquired about assessment considerations using performance-based assessments. Our institution's Institutional Review Board (IRB) approved the study as exempt. We provide more details of the interview protocol in the appendix (Section B).

### 4.2 Participant selection and recruitment

We interviewed all individuals who were part of admissions. For instructors, we ensured diversity in the courses they taught so that we could get an all-round experience about assessments. Thus, our instructor pool covered all the major pillars of the program – programming, mathematics, data ethics, machine learning, and a capstone project. We selected graduate students through snowball sampling but ensured a broad spectrum of backgrounds and experiences in our pool. We included individuals with strong technical expertise, those balancing full-time professional careers alongside part-time study, students grounded in statistics, and recent graduates entering the program directly from their undergraduate students. This diversity among students ensured a well-rounded representation of challenges encountered in transitioning to advanced data science concepts and insights into how students with varying levels of prior experience adapt to the program's demands. In total, we had four admissions, five instructors, and five students, totaling 14 stakeholder interviews. Table 3 illustrates the backgrounds of the various stakeholders.

### 4.3 Data analysis

After conducting the interviews, we reviewed recordings and notes for accuracy, resulting in 17 hours of recordings and corresponding transcripts. Using the Rev tool [1], we transcribed the interviews and applied open coding to the transcripts. Two authors collaboratively refined approximately 920 codes, representing aspects such as holistic review, missing knowledge in students, and assessment workflows. We aggregated these codes into themes using affinity diagramming on MIRO [2], iteratively refining them through team discussions while mitigating biases through reflective journaling. These themes encompassed assessment workflows, fairness, knowledge test formats, and the role of assessments post-admissions, forming four categories for our findings.

### 4.4 Limitations

This study focuses on interdisciplinary applied data science master's programs, which may differ from more theoretical or specialized programs, PhD programs emphasizing research, or undergraduate programs with less diverse educational backgrounds. Additionally, the North American context of our study may limit generalizability to regions with different labor markets or undergraduate curricula.

Our student participants were based in the US and Canada, limiting insights on time zone differences, financial considerations, and language barriers. To mitigate this limitation, we asked our student participants to recount any relevant international student-related experiences of their friends.

[1] https://www.rev.com/

[2] https://miro.com/

Table 3. Participant backgrounds across different roles: Admission personnel, Instructors, and Students.

| Participant Code | Role | Background/Responsibilities |
|---|---|---|
| A01 | Admissions - Content Specialist | Assesses instructors, develops course content, conducts beta tests, fixes errors, and reviews assignment codes |
| A02 | Admissions - Lead Staff | Collaborates with admissions, designs apps, reviews content, manages communications, and handles regulations |
| A03 | Admissions - Director | Oversee admissions processes and ensure alignment with institutional goals |
| A04 | Admissions - Staff | Supports admissions processes, liaises with applicants, manages documentation |
| I01 | Instructor | Teaches data visualization techniques using Matplotlib |
| I02 | Instructor | Teaches project-based capstone course |
| I03 | Instructor | Teaches applied statistics and mathematics |
| I04 | Instructor | Teaches introductory data science focused on problem framing, communication, and ethics. |
| I05 | Instructor | Teaches advanced data manipulation techniques using Python |
| S01 | Full-time student | Background in chemistry with 5 years of industry experience and no coding background |
| S02 | Part-time student | Background in psychology with a minor in programming and no industry experience |
| S03 | Part-time student | Background in aerospace engineering with 30 years of non-coding analysis experience |
| S04 | Part-time student turned full-time | Background in industrial engineering and 20 years of experience in business technology consulting |
| S05 | Part-time student turned full-time | Background in computer science engineering with 22 years of IT experience in data |

We minimized bias through open coding, regular team discussions, and verification with select participants. While our in-depth study of one program provided detailed insights into AI-supported assessments, differences in curricula and prerequisites across programs (Table 1) suggest our findings should be contextualized for other programs.

## 5 Findings

We now describe the three major themes that emerged from our data: 1) Considerations for technical and professional prerequisite skill assessment for graduate programs, 2) Ensuring equity, integrity and efficiency in prerequisite skill assessments and considerations for AI-assisted performance-based assessments, and 3) Evaluation processes for fair, and efficient AI-assisted performance-based assessments. Some of our findings are in the context of the knowledge test, proctoring, and the Autograder tool. As a reminder, we have described these aspects of admissions in Section 3.2.

### 5.1 Considerations for technical and professional prerequisite skill assessment for graduate program

In this section, we describe the current workflow for assessing students, assessment gaps, and AI considerations to support efficient and adequate assessments.

*5.1.1 Prior degrees and skill certifications may not accurately estimate competencies required to succeed in an interdisciplinary applied degree.* As described in Section 3.2, admissions waived the Python and statistics entrance knowledge test if the applicant had an engineering undergraduate degree or if they took the approved Python MOOC. Engineering waivers were based on the fact that engineering degrees teach programming-related competencies. However, curriculums across undergraduate engineering degrees and institutions varied significantly, with some providing extensive programming instruction and others providing no programming courses.

To get around this challenge of variation across degrees, admissions allowed students to take a Python MOOC and use it to waive entrance tests. However, MOOCs *"cater to a broad audience"* I01 ranging from casual learners to those intending to learn new skills for their jobs. Due to this diversity, MOOCs offer a self-paced learning style with limited accountability in learning as they progress through MOOCs. Self-paced learning and limited accountability may work for students who can understand concepts, but students who struggle with the MOOC may need more hands-on guidance. As shown in the quote below, some students also found the MOOC's pace challenging and thus inefficient in helping them learn despite getting the completion certificate.

> *"The MOOC moved too quickly, and there was no reinforcement of tools and techniques taught in the course, but we were expected to apply the concepts."* – S04

Students would also progress quickly through MOOCs without a "*fair assessment of their learning*" (S03), and their certifications may not reflect the competency to apply the programming knowledge to the graduate degree. Consequently, instructors in the course reported that some students with engineering backgrounds or those with MOOC experience had insufficient programming experience and struggled in the courses.

**Consideration for design:** Varied and non-standardized curriculums of prior degrees or MOOCs limit the generalization of technical abilities from such experience, and a domain-specific knowledge test is necessary to evaluate diverse backgrounds in a standardized manner.

*5.1.2 Synchronous timed MCQ entrance tests can assess theoretical knowledge but may not be suitable for assessing technical expertise to apply the knowledge.* Manually designed Python and statistics knowledge tests contained short MCQs testing the topic fundamentals, but they could not adequately assess students' ability to write end-to-end Python code for solving problems or *"debug broken code"* I01. Instructors reported students struggling to complete assignments due to a lack of debugging skills or the ability to think about algorithms as building blocks to solve computational problems, critical thinking and reasoning abilities [74]. Such abilities form the highest levels in Bloom's taxonomy ("synthesis" and "evaluation"). For example, students learning the statistical method assumptions for correct application instead of following rule-based recipes I01 or students learning the history of social interaction theories so that they can reason about the correct application of metrics given the problem context S01.

To help students who lacked such skills in the program, instructors employed various strategies such as *"suggesting them to work on developing a game to increase Python knowledge"* I02 or *"working closely with them to scaffold steps in a complex problem that students were unable to identify"* I03. They even suggested using generative AI tools to *"scaffold complex parts of a problem"* I02 so that students can develop applied problem-solving expertise. Students indicated benefits from feedback on their knowledge or guidance on adequate difficulty problems as exemplified by S03 below.

> "...sometimes I still, a lot of times I still struggle actually with the concepts. And I think it's because I am, I've been learn, I've learned to just like use the ready package and I don't fully understand like, the guts of

> it. Like, so where possible it would've been nice to have some more challenging and, um, like assessments that, or assignments that make you really think about the concepts more. " – S03

**Consideration for design:** Entrance knowledge tests designed to assess proficiency must go beyond knowledge of facts and basic procedural skills. Assessments that evaluate theoretical and technical skills similar to what is expected of them in the degree can provide better preparedness assessments of diverse student backgrounds.

*5.1.3 Standardized tests like TOEFL or essays may not provide an adequate assessment of human-oriented professional skills and learning attitudes necessary for graduate program success.* Online graduate programs for professionals and recent graduates expect professional etiquette, including collaboration on group projects, respectful communication on IM platforms, digital etiquette, and valuing diversity. English proficiency is necessary to succeed in problem-based courses, where students collaborate on data science projects from conception to presenting findings [65, 78]. While most students excel in coding, some struggle with communication skills, such as selecting appropriate presentation tools or effectively conveying ideas. For instance, I02 noted *"suboptimal choice of figures"* and *"irrelevant presentation details"*, while I03 emphasized the need to understand domain-specific explanations for teamwork.

Admissions rely on application essays and TOEFL scores to evaluate applicants' professional skills and motivation. However, essays often lack integrity, *"as applicants may seek external help or use AI tools"* A02. TOEFL assesses general English proficiency but fails to capture domain-specific communication skills, while imposing financial and time burdens, particularly on students from developing countries.

Students in the program also had varying learning attitudes towards different subjects in the course due to varying industry experience and undergraduate degree domains. For example, students with industry experience *"liked being treated like adults"* I01 who show more independent learning skills but also had negative attitudes towards more ethical or theoretical topics in data science. Instructors also found some experienced students to be *"overconfident in their abilities"* I01 related to natural language processing and quantitative research and asked to waive important courses in the program. These attitudes impacted students' approach toward learning and student-instructor dynamics. Such attitudes are difficult to capture in essay-based answers.

**Consideration for design:** Standardized tests burden students and offer limited insight into program-specific skills, and essays fail to reflect student behaviors. Interactive assessments like Situational Judgment Tests (SJT) can more accurately assess professional and communication skills by evaluating responses to program-relevant scenarios in real time.

**Performance-based admissions workflow for assessment of theoretical, technical, and professional skills.** We discussed that while administering and evaluating MCQ-based knowledge tests and essays is convenient, they provide a limited evaluation of students' theoretical, technical, and human-oriented professional skills or have integrity considerations. Work sample evaluations (also called **performance-based assessments**) are an alternative evaluation process that requires candidates to solve an *open-ended* in-course or on-job task to assess their *readiness* for the program. Such work-samples require students to understand the problem, identify solutions and implement them, testing full range of their skills. Situational Judgement Tests (SJT) [53, 54] can evaluate an applicant's professional and communicative ability through their live responses to educational scenarios.

However, manually administering and evaluating students' work sample solutions is challenging because each student may have unique solutions that cannot be assessed automatically like MCQ-based knowledge tests. Admissions staff also lack the training to conduct knowledge and situational assessments of applicants. AI can ease this burden by generating and evaluating questions personalized to students' work-sample solutions that test theoretical, practical, and professional skills.

However, using AI in administering and evaluating student's coursework samples presents additional challenges of fairness and legal considerations as they play a role in deciding who gets admission to the program. We now discuss these considerations through our interviews with the stakeholders.

### 5.2 Considerations for equity, integrity, and efficiency in prerequisite skill assessment workflow

In this section, we discuss the additional considerations for maintaining fairness, integrity, and legal compliance and how it can inform the design of responsible AI-supported assessments.

*5.2.1 Live checks on knowledge of work sample solutions can balance integrity and time commitment burden incurred from synchronous knowledge tests but question difficulty should be managed carefully.* Preparing for a scheduled proctored entrance test burdened applicants' schedules. Many applicants work full-time or have family responsibilities. They face difficulties balancing their jobs and family responsibilities. Additionally, as part of proctoring, applicants had to *"show proof of the integrity of their testing location by showing the camera around or giving permission to record the screen"* S03. These demands placed *"emotional stress on participants"* S01 or *"limited their freedom of choice of testing location"* S04. The need for preparation can also change an applicant's plans for the degree, as preparation places a time burden, like in the following quote recounted by a student participant. Proctoring the entrance knowledge test also costs resources for the program.

> *"So once I found that I needed to take the proficiency exam, I had to start studying. And I think I put off the exams until the end of the summer because I needed that long to prepare while working full time"* – S04

To further preserve the integrity of knowledge tests, the staff maintained multiple versions of the entrance test and periodically checked online to ensure that prior students had not posted questions that could compromise integrity.

Despite the challenges with timed tests, instructors carefully balance question difficulty in tests to build students' confidence as they take the test. Generative AI-supported live knowledge checks could end up with difficult questions that hurt student confidence if not explicitly controlled.

On the student side, all four students in our study favored an interactive live check on their knowledge by a staff or instructor where they could explain their solutions. One student expressed concern about nervousness when appearing for an interactive assessment instead of a take-home assessment. All students also indicated "more trust" in the reliability of an interactive check than in a proctored exam.

**Consideration for design :** Allowing students to do take-home assignments followed by a short live assessment of the student by the staff regarding the student's solution provides integrity assurance while minimizing proctoring or long-timed exams on students.

*5.2.2 Ability to accommodate and evaluate partially correct solutions and diverse programming styles ensures equity and helps identify students with potential in a diverse applicant pool.* In our study of Python and statistics MCQ tests, we found that the Autograder's rigid rule-based grading often penalized students for correct answers containing *"syntax* *errors"* or *"context-dependent correctness"* A03. This occurred because the Autograder strictly matched outputs to instructor-provided references (see Section 3.2), incorrectly grading submissions with formatting inconsistencies *"despite correct numerical answers"* S03. While the admissions team manually reviewed flagged cases to ensure equity, this process added a significant administrative burden.

> *"Minor syntax errors or formatting issues can cause the Autograder to fail even when the student understands the concept."* – A04

These challenges were more evident for non-CS majors, who lacked familiarity with coding conventions and auto-grading nuances, creating equity concerns compared to CS-trained students. While students could demonstrate programming skills like writing loops or functions, they still risked failing due incorrect final outputs. Student solution grading using generative AI could improve equity by *"analyzing students' process of problem-solving"* I04 in addition to their final answer.

**Consideration for design:** Generative AI accommodates diverse student responses by evaluating open-ended and partial answers. However, it may introduce biases influenced by superficial factors such as coding style, interview accents, or response delivery. Combining AI evaluations with effective human oversight can enhance both robustness and equity in assessments.

### 5.3 Processes to ensure fair, effective and efficient AI-assisted prerequisite skill assessments

In the previous sections, we discussed opportunities and considerations for efficient and effective assessments and trade-offs for fairness, integrity, and legal compliance when designing them. In this section, we discuss the processes necessary for fair, effective, and efficient AI-assisted evaluation of prerequisite skill assessments.

*5.3.1 Design and administration of live knowledge check and professional skills assessment need to balance personalization with standardization for fairness and legal compliance.* When students submit solutions for coursework samples, the solutions can vary across students. Depending on the solution, even the follow-up live check on the student's knowledge will show variations. While a live knowledge check mitigates integrity challenges and takes away the proctoring burden on students, but administering and evaluating such tests causes fairness concerns for admissions. They favored an interview-based test protocol with a *"highly prescriptive approach with a little bit of flexibility"* A03 where the interviewer has limited discretion when asking questions, which can aid in reducing bias in evaluations.

The consideration for standardization becomes even more challenging as admissions consider assessing professional competencies through interview-based assessments. For example, A02 indicated the difficulty of consistently evaluating applicants' professionalism across individuals through an interview-based assessment. Designing questions that elicit students' professional skills while maintaining consistency across applicants requires rubrics and guidelines.

Staff's current practices to maintain knowledge test design reflect a viable approach for AI-based assessment. The program ensures standardization of the current MCQ-based entrance Python knowledge test through careful assessment design and review by multiple instructors. The questions in the assessment are carefully curated to follow the increasing difficulty of maintaining student confidence (A02). Questions in live-check of students' solutions and professional assessments need to follow similar standards. Human interviewers need to be provided guidelines, or AI-based QA assessments need to be explicitly constrained in the range of questions that they can generate.

*5.3.2 Assessment grading needs to be guided by well-defined and unambiguous rubrics for consistent and fair evaluations.* The open-ended nature of student's technical solutions to performance-based assessments and the subjectivity in professional skills assessment implies that their grading (human or AI-based) needs processes to minimize discrepancy across evaluators. A02 suggested that *"there could be bias and concerns about inter-rater reliability if there is more than one evaluator"* for the performance-based asssessment results. This bias becomes more significant when evaluating professional skills due to subjectivity. Due to this subjectivity, admissions also stressed that professional skills assessment should *"have minimal impact on admission decisions, rather only be used for advising students initially"* A01. A02 indicated the difficulty of defining professionalism for evaluators.

> *"What professionalism is to me and to you and to you could be 3 different things. So that's where we started to turn away from the idea that this interview could accomplish that objective."* A02

While manually graded assignments may show disagreements across students due to varying interpretations of guidelines and subjectivity, AI-based grading *"can be biased by its training data or evaluation instructions"* I02. Suggestions from the legal team indicate *using rubrics for human and AI evaluations* A01 as a potential way to ensure consistency and fairness in evaluations. For example, a rubric "Assess whether the code is clean and well-structured" leaves ample room for interpretation than one that defines the interpretations - "Evaluate based on the following criteria: (1) proper use of functions to avoid repetitive code, (2) meaningful variable and function names." In addition to rubrics, legal requirements dictate that human reviewers should be properly trained. They also recommend periodic audits of reviewer ratings for systematic bias, low inter-reviewer agreement, and oversight of the student-facing AI outputs. During the evaluation, it is necessary to have a rationale for the questions the student was asked, and the overall assessment should show some diversity in the dimensions of questions (e.g., difficulty and concepts covered).

**Consideration for design:** Out-of-the-box use of generative AI to administer and evaluate coursework sample assessments can lead to hyper-personalization and disparate outcomes. Explicit measures to ensure consistency in the difficulty of questions generated and evaluation protocols are necessary for standardization.

## 6 Discussion

We now summarize our findings from the research questions and discuss their implications for 1) (**RQ1**) How can AI-supported prerequisite skill assessments address the challenges of increasing diversity in graduate program applicants?, 2) (**RQ2**) How do decision-makers ensure equity, integrity, and compliance in prerequisite skill assessments, and how do these affect the use of AI-supported formats?, 3) (**RQ3**) What evaluation processes ensure fair and effective AI-assisted prerequisite skill assessments for an interdisciplinary graduate program?

### 6.1 RQ1: Incorporating AI-supported prerequisite assessments for diverse graduate applicant evaluations

Our findings show that while traditional assessment measures, such as prior degrees and standardized tests, may indicate some preparedness for traditional degrees, such measures are largely insufficient for interdisciplinary applied graduate programs that are coming up to address the rapidly changing skill demands [10, 57] (Section 5.1.1). Our study of the program's consideration for performance-based assessments [61] suggests its potential for assessing students' theoretical and technical proficiency through open-ended coursework sample problems and short live knowledge checks (Section 5.1.2, 5.1.3). While performance-based assessments reduce the proctoring and synchronous test-taking burden on students and enable in-depth knowledge assessments, solving work samples offline raises integrity concerns.

Follow-up AI-assisted live checks on work-sample solutions and interactive assessment of professional skills can *efficiently* evaluate students' theoretical, technical, and professional skills that applied graduate programs demand (Section 5.2.1).

### 6.2 RQ2: Ensuring equity, integrity, and compliance in AI-assisted prerequisite skill assessment workflows

We found that equity and integrity concerns lead to significant resources spent on proctoring, curating test versions, and spot-checking MCQ evaluations (Section 5.2.1). These challenges increase with the use of generative AI for interactive work-sample and professional skills assessments. Legal requirements for standardized AI-generated questions and balanced difficulty reveal tensions with personalization, similar to issues in recommendation systems [63, 77] (Section 5.3.1). The need to maintain standardization through periodic testing of AI on representative test cases, using simulated student profiles and historical data, echoes prior work [28] on contextual fairness against more domain-general fairness metrics [62]. On the other hand, we also identify the potential for carefully design AI grading to improve equity by analyzing even partially correct student answers (Section 5.2.2).

### 6.3 RQ3: Processes to ensure Fair and Effective AI-assisted Prerequisite Skill Assessments

We identified the tension between personalizing performance-based assessments and AI-assisted live knowledge checks and the need to ensure fairness by providing questions of comparable difficulty to all students (Section 5.3.1). Personalized assessments also challenge standardizing grading, unlike MCQs where all students are assessed on the same questions (Section 5.3.2). To maintain consistency, practices like grade audits and legal requirements for reviewing inter-annotator agreements highlight the need for logging decision points for review and accountability, as noted in automated decision systems (ADS) studies [17]. This data aids algorithmic impact assessments and ensures alignment with institutional values [48]. Our finding of using well-calibrated rubrics and a Human-AI workflow to improve decision quality in educational assessments reinforces prior such calls in more generalized settings [12, 69] (Section 5.3.2).

## 7 Conclusion and Future Work

In this work, we studied entrance admissions assessment workflow to understand the design of interactive performance-based assessment formats and the responsible use of AI for fair, efficient, and effective assessments. Through interviews with admissions, instructors, and students in a large online master's program in Data Science, we determined the processes for assessing student preparedness for the program, considerations for fairness, equity, and legal compliance designing and administering AI-supported assessments, and insights for the responsible use of AI in evaluating personalized performance-based assessment workflows.

Our work opens up future directions for the responsible use of AI in multi-stakeholder, high-stakes decision contexts. Research in HCI is necessary to understand how decision-makers can use AI-based evaluations of performance-based assessments and quantitative analysis of how personalization in assessments impacts standardization in administering and evaluating assessments. Work that explores how personalized AI-supported performance-based assessments align with institutional objectives of fairness and equity will inform the responsible deployment of these technologies in practice.

## References

[1] 2011. Linked Learning Performance-Based Assessment. https://edpolicy.stanford.edu/sites/default/files/events/materials/2011-06-linked-learning-performance-based-assessment.pdf Accessed: 2025-01-08.

[2] 2025. New Report Explores How States Approach Performance-Based Assessments. https://www.wested.org/blog/new-report-explores-how-states-approach-performance-based-assessments/ Accessed: 2025-01-08.

[3] 2025. Who Benefits from Performance-Based Admissions? https://www.onlineeducation.com/features/who-benefits-from-performance-based-admissions Accessed: 2025-01-08.

[4] Fahrobby Adnan, Maulida Dwi Agustiningsih, and Lutfi Ariefianto. 2022. The Analysis of Readiness and Acceptance of Learning Management System (LMS) Usage in Universities of East Java. In *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. 198–203. https://doi.org/10.23919/EECSI56542.2022.9946579

[5] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300760

[6] Richard C. Atkinson and Saul Geiser. 2009. Reflections on a Century of College Admissions Tests. *Educational Researcher* 38, 9 (2009), 665–676. https://doi.org/10.3102/0013189X09351981

[7] William G Bowen, Michael McPherson, and Matthew M Chingos. 2009. Crossing the finish line: Completing college at America's public universities. (2009).

[8] Nicholas A. Bowman and Michael N. Bastedo. 2018. What Role May Admissions Office Diversity and Practices Play in Equitable Decisions? *Research in Higher Education* 59 (2018), 430–447. https://doi.org/10.1007/s11162-017-9468-9

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[10] Longbing Cao. 2017. Data Science: A Comprehensive Overview. *ACM Comput. Surv.* 50, 3, Article 43 (jun 2017), 42 pages. https://doi.org/10.1145/3076253

[11] Robert E Carlson, Paul W Thayer, Eugene C Mayfield, and Donald A Peterson. 1971. Improvements in the selection interview. *Personnel Journal* (1971).

[12] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating Classifier Error Discovery through Interactive Semantic Data Exploration. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 269–280. https://doi.org/10.1145/3172944.3172950

[13] Youjie Chen, Annie Fu, Jennifer Jia-Ling Lee, Ian Wilkie Tomasik, and René F. Kizilcec. 2022. Pathways: Exploring Academic Interests with Historical Course Enrollment Records. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 222–233. https://doi.org/10.1145/3491140.3528270

[14] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[15] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (apr 2020), 82–89. https://doi.org/10.1145/3376898

[16] Melissa Clinedinst and Anna Maria Koranteng. 2017. State of college admission. (2017).

[17] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 598–609. https://doi.org/10.1145/3442188.3445921

[18] Sarah S Conrad, Amy N Addams, and Geoffrey H Young. 2016. Holistic review in medical school admissions and selection: a strategic, mission-driven response to shifting societal needs. *Academic Medicine* 91, 11 (2016), 1472–1474.

[19] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[20] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[21] DM Dunleavy and KM Whittaker. 2011. The evolving medical school admissions interview. *AAMC Analysis in Brief* 11, 7 (2011), 1–2.

[22] Mary K. Enright and Drew Gitomer. 1989. TOWARD A DESCRIPTION OF SUCCESSFUL GRADUATE STUDENTS. *ETS Research Report Series* 1989, 1 (1989), i–37. https://doi.org/10.1002/j.2330-8516.1989.tb00335.x

[23] Sarah Albers Fester. 2024. *An Examination of Master's Level Data Science Programs Across the United States.* Master's thesis. University of California, San Diego.

[24] Robert M Gagne. 1970. Learning Theory, Educational Media, and Individualized Instruction. (1970).

[25] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563

[26] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[27] James J Heckman and Yona Rubinstein. 2001. The importance of noncognitive skills: Lessons from the GED testing program. *American economic review* 91, 2 (2001), 145–149.

[28] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I* (Chicago, IL, USA). Springer-Verlag, Berlin, Heidelberg, 157–171. https://doi.org/10.1007/978-3-030-23204-7_14

[29] Edmond A. Hooker, Peter Mallow, Cordell Downes, and Sonal Baidwan. 2022. Abandoning the GRE and GMAT Improved Diversity in a Graduate Program for Health Administration. *Journal of Health Administration Education* 38, 4 (2022), 895–912. https://www.ingentaconnect.com/content/aupha/jhae/2022/00000038/00000004/art00004

[30] Eric Hoover. 2018. U. of Chicago will no longer require ACT or SAT tests. *Chronicle of Higher Education* (2018).

[31] Don Hossler and David Kalsbeek. 2009. Admissions testing & institutional admissions processes: The search for transparency and fairness. *College and University* 84, 4 (2009), 2.

[32] Silas Hsu, Tiffany Wenting Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. 2021. Attitudes Surrounding an Imperfect AI Autograder. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 681, 15 pages. https://doi.org/10.1145/3411764.3445424

[33] David A. Joyner. 2022. Meet Me in the Middle: Retention in a "MOOC-Based" Degree Program. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 82–92. https://doi.org/10.1145/3491140.3528283

[34] Rais Kamis, Jessica Pan, and Kelvin KC Seah. 2023. Do college admissions criteria matter? Evidence from discretionary vs. grade-based admission policies. *Economics of Education Review* 92 (2023), 102347.

[35] Evangelos Katsamakas, Oleg V Pavlov, and Ryan Saklad. 2024. Artificial intelligence and the transformation of higher education institutions. *arXiv preprint arXiv:2402.08143* (2024).

[36] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439

[37] Graham Keir, Willie Hu, Christopher G Filippi, Lisa Ellenbogen, and Rona Woldenberg. 2023. Using artificial intelligence in medical school admissions screening to decrease inter-and intra-observer variability. *JAMIA open* 6, 1 (2023), ooad011.

[38] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[39] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.

[40] Suzanne Lane and Clement A Stone. 2006. Performance assessment. *Educational measurement* 4 (2006), 387–431.

[41] Melissa Laufer, Anne Leiser, Bronwen Deacon, Paola Perrin de Brichambaut, Benedikt Fecher, Christian Kobsda, and Friedrich Hesse. 2021. Digital higher education: a divider or bridge builder? Leadership perspectives on edtech in a COVID-19 reality. *International Journal of Educational Technology in Higher Education* 18 (2021), 1–17.

[42] Nicholas Lemann. 1999. The big test: the secret history of the American meritocracy. (1999).

[43] Filip Lievens and Fiona Patterson. 2011. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology* 96, 5 (2011), 927.

[44] John C Lin, Christopher Shin, and Paul B Greenberg. 2024. The impact of the medical school admissions interview: a systematic review. *Canadian Medical Education Journal* 15, 1 (2024), 68–74.

[45] Kelly McConvey, Shion Guha, and Anastasia Kuzminykh. 2023. A Human-Centered Review of Algorithms in Decision-Making in Higher Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 223, 15 pages. https://doi.org/10.1145/3544548.3580658

[46] Michael A McDaniel, Deborah L Whetzel, Frank L Schmidt, and Steven D Maurer. 1994. The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of applied psychology* 79, 4 (1994), 599.

[47] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*. PMLR, 107–118.

[48] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 735–746. https://doi.org/10.1145/3442188.3445935

[49] Kristen M. Glasener Michael N. Bastedo, Nicholas A. Bowman and Jandi L. Kelly. 2018. What are We Talking About When We Talk About Holistic Review? Selective College Admissions and its Effects on Low-SES Students. *The Journal of Higher Education* 89, 5 (2018), 782–805. https://doi.org/10.1080/00221546.2018.1442633

[50] Martin Mulder, Judith Gulikers, Harm Biemans, and Renate Wesselink. 2009. The new competence concept in higher education: error or enrichment? *Journal of European industrial training* 33, 8/9 (2009), 755–770.

[51] Fawad Naseer, Muhammad Usama Khalid, Nafees Ayub, Akhtar Rasool, Tehseen Abbas, and Muhammad Waleed Afzal. 2024. Automated Assessment and Feedback in Higher Education Using Generative AI. In *Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation*. IGI Global, 433–461.

[52] MIT News Office. 2022. *MIT reinstates SAT/ACT requirement for future admissions cycles.* https://news.mit.edu/2022/stuart-schmill-sat-act-requirement-0328 Accessed: 2024-09-04.

[53] Fiona Patterson, Victoria Ashworth, Lara Zibarras, Philippa Coan, Maire Kerrin, and Paul O'Neill. 2012. Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical education* 46, 9 (2012), 850–868.

[54] Fiona Patterson and Rachel Driver. 2018. Situational judgement tests (SJTs). *Selection and recruitment in the healthcare professions: Research, theory and practice* (2018), 79–112.

[55] Julie R. Posselt. 2014. Toward Inclusive Excellence in Graduate Education: Constructing Merit and Diversity in PhD Admissions. *American Journal of Education* 120, 4 (2014), 481–514. https://doi.org/10.1086/676910 arXiv:https://doi.org/10.1086/676910

[56] Rebecca M. Quintana, Chris Quintana, Jacob Fortman, and Elisabeth R. Gerber. 2020. ViewPoint: Student Experiences with Technology Supporting Role-Based Educational Simulations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3383086

[57] Daphne R Raban and Avishag Gordon. 2020. The evolution of data science and big data research: A bibliometric analysis. *Scientometrics* 122, 3 (2020), 1563–1581.

[58] Marco Rieckmann. 2012. Future-oriented higher education: Which key competencies should be fostered through university teaching and learning? *Futures* 44, 2 (2012), 127–135. https://doi.org/10.1016/j.futures.2011.09.005 Special Issue: University Learning.

[59] Gretchen W Rigol. 2003. Admissions Decision-Making Models: How US Institutions of Higher Education Select Undergraduate Students. *College Entrance Examination Board* (2003).

[60] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 589, 14 pages. https://doi.org/10.1145/3411764.3445748

[61] Paul R Sackett. 2013. Performance assessment in education and professional certification: Lessons for personnel selection? In *Beyond multiple choice*. Psychology Press, 113–129.

[62] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[63] Jessie J. Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. Recommend Me? Designing Fairness Metrics with Providers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2389–2399. https://doi.org/10.1145/3630106.3659044

[64] Adele Smolansky, Andrew Cram, Corina Raduescu, Sandris Zeivots, Elaine Huber, and Rene F. Kizilcec. 2023. Educator and Student Perspectives on the Impact of Generative AI on Assessments in Higher Education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 378–382. https://doi.org/10.1145/3573051.3596191

[65] Il-Yeol Song and Yongjun Zhu. 2017. Big data and data science: Opportunities and challenges of iSchools. *Journal of Data and Information Science* 2, 3 (2017), 1–18.

[66] Steven E. Stemler. 2012. What Should University Admissions Tests Predict? *Educational Psychologist* 47, 1 (2012), 5–17. https://doi.org/10.1080/00461520.2011.611444

[67] Lisa M. Sullivan, Amanda A. Velez, Nikki Longe, Ann Marie Larese, and Sandro Galea. 2022. Removing the Graduate Record Examination as an Admissions Requirement Does Not Impact Student Success. *Public Health Reviews* 43 (2022). https://doi.org/10.3389/phrs.2022.1605023

[68] Poorna Talkad Sukumar, Ronald Metoyer, and Shuai He. 2018. Making a Pecan Pie: Understanding and Supporting The Holistic Review Process in Admissions. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 169 (nov 2018), 22 pages. https://doi.org/10.1145/3274438

[69] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[70] Kenneth E Vogler. 2002. The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education* 123, 1 (2002), 39–56.

[71] Xiaomei Wang, Ann M. Bisantz, Matthew L. Bolton, Lora Cavuoto, and Varun Chandola. 2020. Cognitive Work Analysis and Visualization Design for the Graduate Admission Decision Making Process. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 815–819. https://doi.org/10.1177/1071181320641189

[72] Grant Wiggins. 2011. A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan* 92, 7 (2011), 81–93.

[73] Anne Wildermuth. 2022. Predictive Ability of Multiple Mini-Interviews in Admissions on Programmatic Academic Achievement: A Systematic Review. *Journal of Allied Health* 51, 2 (2022), 154–159. https://www.ingentaconnect.com/content/asahp/jah/2022/00000051/00000002/art00012

[74] Jeannette M. Wing. 2006. Computational thinking. *Commun. ACM* 49, 3 (mar 2006), 33–35. https://doi.org/10.1145/1118178.1118215

[75] Jack M. Wolfe. 1971. Perspectives on testing for programming aptitude. In *Proceedings of the 1971 26th Annual Conference (ACM '71)*. Association for Computing Machinery, New York, NY, USA, 268–277. https://doi.org/10.1145/800184.810494

[76] Lingrui Xu, Zachary A. Pardos, and Anirudh Pai. 2023. Convincing the Expert: Reducing Algorithm Aversion in Administrative Higher Education Decision-making. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing

Machinery, New York, NY, USA, 215–225. https://doi.org/10.1145/3573051.3593378

[77] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

[78] Yin Zhang, Dan Wu, Loni Hagen, Il-Yeol Song, Javed Mostafa, Sam Oh, Theresa Anderson, Chirag Shah, Bradley Wade Bishop, Frank Hopfgartner, et al. 2023. Data science curriculum in the iField. *Journal of the Association for Information Science and Technology* 74, 6 (2023), 641–662.

[79] Chang Zhu and Nadine Engels. 2014. Organizational culture and instructional innovations in higher education: Perceptions and reactions of teachers and students. *Educational Management Administration & Leadership* 42, 1 (2014), 136–158. https://doi.org/10.1177/1741143213499253

[80] Rebecca Zwick. 2019. Assessment in American Higher Education: The Role of Admissions Tests. *The ANNALS of the American Academy of Political and Social Science* 683, 1 (2019), 130–148. https://doi.org/10.1177/0002716219843469

## A Admissions process

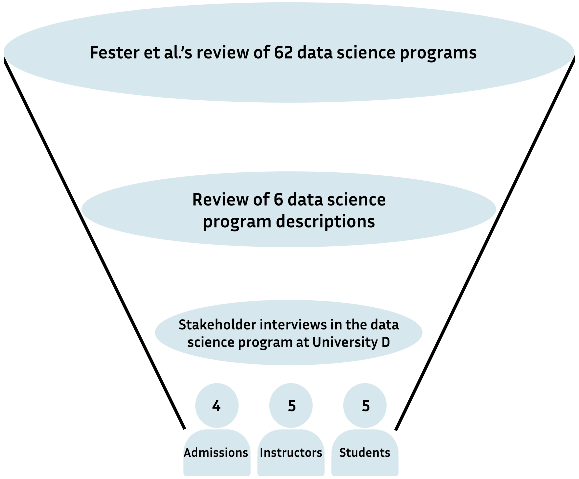

Fester et al.'s review of 62 data science programs
Review of 6 data science program descriptions
Stakeholder interviews in the data science program at University D
4
5
5
Admissions
Instructors
Students

Fig. 2. Illustration of our study method to identify the data science programs to study and for the in-depth stakeholder interviews.

### A.1 Admissions flowchart

### A.2 Python knowledge test illustration

## B Study protocol

We followed a qualitative research methodology, using semi-structured interviews to explore the experiences of admissions personnel, instructors, and students within the online applied data science graduate program. Two authors conducted each interview session: one led the discussion, and the other took notes. The lead interviewer explained the interview process, obtained consent to record the session, and ensured the meeting transcript could be reviewed and analyzed later. This dual approach allowed for a more thorough capture of the participants' responses and insights, ensuring that critical details were not missed during the interview.

We designed three different interview scripts tailored to each of the three different stakeholders— admissions personnel, instructors, and students. We structured these scripts to investigate each group's specific challenges and perspectives. The interview questions inquired about three main aspects: (1) understanding the participant's background, (2) their current admissions processes (for admissions personnel), their current teaching methods (for instructors), their

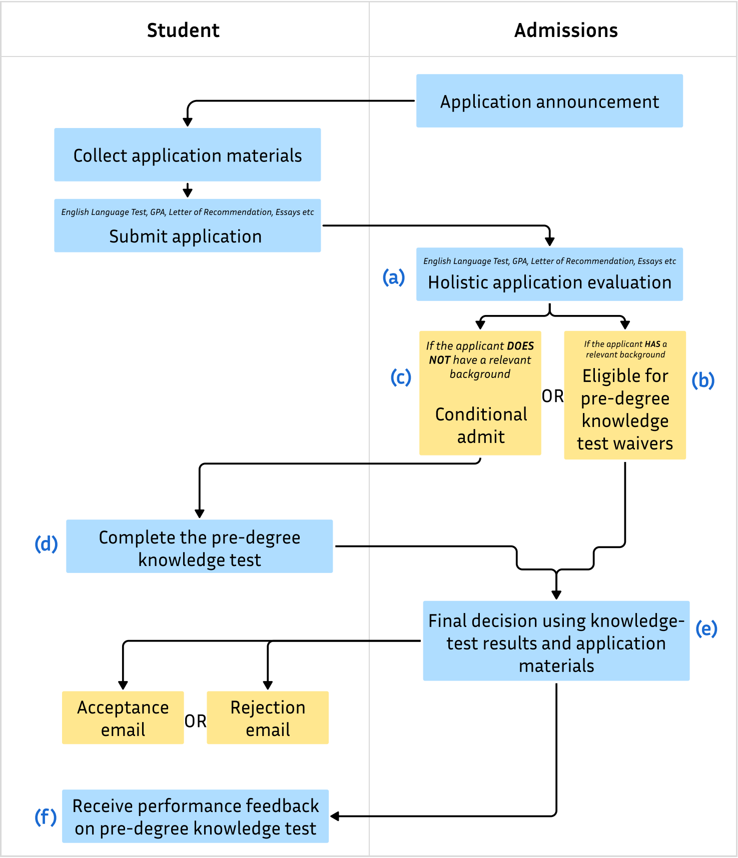
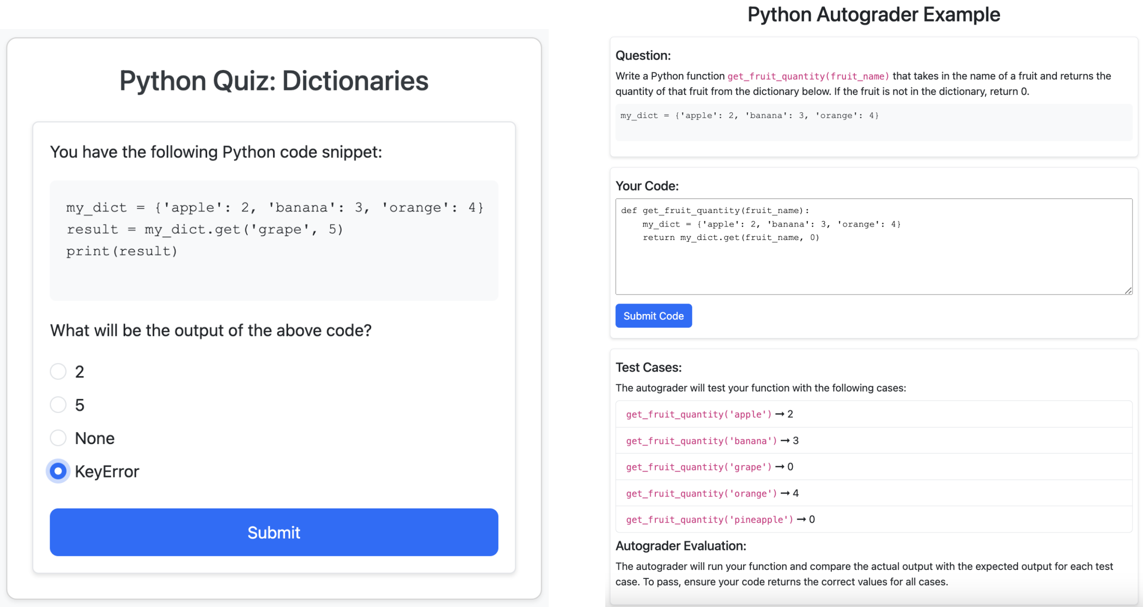

Student
Admissions
Application announcement
Collect application materials
English Language Test, GPA, Letter of Recommendation, Essays etc
Submit application
(a)
English Language Test, GPA, Letter of Recommendation, Essays etc
Holistic application evaluation
(c)
If the applicant DOES NOT have a relevant background
Conditional admit
OR
If the applicant HAS a relevant background
Eligible for pre-degree knowledge test waivers
(b)
(d)
Complete the pre-degree knowledge test
Final decision using knowledge-test results and application materials
(e)
Acceptance email
OR
Rejection email
(f)
Receive performance feedback on pre-degree knowledge test

Fig. 3. Workflow illustrating different steps that graduate data science programs may have. Applicants submit their application materials by the admissions deadline. Admissions may conduct a holistic evaluation of the applications and determine if they need a knowledge test or could be given a test waiver. Then, based on the knowledge test results and other application materials, admissions finally decide to admit or reject the applicant. Programs may provide post-assessment feedback to applicants, but most feedback is limited to the results of student's correct and incorrect questions on the knowledge test.

motivation to enroll in the program (for students), and (3) gathering each stakeholder's perspective on the current admissions process and knowledge assessment.

For admissions, we asked about 1) their background and specific tasks as admissions staff, 2) the current steps in prerequisite skill assessments, their challenges, and how they collaborate with instructors on assessments, and 3) what resources they provide students to support their preparation and why. Around the time of the admissions interview, admissions were considering alternative knowledge test formats like an interview-based knowledge assessment that mitigates integrity concerns and could better estimate the applicant's knowledge. We also asked admissions about their trade-offs for an interview-based knowledge assessment against the currently used static entrance knowledge test.

For instructors, we inquired about 1) their backgrounds, instructional activities, and teaching styles, 2) the prerequisite knowledge necessary to perform well in their courses, 3) instances where students lacked prerequisite course knowledge and how they handled such situations, and 4) their strategies to understand student's knowledge level. These questions



Python Quiz: Dictionaries

You have the following Python code snippet:

```
my_dict = {'apple': 2, 'banana': 3, 'orange': 4}
result = my_dict.get('grape', 5)
print(result)
```

What will be the output of the above code?

○ 2

○ 5

○ None

● KeyError

Python Autograder Example

Question:

Write a Python function `get_fruit_quantity(fruit_name)` that takes in the name of a fruit and returns the quantity of that fruit from the dictionary below. If the fruit is not in the dictionary, return 0.

```
my_dict = {'apple': 2, 'banana': 3, 'orange': 4}
```

Your Code:

```
def get_fruit_quantity(fruit_name):
    my_dict = {'apple': 2, 'banana': 3, 'orange': 4}
    return my_dict.get(fruit_name, 0)
```

Test Cases:

The autograder will test your function with the following cases:

`get_fruit_quantity('apple')` → 2

`get_fruit_quantity('banana')` → 3

`get_fruit_quantity('grape')` → 0

`get_fruit_quantity('orange')` → 4

`get_fruit_quantity('pineapple')` → 0

Autograder Evaluation:

The autograder will run your function and compare the actual output with the expected output for each test case. To pass, ensure your code returns the correct values for all cases.

(a) An illustration of a question testing Python fundamentals from the Python entrance knowledge test. The question in the illustration resembles the questions used in the program's knowledge test. We did not use any original questions for illustration to preserve integrity.

(b) An illustration of how the Autograder tool uses instructor-defined test cases to evaluate the student's code. In actual tests, evaluation test cases are hidden from the students and only accessible to the instructor. The tool strictly evaluates the student code's output against reference answers. E.g., if the student code outputs "2 " (extra space) and the reference answer is "2", the code will be marked incorrect, even though the student implemented the correct solution.

Fig. 4. Illustration of the Python entrance knowledge test and how the Autograder tool uses test cases behind the scenes to evaluate the student's code. We created custom illustrations to preserve anonymity.

helped us determine what prerequisite knowledge is necessary to be assessed in admissions and what alternative strategies domain experts (instructors) might use to assess students' knowledge.

For students, we asked about their: 1) educational background, work experience, and residency status(international or domestic), 2) their reasons for applying and choosing this specific program, including their career goals and expectations, 3) their experiences with the admissions process, specifically whether they took the pre-degree knowledge test or waived it due to relevant background, 4) the challenges they faced due to lack of any prerequisite knowledge in the program, and 5) how university provided resources helped them get over these challenges, or what other resources could have benefited them. We also asked students to compare appearing for an interview-based assessment conducted by an instructor with the entrance knowledge test regarding comfort and trust.

The study was reviewed by our institution's Institution's Review Board (IRB) and deemed exempt from IRB oversight.

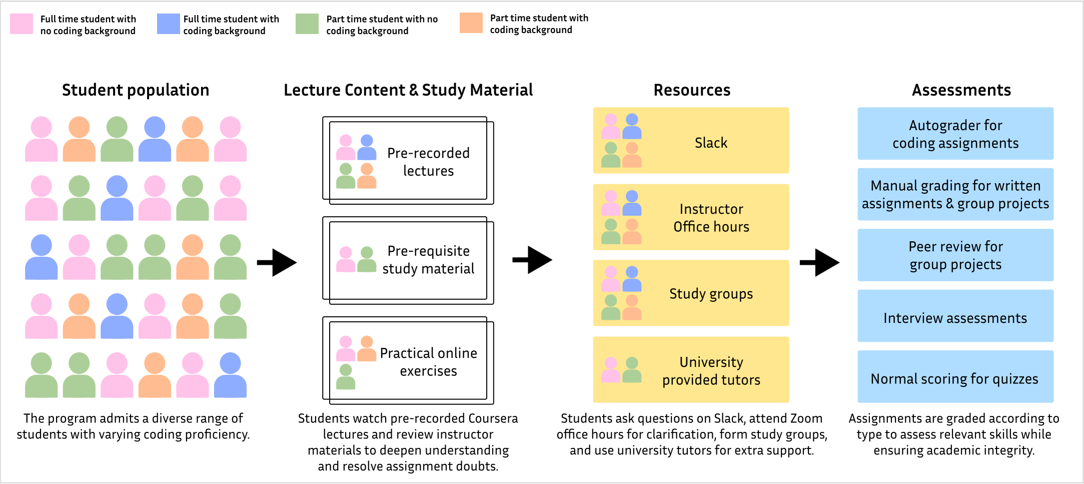

Full time student with no coding background
Full time student with coding background
Part time student with no coding background
Part time student with coding background
Student population
Lecture Content & Study Material
Resources
Assessments
Pre-recorded lectures
Pre-requisite study material
Practical online exercises
Slack
Instructor Office hours
Study groups
University provided tutors
Autograder for coding assignments
Manual grading for written assignments & group projects
Peer review for group projects
Interview assessments
Normal scoring for quizzes
The program admits a diverse range of students with varying coding proficiency.
Students watch pre-recorded Coursera lectures and review instructor materials to deepen understanding and resolve assignment doubts.
Students ask questions on Slack, attend Zoom office hours for clarification, form study groups, and use university tutors for extra support.
Assignments are graded according to type to assess relevant skills while ensuring academic integrity.

Fig. 5. Overview of student experience within the program for students from diverse backgrounds.