AI-ENABLED CLASSROOM TOOL FOR VISUAL LEARNING ANALYTICS

by

Ajay Kulkarni

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Sciences and Informatics

Committee:

_____	Dr. Olga Gkountouna, Dissertation Director
_____	Dr. Andrew Crooks, Committee Member
_____	Dr. Aditya Johri, Committee Member
_____	Dr. Feras Batarseh, Committee Member
_____	Dr. Jason Kinser, Department Chair
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Fernando Miralles-Wilhelm, Dean, College of Science

Date: _____ Spring 2022
George Mason University
Fairfax, VA

AI-Enabled Classroom Tool For Visual Learning Analytics

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Ajay Kulkarni
Master of Science
George Mason University, 2018
Bachelor of Engineering
University of Pune, India, 2013

Director: Dr. Olga Gkountouna, Assistant Professor
Department of Computational and Data Sciences

Spring 2022
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my family and many friends who supported me in this journey.

Acknowledgments

I would like to express my appreciation and gratitude to my advisor Dr. Olga Gkountouna for her constant support and encouragement. I also would like to thank the committee members, Dr. Feras Batarseh, Dr. Aditya Johri, and Dr. Andrew Crooks, for providing unique prospectives and advice to improve my work.

I am incredibly grateful to Dr. Olga Gkountouna for providing financial support to help my research. I also would like to thank Dr. Jason Kinser, Chair of the Computational and Data Sciences department, George Mason University, for providing teaching and funding opportunities. I am also thankful to DataLab and the Office of the Provost at George Mason University for their partial financial support. A special thanks to Natalie Lapidot-Croitoru, Karen Underwood, and Hillary Hamm for providing guidance on understanding and smoothly completing necessary paperwork throughout this period.

I wish to thank my friends Dr. Gideon Gogovi, Dr. Swabir Silayi, Dr. Ron Mahabir and Taylor Stevens for their constant support, direction and motivation throughout the program. I again would like to say thanks to Natalie Lapidot-Croitoru for being a good friend especially for meeting once a week for coffee.

Finally, I would like to express my gratitude towards my parents, who gave me everything and sacrificed a lot for me to fulfill my dreams. I forever will be grateful to them for the opportunities which they provided.

Table of Contents

List of Tables

List of Figures

Abstract

AI-ENABLED CLASSROOM TOOL FOR VISUAL LEARNING ANALYTICS

Ajay Kulkarni, PhD

George Mason University, 2022

Dissertation Director: Dr. Olga Gkountouna

This work focuses on simulation, design, development, and evaluation of a visual Learning Analytics (LA) tool - Real-time Educational AI-powered Classroom Tool (REACT) - to support educators' data-driven decision-making. The educational institutions face one of the biggest challenges, such as predicting student performance, detecting undesirable student behaviors, profiling or grouping students, etc., due to the exponential growth of educational data. The educators play a crucial role, where one of their primary responsibilities is effective, high-quality teaching. To do so, they must stay updated with students' responses, efforts, and outcomes, for providing timely feedback to promote students' improvement. Additionally, some of these educators are also academic advisors who provide advice to students, which is a critical aspect of judging institutional effectiveness. Considering these challenges, a solution in terms of an Artificial Intelligence (AI) driven visual LA tool is proposed in this work.

This work begins with a simulation approach for understanding the effects of a LA tool with alerts and recommendations on student performance. These simulations are performed by developing and testing an Agent-based Model (ABM) for the Department of Physics and Astronomy at a large public university. The positive results from this simulation study

indicated that the alerts and recommendations might help to increase student performance. Further, to understand the importance of the tool's design and its features, a high-fidelity prototype of REACT is developed using Shiny framework in R. The design and development of this tool followed recommendations from the golden rules of interface design and the Gestalt principles from visualization literature. Furthermore, considering the involvement of humans in educational applications, model-agnostic explanations are included on REACT for bringing explainability and interpretability in the process of decision-making. Finally, a study was conducted to understand the effectiveness, experience, and usability of REACT. The participants were 33 educators from Science, Engineering, and Humanities & Social Sciences. This study was performed using a hybrid approach of think-aloud interviews and questionnaires for exploring educators' perceptions. The study concludes that REACT was rated as highly usable by educators from the Science and Engineering domains who perceive their experience similarly. Their perception and experience in using this technology-focused tool differed from the educators of the Humanities & Social Sciences domain due to the technological knowledge gap in these fields, as exposed by the study's findings. The results also demonstrated that REACT has higher effectiveness and a higher likelihood of motivating behavior changes in educators from the Science and Engineering domains.

Chapter 1: Introduction

1.1 Artificial Intelligence (AI) and Learning Analytics (LA)

Educational institutions face one of the biggest challenges due to the exponential growth of educational data [5]. These challenges include but are not limited to predicting student performance, detecting undesirable student behaviors, profiling or grouping students, planning & scheduling, providing alerts & recommendations to stakeholders, etc. [6]. Considering these challenges, Artificial Intelligence (AI) techniques can help to provide personalized guidance as well as feedback to students and assist educators or policymakers in making decisions [7]. It has been noted [8,9] that these AI techniques enable computers to perform tasks via simulating intelligent human behaviors, such as inference, analysis, and decision making. Recently, due to a rapid growth in the advancement of computing and information processing, the AI applications in educational settings to facilitate teaching, learning, or decision making have also increased [7]. These applications are mainly to support humans, i.e., educators, students, etc. Thus, this makes it essential to understand human perceptions on interpretability, explainability, and trust before using these AI-powered applications in practice [10]. Further, it is vital to think about utilizing educational data from the social/pedagogical dimension [1], which makes it crucial to focus on Learning Analytics (LA). Therefore, considering these aspects, this work deals with simulation, design, development, and evaluation of a LA tool that combines LA and AI.

LA is "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [11]. LA is focused on data-driven decision-making and integrates the technical, social and pedagogical dimensions of learning by applying known predictive models [1]. Based on Larusson et al. [12], LA focuses on the six aspects.

1. Enhancing learner and faculty performance.

2. Finding, assessing, and attending to the needs of struggling learners.

3. Allowing instructors to determine and develop their strength.

4. Improving learners understanding of course material.

5. Helping to improve accuracy in grading.

6. Encouraging more efficient use of resources at the institutional level.

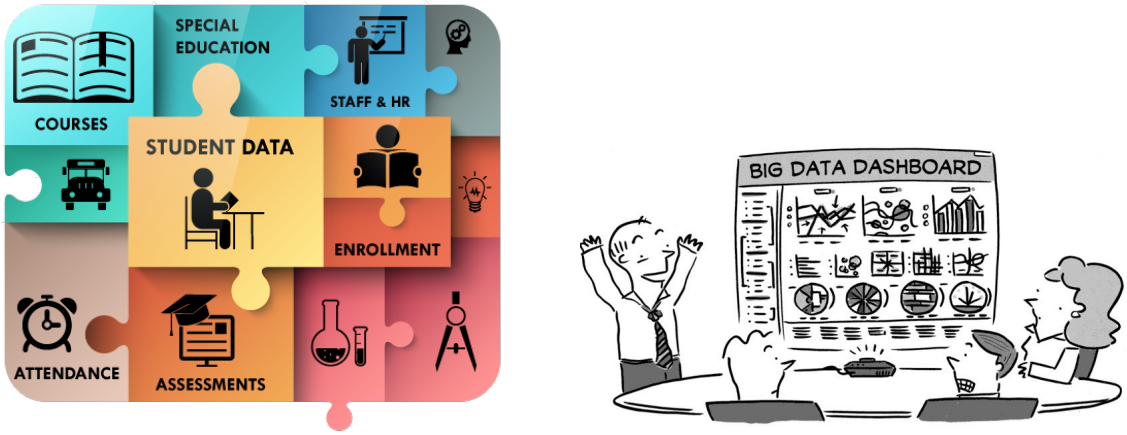


Figure 1.1: Educational data (Source: `https://ensemblelearning.org/`) and decision-making (Source: `https://thisisgraeme.me/`).

One way to achieve the above specified aspects is the use of AI techniques to find new and useful insights from the available educational data [13] as shown via a cartoon in Figure 1.1. It has been seen that in recent years the use of AI techniques on datasets from educational environments is common to answer critical educational questions [14–16] which emerged as AI in Edcucation (AIEd) [17]. Zhang and Aslan [17] investigated selected articles from

1993–2020 on AIEd and categorized six types of learning technologies - Chatboat, Expert systems, Intellignet tutors or agents, Machine Learning (ML), Personalized learning systems or environments and Visualizations. The authors also concluded four key challenges in AIEd which needs to be addressed to provide potential benefits for teaching and learning. These four key challenges are as follows.

1. Lack of actionable guidelines for educators.
2. Lack of AI expertise among educators.
3. Ethics and privacy.
4. Cost and scalability.

Recently, Aldowah et al. [18] surveyed 402 articles to understand the different applications of LA in education domain. Based on the survey the authors categorized these applications into four dimensions – Computer Supported Learning Analytics (CSLA), Computer Supported Predictive Analytics (CSPA), Computer Supported Behavioral Analytics (CSBA) and Computer Supported Visualization Analytics (CSVA) - and exposed that the majority of the research were focused on CSPA applications (253 articles, 63%) while the least amount of research were found on CSVA applications (38 articles, 9.50%). Further, the authors also identified twelve techniques - Classification, Clustering, Visual data mining, Statistics, Association rule mining, Regression, Sequential pattern mining, Text mining, Correlation mining, Outlier detection, Causal mining, and Density estimation - which has been utilised in LA applications. Based on the collected data the authors concluded, the CSVA area is still under-researched in education and encouraged researchers to utilise it with classification or clustering technique. In educational applications, cluster analysis or clustering can help to group/cluster students based on various characteristics such as their learning style preferences, academic performance, behavioral interaction, etc. This can help to explore collaborative learning opportunities and identify at-risk students at an early

stage [18]. Therefore, considering the importance of clustering and the need for CSVA applications, this work utilizes clustering in a CSVA application.

1.2 LA Knowledge Discovery Cycle

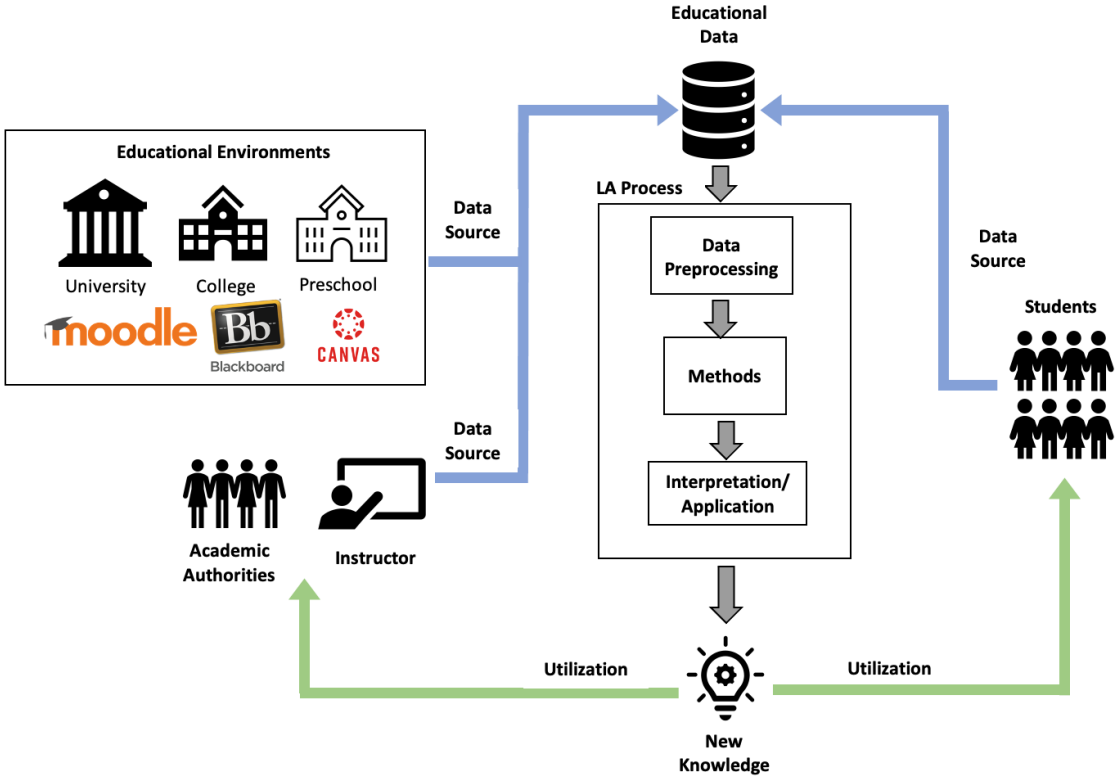


Figure 1.2: LA Knowledge Discovery Cycle process. This is a recreated figure based on an article published by Romero et al. [1].

The LA knowledge discovery cycle process is shown in Figure 1.2. It consists of five components – Educational environment, Educational data, Preprocessing, Methods, and Interpretation or application of new knowledge. The educational environment can be of any type, such as traditional face-to-face, online, or hybrid. In addition to that, Learning Management Systems (LMSs), Intelligent Tutoring Systems (ITSs) as well as the Massive Open Online Courses (MOOCs) are also considered as educational environments. The data generated from these educational environments can be of different types and granularity, as shown in Figure 1.3. It may include interactions between instructor and students, interactive exercises, in-class activity responses, administrative data, demographic data, etc. Romero

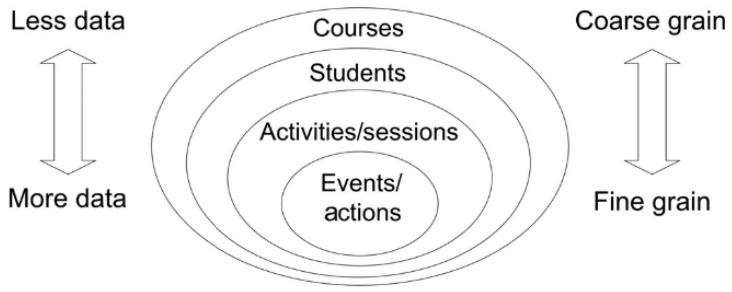


Figure 1.3: Levels of granularity and their relationship to the amount of data. (Source: Romero et al. [1]).

et al. [1] noted that the educational environments generate data from coarser grain level to fine-grain level along with multiple meaningful hierarchies such as answer-level, student-level, classroom-level, school-level, etc. Therefore, it is crucial to convert raw educational data in an appropriate format in the preprocessing phase. After the preprocessing phase, the data will be tidy, and ready for LA or AI methods. There are a variety of methods noted by Romero et al. [1] that include but are not limited to Clustering, Classification, Outlier detection, Process mining, Recommendations, Relationship mining, Visualizations,

Text mining, etc. which results in new knowledge discovery. The objective of this new knowledge discovery is its utilization in taking actions, making interventions, or making data-driven decisions to help learners, educators, and academic institutions where needed. This work closely follows the LA knowledge discovery cycle process by using student activity-level data, performing prepossessing & cluster analysis, and then, providing access to this new knowledge on a LA tool to educators.

1.3 Modeling & Simulation

Modeling & simulation is a technique for designing and evaluating complex systems. William Menner [19] said that it allows the construction of abstraction of systems and experimentation that otherwise would be cumbersome or impossible. Thus, modeling is "a representation of the construction and working of a system of interest", while the simulation is "a tool to evaluate the performance of an existing or a proposed system under different configurations of interest and over a long period of real-time" [20]. Modeling & simulation has four main applications [21] - scientific understanding, system development in technology, system management, and development planning – and can also help in cost reduction [22]. The models can be broadly classified into two types – 1) physical models, which are actual physical systems, and 2) mathematical models, which represents a set of computational or logical association [20]. Further, the models can be static (represent a system at a particular

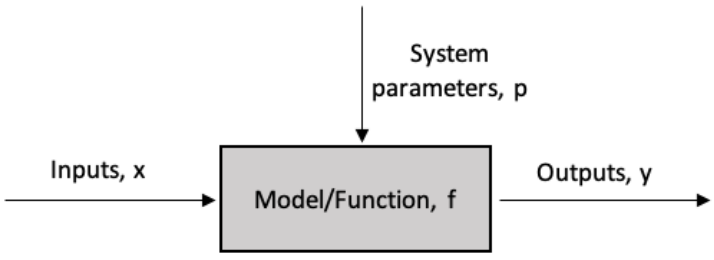


Figure 1.4: A functional view of the model.

point time) or dynamic (represent how a system changes with time) and stochastic (at least one random variable is present), or deterministic (when random variables are absent) [20]. A model can often viewed as a function $y = f(x, p)$ that produces output y from input x and system parameters p [23], as shown in Figure 1.4.

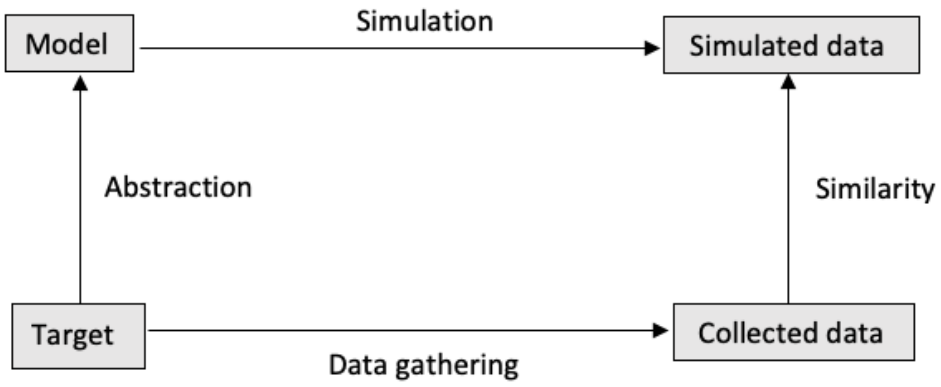


Figure 1.5: The simulation logic as a method consists of four components - Target, Model, Simulated data, and Collected data.

Gilbert and Troitzsch [24] presented simulation logic as a method indicated in Figure 1.5. It involves Target, Model, Simulated data, and Collected data. This process begins with developing an abstract model based on a social process, which is a computer program. This model is then simulated based on the conditions in the model for understanding the behavior. The simulations from the model generate simulated data which then compared with the collected data. Shiflet and Shiflet [25] explained five different approaches for model development in Computational Science. These five approaches along with their descriptions, are provided below.

1. System dynamics models - These models indicate global views of major systems that change with time.

2. Cellular automation simulation - These models present local views of individuals affecting individuals.

3. Agent-based simulations - These models include autonomous, decision-making agents who assess their situation and make decisions based on a set of if-else rules.

4. Empirical modeling - These models deal with finding a function that captures the trend of the data and then using this function to make predictions.

5. Matrix models - These models incorporate probabilities and averages used to make long-term predictions about system behaviors and populations.

Considering the different types of models, this work focuses on the Agent-based modeling approach because it is the best suitable approach for conducting detailed hypothesis-testing in the simulation experiments [26]. This may help to improve understanding of complex systems, interactions, and/or processes. Finally, Agent-based models are flexible and allow variations in the behavioral rules making them more suitable for educational applications. The most adopted paradigm for Agent-based modeling is Object Oriented Programming (OOP) [27]. As noted by Rob Allen [27], in OOP, the agents are considered self-directed objects that can choose actions autonomously based on the environment. For these reasons, Python (version 3) programming language is used for modeling & simulating an Agent-based Model (ABM).

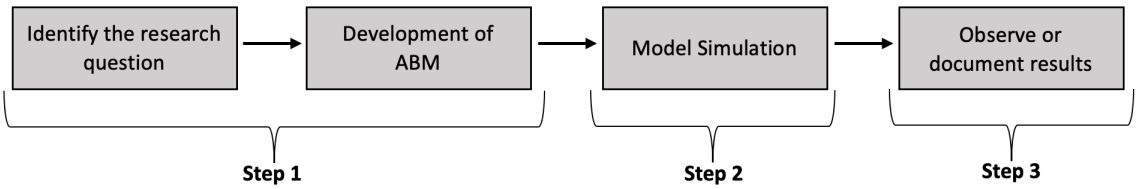


Figure 1.6: A three-step approach for developing an ABM. This diagram is motivated from Crooks et al. [2].

Crooks et al. [2] explained the three-step approach for developing an ABM, shown in Figure 1.6. The first step deals with identifying the research question and an output metric that will be studied based on the simulations of the model. Additionally, model parameters, initial conditions, and assumptions behind the model are decided in this step. The first step also deals with the creation and implementation of the rules. These rules are generally developed using if-then-else computer statements, mainly programmed using the OOP paradigm. The second step simulates the model, i.e., running or executing the model until a certain threshold or criterion is met. In the last step, the results from the simulations are recorded and can be used for evaluation. This three-step approach has been used for developing ABM.

1.4 Research Questions (RQs)

This section explains the research questions which are explored in this study. The main objective behind this study is to develop a LA tool for educators to support data-driven decision-making and understand its impact on them. This work focuses on the three aspects of LA - (i) Enhancing learner and faculty performance, (ii) Finding, assessing, and attending to the needs of struggling learners, (iii) Allowing instructors to determine and develop their strength. Additionally, this work also sheds light on one of the key challenges in AIEd - the lack of actionable guidelines for educators. The data-driven decision-making for educators can be considered as a part of institutional and teaching analytics. Institutional analytics generates institutional insight based on the courses, degree programs, research, revenue of students' fees, course evaluation, retention, graduation rate, resource allocation, and management [28, 29]. The teaching analytics deals with analyzing teaching activities, students' performance data, design, development, and evaluation of teaching activities [30]. Thus, this work focuses on both of these aspects. The first part of this work focus on institutional analytics, which utilizes a simulation approach via an ABM approach. The later part of the work focuses on designing, developing, and evaluating a visual LA tool for instructors which covers teaching analytics aspect.

1.4.1 Effects of a Learning Analytics (LA) Tool Via Simulation

The objective of the first research question is to understand the effects of a LA tool on graduation rates via simulations. The first research question which is answered in this work is as follows.

RQ1 - If educators used a LA tool for advising what effect will it have on graduation rates?

The above research question - RQ1 - is based on the following hypothesis.

H_1: If educators use a LA tool with alert and recommendation components, then it can help to increase the college graduation rate.

The validity of the above hypothesis is tested by simulating the behavior of a LA tool with alert and recommendation components. To achieve this, an ABM is developed based on the core courses from the Department of Physics and Astronomy at a large public university. ABM is a simulation technique that contains a collection of autonomous decision-making agents [31]. Every agent in ABM has specific attributes and interacts with the environment based on the provided rules [32]. Triulzi et al. [33] stated that the true dynamics in universities are difficult to model with conventional quantitative analysis. Thus, ABM can be one of the approaches used to model the true dynamics. Therefore, for these reasons, a simulation-based approach is utilized to model a LA tool's effect and aims to answer the RQ1. This modeling & simulation experiment explores four different situations - baseline, alerts, recommendations, and alerts & recommendations. To test the hypothesis and answer RQ1, the ABM is simulated for 100 runs using 100 agents, and the average college graduation rates are calculated for every situation. Further paired t-tests are used to test and confirm the validity of the hypothesis.

1.4.2 Usefulness and Perceptions of a LA tool

The results from the RQ1 indicated that a LA tool with alert and recommendation components may create a positive impact. These results direct towards a positive direction in developing a LA tool for educators and emphasize that features and design will play a significant role. The development of a tool is an iterative and time consuming process which includes - Analysis, Design, Implementation & Coding, Testing, Deployment and Maintenance [34]. Shum et al. [35] mentioned that it is important to account a range of human factors, including why and how they will use it while incorporating analytics on a LA tool. Thus, the process of development should be human-centred. Ahn et al. [36] suggested to conduct usability analyses to understand interface utilisation by using common data collection techniques, such as user interviews and think-aloud, which will also give insights into educators sensemaking needs. A study conducted by Wise and Jung [37] explored that involvement of educators throughout a LA tool development and conducting early studies can provide important insight into tool design for local actionability. Finally, Holstein et al. [38] recommended to understand the behaviour of LA tools using real-world datasets. Thus, at present, considering these guidelines, time constraints, and privacy aspects of obtaining real-world data of students, it is not possible to develop a LA tool to cover all the responsibilities of advisors presented in the literature [39–41]. Although, it is possible to develop a prototype of a LA tool that can aim to fulfill two responsibilities - (i) detecting and advising students whose performance is degrading, and (ii) recommending academic resources - which also coincides with the responsibilities of instructors. This approach also gives a possible solution to provide actionable guidelines for educators, which is one of the key challenges in AIEd. Thus, a visual a LA tool which can provide alerts, recommendations, and other analytical insights to instructors is designed, developed and evaluated to fulfill the above requirements. The RQ2 and RQ3 focused on the above aspects are mentioned below.

RQ2 - How can a visual LA tool - REACT (Real-time Educational AI-powered Classroom Tool) - that incorporates Visual Analytics (VA) and AI be helpful in classrooms?

The above research question highlights REACT's design & development process, features, and usefulness. To answer RQ2, a high-fidelity prototype of REACT is developed, and evaluation experiments are performed. These experiments were performed with 19 educators by conducting think-aloud studies. The participated educators were divided into three domains - Science, Engineering, and Humanities & Social Sciences - based on their field of work. The comments and activity patterns from the educators were recorded in a logbook.

RQ3 - How do educators from different domains perceive the integration of VA and AI in real-time on REACT?

The above research question - RQ3 - is based on the following hypothesis.

H_2: If educators from different domains use REACT, then they will show a similar perception based on Kirkpatrick's four-level evaluation model.

To answer the RQ3 and verify the above hypothesis, questionnaire studies were conducted with 33 educators from three domains – Science, Engineering, and Humanities & Social Sciences. A combined questionnaire based on Kirkpatrick's four-level evaluation model [42] and System Usability Scale (SUS) [43] is used to explore educators' perceptions. Finally, composites scores are calculated, and acceptance thresholds are used as a metric for testing the hypothesis.

Additionally, a statistical hypothesis is also tested based on the collected data from questionnaires. This hypothesis is based on the correlation coefficient and provided below.

H_3: The participants' scores on usability have significant correlations with Reaction,

Learning, Behavior, Result, and Effectiveness criteria[1].

$$\textit{Null Hypothesis: } H_0 : r = 0$$

$$\textit{Alternate Hypothesis: } H_\alpha : r \neq 0$$

The above *Null* and *Alternate* hypothesis can be expressed in words as follows.

> *Null Hypothesis: There is not a significant correlation between the two variables.*
>
> *Alternate Hypothesis: There is a significant correlation between the two variables.*

The main objective behind the above statistical hypothesis testing is to shed light on the strength of A relationship between SUS, Kirkpatrick's four-level evaluation model, and the effectiveness of a visual LA tool - REACT.

Research Techniques

This work utilizes qualitative and quantitative research techniques for answering the RQ2 and RQ3, respectively. A qualitative research technique - observation - has been used to explore the RQ2. Qualitative research deals with gathering data that is non-numerical, and it attempts to make sense or interpret phenomena in terms of the meaning people bring to them [44]. There are different approaches for quantitative research, including direct observation, user interviews (audio/video recording or interviewers notes), open-ended questions via questionnaires, analysis of artifacts, etc. [45]. The observation technique is helpful, especially in evaluating prototypes and investigating their support to achieve tasks and goals. Russell Bernard [46] mentioned that the users interact with a prototype and perform activities that investigators further study in observation technique. Therefore, due to these reasons, the think-aloud technique is utilized to answer RQ2. The activity patterns and comments were noted in a logbook during think-aloud studies. This collected data is then analyzed by categorizing data, an inductive analysis technique - extracting concepts from the data [47] – that helps to identify the usability problems.

[1] based on Kirkpatrick's four-level evaluation model.

In our efforts to answer RQ3, we have employed two quantitative research techniques. Quantitative research deals with the gathering of data which is in a numerical form and is defined as "explaining phenomena by collecting numerical data that are analyzed using mathematically based methods" [48]. There are different quantitative research approaches: survey research, correlation research, experimental research, and causal-comparative research. To answer the RQ3, survey and correlation research techniques are utilized. The data from the educators were collected based on a survey, and then composite scores were calculated. These composite scores are more reliable and representative of educators' attitudes than the individual scores [49, 50]. Next, these calculated composite scores are used for understanding the perceptions of different educators from different domains. The correlation research helps investigate the extent to which two or more variables are related [51]. Grove et al. [52] mentioned three types of correlation research design – descriptive, predictive, and model testing. The descriptive correlation research technique is most suitable and used to answer RQ3 by considering the need to understand the relationship between different variables.

1.5 Relevance to the Computational Sciences & Informatics (CSI)

This work is at the intersection of LA and AI combined using Computational Science tools and techniques from Informatics. Computational Science is an interdisciplinary discipline that is "at the intersection of the sciences, computer science, and mathematics" [25]. It mainly uses modeling which is "the application of methods to analyze complex, real-world problems in order to make predictions about what might happen with various actions" [25]. As noted by Shiflet and Shiflet [25] there are various approaches for modelling which includes but are not limited to System Dynamics models, Data-driven models, Agent-based models, Matrix models etc. The work which presented here focuses on Agent-based and Data-driven models. Thus, RQ1 and RQ2 are directly connected to the Computational Science aspect.

Further, Informatics is focused on Data Science[2] which involves "principles, processes, and techniques for understanding phenomena via the (automated) analysis of data" [53]. The ultimate goal of Data Science is data-driven decision making which is "the practice of basing decisions on the analysis of data rather than purely on intuition" [53]. As shown in Figure 1.7, there are different elements which contributes to Data Science and their contribution depends on the problem. Considering these elements, this work directly amalgamates Statistics, Algorithms, Machine Learning, and Visualisation & Visual Analytics. Thus, RQ2 and RQ3 cover the Informatics aspect of CSI.

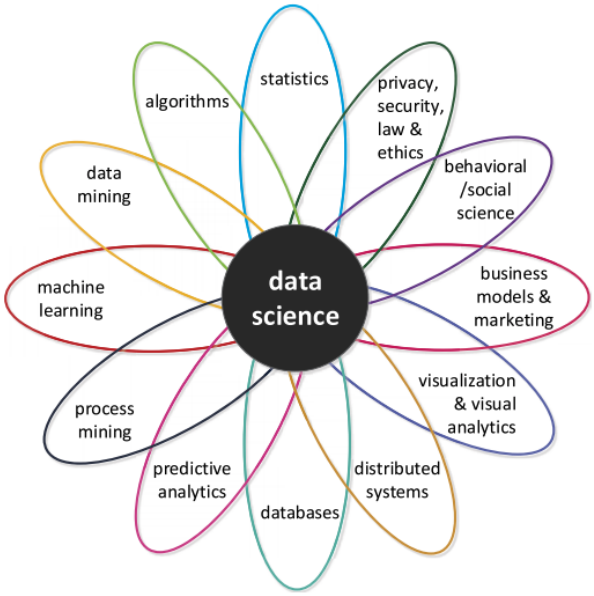


Figure 1.7: Different elements which contributes in Data Science. This figure is taken from a book chapter from "Data science in action" authored by Wil Van Der Aalst [3].

The overall structure of this dissertation is as follows. The detailed discussions on ABM design, simulated situations, results, and other related information are presented in

[2]https://tinyurl.com/e364hzu5

Chapter 2. The background on LADs and unsupervised learning - mainly focused on clustering - are provided in Chapter 3. Additionally, the information on proposed architecture, features, and utilization of Model-agnostic explanations is also presented in Chapter 3. Finally, the proposed LAD's effectiveness, usability, and impact are studied by conducting questionnaires and think-aloud-based studies. The details on experiment design and results obtained from these studies are documented in Chapter 4. In the end, the conclusions and a future direction of research are explained in Chapter 5.

Chapter 2: Estimating Effects of a Learning Analytics (LA) tool on Educational Agents With Simulations

2.1 Introduction

This chapter presents details on the development of an Agent-based Model (ABM) and explains the importance of simulations while designing a tool. The Agent-based modeling is a bottom-up modeling approach that contains a collection of autonomous decision-making agents [31, 54]. Every agent in an ABM is an entity having attributes and interacts with the environment based on the provided rules [32]. The ABM is an advantageous approach for depicting real-world scenarios becuase these scenarios are more complex and involve complex interactions. Eric Bonabeau [31] specified three benefits of ABM - (i) it captures emergent phenomena, (ii) it provides a natural description of a system, and (iii) it is a flexible approach. There are various applications of ABM which includes but are not limited to Epidemic Modeling, Anthropology, Biomedical Research, Chemistry, Crime Analysis, Ecology, Market Analysis etc. [55] but applications in the domain of education are limited [54]. X. Gu and K.L. Blackmore [54] conducted a systematic review of ABM and simulations in the education domain. The authors identified six applications - university system, university collaboration, academic activities, application & enrolment, student performance, and teaching & learning – and for student performance, authors found only three articles. These three articles are based on online peer support [56], students' grades, and graduate employment [57, 58]. Considering this information this study focuses on one of the applications of student performance enhancement. It is pursued by understanding the effects of utilization of alerts and recommendations on a Learning Analytics (LA) tool. To understand these effects a scenario for undergraduate academic advising is considered because it

is one of the critical aspects to enhance students' academic performance and institutional effectiveness [39, 59]. In this study, the agents are students who interacts with advisors in the environment which is an institution. This is studied by considering four different situations and simulated results are compared based on the calculated graduation rates. The results from this chapter answers – *if educators used a LA tool for advising what effect will it have on graduation rates?*

2.2 Background

College graduation plays an essential role for students as well as for the institutions [59,60]. Along with the college graduation rate, the second most crucial aspect for any institution is academic advising [39]. Academic advising positively affects student retention [61–63] and it helps students in decision making, resource identification, problem-solving as well as for goal (personal, professional and academic) setting [64–67]. A study conducted by Swecker et al. [66] based on 363 first-year first-generation students concluded that every meeting with an academic advisor increased 13% chance of student retention. Kirk-Kuwaye et al. [68] conducted a study to understand the effect of low and high advisor involvement on the academic performance of probated students. To perform these experiments, low involvement and high involvement groups were created. For high involvement group, activities such as mandatory meetings, agreement to use resources, study strategy materials/web sites, reminder phone calls, and assignments along with a letter of notification were provided. For the low involvement group, only the letter of notifications was provided by the advisors. Based on the results, authors concluded that there was a high semester mean GPA for students from high involvement group as compared to the low involvement group. Also, students from high involvement group never felt annoyed by the involvement of the institution at a higher level. Based on another study conducted using surveys of 611 students [69], it was reveled that that meeting with an advisor at least one time during a semester contributed to the multiple factors affecting students' success. Felly Chiteng

Kot [70] also conducted a study to understand the impact of centralized advising on the performance of First-Year students and its effect on the enrollment for the second year. Based on the conducted study author noted that for the students who utilized the service of centralized advising there was a net gain of 31 percentage points in the first-term GPA, 22 percentage points on average in the second-term GPA and 25 percentage points on average in the Cumulative GPA at the end of the first academic year. This study also showed that students who used centralized advising were more likely to return in the second year. It also has been noted that the advisors who are well organized, on time, and prepared for meetings are highly effective [39]. Further, based on the literature [39–41] the five most important responsibilities of the advisors are - guiding students for selecting their majors and minors, helping them to choose relevant and useful courses, detecting and advising students whose performance is degrading, providing valuable academic resources, and informing sources of help & activities offered through Student Affairs Division to students. This collected literature reflects the importance of advisors not only for students but also for the institutional effectiveness. Therefore, in this study through a lens of ABM & simulation a possible solution for supporting advisors using a LA tool is explored.

2.3 Data and Course Dependency Maps

The ABM is designed for the department of Physics and Astronomy at a large public university. For the simulation experiments, information on core courses is used and taken from the advising webpage of the university. The detailed list of courses and prerequisites used in the ABM is provided in Table 2.1. From the list it can be seen that the requirement of the prerequisites starts from the second semester and thus, indicates a need of data for defining scores of agents in the first semester. The literature [71–73] indicated that the Scholastic Assessment Test (SAT) scores are one of the critical predictors of the student's performance in the first semester and are useful to measure a student's potential for academic success in college. Thus, the average SAT scores [74] for Mathematics, Critical Reading, and Writing sections are utilised as initial conditions for simulating the performance of students in the

first year. For maintaining diversity in the SAT scores, they are generated using random normal distribution with different means and standard deviations for different sections. The parameters used for generating SAT scores are given in Table 2.2.

Table 2.1: Core courses and prerequisites for the department of Physics and Astronomy.

Semester	Core courses	Prerequisites
Fall (Year 1)	MATH 113(Analytic Geometry and Calculus)	
	PHYS 122/123 (Inside Relativity/ Inside the Quantum World)	
	ENGH 101 (Composition)	
Spring (Year 1)	MATH 114 (Analytic Geometry and Calculus II)	MATH 113
	ASTR 124 (Introduction to Observational Astronomy)	
	PHYS 160 (University Physics I)	MATH 114
	PHYS 161 (University Physics I Laboratory)	PHYS 160
Fall (Year 2)	MATH 213 (Analytic Geometry and Calculus III)	MATH 114
	PHYS 260 (University Physics II)	PHYS 160, MATH 213
	PHYS 261 (University Physics II Laboratory)	PHYS 161, PHYS 260
	PHYS 251 (Introduction to Computer Techniques in Physics)	PHYS 160
Spring (Year 2)	MATH 214 (Analytic Geometry and Calculus II)	MATH 213
	PHYS 307 (Thermal Physics)	PHYS 260
	PHYS 308 (University Physics I)	PHYS 260
	ASTR 210 (Introduction to Astrophysics)	PHYS 160
	Elective – PHYS 265 (Intermediate University Physics Laboratory)	PHYS 251, PHYS 260
Fall (Year 3)	PHYS 301 (Analytical Methods of Physics)	MATH 214
	PHYS 303 (Classical Mechanics)	PHYS 260, PHYS 301
	PHYS 305 (Electromagnetic Theory)	PHYS 260, PHYS 301
	PHYS 311 (Instrumentation)	PHYS 251, PHYS 261
	ENGH 302 (Advanced Composition)	ENGH 101
Spring (Year 3)	PHYS 306 (Wave Motion and Electromagnetic Radiation)	PHYS 305
	PHYS 312 (Waves and Optics)	PHYS 251, PHYS 261
	PHYS 402 (Introduction to Quantum Mechanics and Atomic Physics)	PHYS 303, PHYS 305, PHYS 308
Fall (Year 4)	PHYS 403 (Quantum Mechanics II)	PHYS 402
	PHYS 407 (Senior Laboratory in Modern Physics)	PHYS 251, PHYS 311, PHYS 312, PHYS 402
	PHYS 408/409 (Senior Research/ Physics Internship)	PHYS 251, PHYS 301, PHYS 303, PHYS 305
	PHYS 410 (Computational Physics Capstone)	PHYS 303, PHYS 305, PHYS 251, PHYS 265
	PHYS 416 (Special Topics in Undergraduate Physics)	
Spring (Year 4)	PHYS 412 (Solid State Physics and Applications)	PHYS 402/502
	PHYS 428 (Relativity)	PHYS 303, PHYS 305

Table 2.2: The average SAT scores, mean and standard deviation values for Mathematics, Critical Reading, and Writing sections.

	Variable	**Value**
SAT score parameters	Average critical reading score	558
	Standard deviation for critical reading	10
	Average math score	585
	Standard deviation for math score	4
	Average writing score	540
	Standard deviation for writing score	10

It can also be observed that there are some courses which need two or more than two prerequisites. For example, PHYS 260 needs two prerequisites while PHYS 407 needs four prerequisites. Based on this data i.e., the core courses and prerequisites, course dependency maps are created for every year which graphically represent the data provided in Table 2.1. An example of a dependency map for the third year is shown in Figure 2.1. These dependency maps are then used while implementing different rules in the ABM.

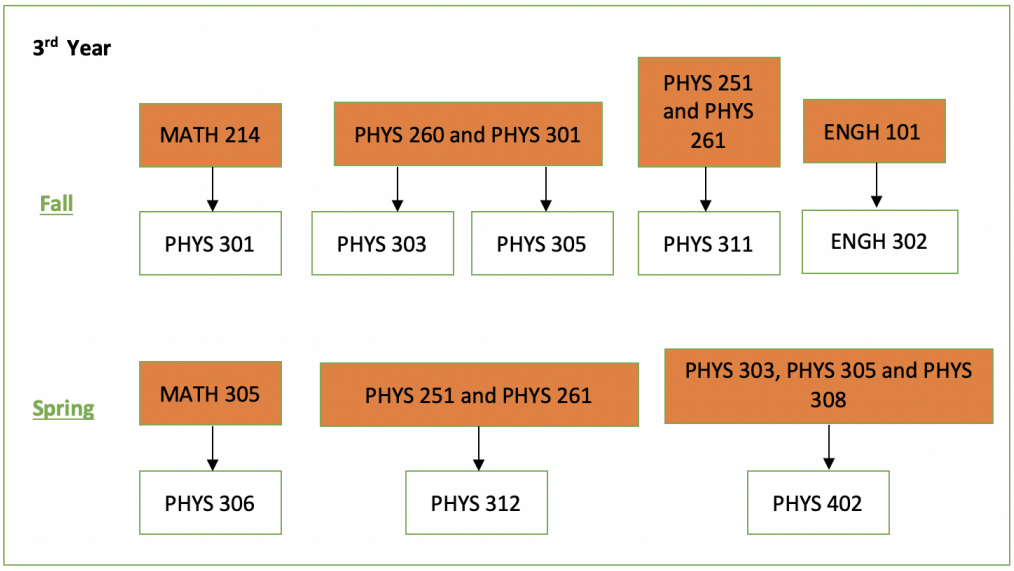


Figure 2.1: Course dependency map indicating prerequisites and core courses for the Fall and Spring semesters of the third year.

2.4 Agent-based Model design

The agents in the ABM were simulated based on the core course dependency maps, which are developed using the information of core courses and prerequisites from Table 2.1. Every agent in the ABM is associated with 56 attributes, which includes their id, SAT scores

(Mathematics, Critical Reading, and Writing), core courses, semester wise GPA, final GPA, and a flag variable – "fail" - for understanding count of students who failed. The rules of the ABM are developed for five unique cases which includes - first semester, courses with one prerequisite, courses with two prerequisites, courses with three and four prerequisites. The ABM explores following four situations which help to understand and compare the effects of a LA tool.

Situation 1 (Baseline): In this situation it is assumed that students will take initiative to contact advisors and a LA tool will not be available to advisors.

Situation 2 (Alerts): The second situation simulates LA tool with an alert system that will provide alerts to advisors on poorly performing students. The LA tool will provide alerts based on two conditions - (i) if the difference between any two consecutive grades is greater than or equal to 0.33 or (ii) if a student earns less than a "C" grade in one or more core courses. The difference between any two consecutive grades is 0.33 at a large public university. For example, the difference between grades A (quality points = 4) and A- (quality points = 3.67) is 0.33. Thus, in the first condition the difference of 0.33 is used for comparison. Additionally, one of the requirements of the core courses is to earn at least a "C" grade and this requirement is covered in the second condition.

Situation 3 (Recommendations): The third situation assumes that advisors will only receive recommendations on courses that can help students to increase their GPA. These recommendations will be based on the student's performance and advisors will provide these recommendations to students if they think they are helpful.

Situation 4 (Alerts & Recommendations): The fourth situation assumes that advisors have access to a LA tool which can provide alerts as well as recommendations.

The outlined situations are simulated 100 times using 100 agents for three runs. This ABM is designed on the hypothesis — "alert and recommender systems can help to increase

the college graduation rate at the end of the fourth academic year" and thus the average change in the graduation rate is used as a metric. The first situation is a baseline situation, and other situations are compared to understand which situation majorly contributes towards the improvement in the graduation rate. For situations 2 and 3, the results are simulated to understand how 5%, 10%, 15%, and 20% performance increase in every semester affects the graduation rate. The situation 4 is a combination of situations 2 and 3. In the first case of situation 4, it is assumed that there will be a 5% increase in the performance, and in the second case, it is assumed that there will be a 15% increase in the performance. The results from these experiments give insight into a crucial aspect – Will these scenarios affect graduation rate? The initial conditions, rules, and simulated situations depicted in this study are motivated from an article published by Kulkarni and Eagle [75].

2.5 Limitations

There are several limitations to this experimental design. First, this model is developed, i.e., the rules are designed by considering the SAT scores for the Science domain and using the core courses for the Department of Physics and Astronomy. Second, we are assuming that students will get enrolled only in the Fall semester. It is also possible to enroll in academic institutions in the Spring semester, but this condition is not included in the current model. Third, there may be a possibility that some students can take a break and return to school. This possibility is also not incorporated in the model. Finally, student-student interactions are not depicted in the model, which happens in academic institutions. However, despite these limitations, we believe that this study may be a good step in designing and developing ABMs for educational applications. It may raise more awareness in the education domain to explore modeling & simulation techniques.

2.6 Results

The simulated results are presented in Table 2.3, Table 2.4 and Table 2.5. During simulations it was observed that overall failure rate is high in the first academic term, and it is highest at the end of the second year. After the second academic year, there is a relatively less percentage of failure students. Next, for the baseline situation, 61.38% average graduation rate was noted. Further, it has been found that for alerts (situation 2), there is an overall increase in the average graduation rate. For recommendations (situation 3), one of the scenarios indicates a decrease in the graduation rate, suggesting that recommendations may not always be useful. For the combination of alerts & recommendations (situation 4), a gradual increase in the graduation rate is observed, suggesting that a combination of alerts and recommendations may provide positive results. To confirm the statistical validity of the results, the paired t-tests are performed. In paired t-test, the null hypothesis indicates there is no difference between the two means i.e., $\mu_1 = \mu_2$ and alternative hypothesis specifics there is a difference between the means of the two samples, i.e., $\mu_1 \neq \mu_2$. To perform a paired t-test, 100 samples for every scenario are used. Before performing a paired t-test, the assumption of normality is checked using the Shapiro-Wilk normality test for the differences. In the Shapiro-Wilk normality test, the null hypothesis indicates that the data are normally distributed, and the alternative hypothesis specifies that the data does not fit the normal distribution. The assumption of normality checked using an alpha level of .05 and observed that all the variables resulted in p-values greater than .05. It indicates the fulfillment of the normality assumption. After testing the assumption of normality, paired t-tests conducted with an alpha level of .05 to determine the significance of differences in the mean graduation rate for different scenarios. Situation 1 used as a baseline and other situations were compared to check the significance of the differences in the graduation rate. There was a significant difference in the graduation rate for scenario 1 (M = 60.76, SD = 6.95) and scenario 3 when there is 15% increase in the performance (M = 64.21, SD = 7.77); t(99) = -3.01, p = .003. Significant difference is also noted in the graduation rate of

scenario 1 (M = 60.76, SD = 6.95) and scenario 4 (case 2) when there is 20% increase in the performance (M = 63.64, SD = 7.58); t(99) = -2.98, p = .003. These results indicate that a LA tool with alert and recommendation components may create a positive impact. Thus, the next chapter, provides details on designing and developing a Learning Analytics Dashboard (LAD) - a LA tool - that can provide alerts, recommendations, as well as other analytical insights to educators.

Table 2.3: For situation 1, the average graduation rate is 61.38%.

Number of runs	**Average graduation rate at the end of 4th year**
1	60.76%
2	62.29%
3	61.11%

Table 2.4: The overall average graduation rate increases from 0.1% to 0.70% in Scenario 2, and in Scenario 3, there is an increase in the graduation rate for all cases except a 10% increase in the performance.

Number of runs	**% increase in the performance**	**Average graduation rate at the end of 4th year**	
		Scenario 2	**Scenario 3**
1	5%	61.41%	60.87%
	10%	62.74%	60.44%
	15%	62.69%	62.02%
	20%	62.43%	61.43%
2	5%	61.58%	61.52%
	10%	61.72%	61.87%
	15%	62.28%	64.21%
	20%	62.22%	62.62%
3	5%	61.46%	62.64%
	10%	61.46%	60.34%
	15%	61.09%	61.92%
	20%	61.62%	63.23%

Table 2.5: The overall average graduation rate gradually increases from 0.16% to 0.53% and 0.07% to 0.73% in Case 1 and Case 2 of Scenario 4 respectively.

% increase in the performance	**Mean percentage change in the graduation rate (Scenario 4)**	
	Case 1	**Case 2**
5%	0.16%	0.07%
10%	0.37%	0.31%
15%	0.43%	0.43%
20%	0.53%	0.73%

Chapter 3: Design & Development of Real-time Educational AI-powered Classroom Tool (REACT)

3.1 Introduction

People are visual learners, and data visualizations help them experience information [76,77]. Visual Analytics (VA) is an application of data visualization which deals with helping users to gain insights into complex data [78,79]. It employs interactive visualizations as interfaces between users and their data, making data-related tasks more effective and efficient [80]. VA tools have been successfully applied in many domains, but their applications in education are still limited [18,81]. There are many web-based environments (e-learning systems, intelligent web-based educational systems, learning management systems, etc.) that generate large amounts of educational data but most of them does not provide suitable tools for utilizing these data to improve learning or teaching [82]. A well-designed VA tool in the educational context can potentially provide useful information to instructors by helping them provide formative feedback and understand as well as optimize the students' learning process [78]. Further, to promote students' improvement with effective high-quality teaching educators must stay updated with students' responses, efforts, and outcomes [83, 84]. One of the solutions to achieve these objectives can be clustering students into groups based on various characteristics (learning style preferences, academic performance, behavioral interaction, etc.) that can be utilized to explore collaborative learning opportunities and identify at-risk students at an early stage [18]. The above approach can be effectively executed by combining VA with Artificial Intelligence (AI) on a Learning Analytics Dashboard (LAD).

Learning Analytics Dashboards (LADs) are a "special kind of display of multiple visualizations about different indicators of learner(s), learning process(es) and/or learning

context(s)" [85] that can aid to improve learning. They may promote awareness, reflection, and sensemaking as they help to process large amounts of data in a meaningful way by visualizing the traces of the learning activities [86, 87]. These traces can help instructors to identify weak spots in learning and topics in which students struggle. Showing this information in a timely and accurate way is of utmost importance for achieving teaching objectives in the classroom [85]. A well-designed LAD can aid instructors regarding potential pedagogical strategies, instructional guidance, actions, and interventions to support students' participation and promote their success [82]. Thus, considering these factors this work proposes Real-time Educational AI-powered Classroom Tool (REACT), a web-based VA tool in the form of an interactive LAD. It aims to help instructors by tracking students' activities and providing detailed insights on their responses. It can also support the instructors' decision-making process with VA, contextualized alerts, and recommendations in real-time. This chapter provides information on the architecture of REACT, features, and utilization of Model-agnostic explanations via a use-case.

3.2 Background

3.2.1 Learning Analytics Dashboards (LADs) & Artificially Intelligent (AI) tools

LADs leverage information visualization and data analytic techniques to explore log data recorded by Learning Management Systems (LMSs) [88]. This data can help educators to diagnose problems concerning participation of students [89]. Visualising these data on a LAD help to maintain situational awareness [90], that is understanding what is happening in our surroundings and how to use this information now and in the future [91]. Park and Jo [92] mentioned that a LAD's visual attraction significantly affect the level of understanding of the presented information. Further, the level of understanding affects the perceived usefulness, which substantially affects potential changes in users' behavior. Thus, these factors need to be considered while designing LADs. Recently, a growing interest in the

design and development of real-time LADs is observed such as RAED [93], MTClassroom and MTDashboard [94], DREAM, REALTO [95], My Learning Progress [96] etc. which can provide actionable teaching analytics in real-time for decision making. These real-time LADs are beneficial because they give more time to instructors to provide one-on-one support to students [97]. Additionally, a variety of other applications of LADs has also been noted in the literature such as to facilitate communication between advisors and students by visualizing grade data [98], tracking and visualizing learners' emotions during online classes [99], academic advising [100], adaptive support for face to face collaborative argumentation [101], and adaptive guidance in mathematics classrooms [102]. However, none of the above LADs considered the design principals such as golden rules [85] and Gestalt principles [103] for interface design or provides personalized recommendations to instructors. Contrary to the aforementioned examples, REACT is not only a LAD which provides reporting functionalities, but it is also an AI based decision support tool.

AI tools are used for decision making in a broad range of industries. The applications of AI in the educational domain, such as predicting student performance, detecting undesirable student behavior, or providing feedback for supporting instructors and students, are gradually gaining popularity [1]. However, many of these tools are seen as black boxes, meaning it is difficult to get insights into the workings of their methods [104]. The lack of transparency of these tools may result in unchecked bias that can negatively affect the quality of decision making [105, 106]. Thus, it is important that the tools which utilizes AI should be interpretable, explainable, and, ultimately, trustworthy for supporting human learning and teaching [10, 107]. Recently, the European Union (EU) passed a regulation requiring algorithms to provide explanations that can significantly affect users based on their user-level predictions [108]. From the AI context, explainability denotes the action taken by a model to detail its internal functions, while interpretability is an ability to provide meaning in terms that are understandable to a human [109]. Explainable components can be included in AI by utilizing text and visual explanations (model-agnostic explanations) [109]. Further, Fabio Zanzotto [110] stated the simple idea of including human-in-the-loop (HitAI) in

which the decision power is given to the specialized professionals who utilize machines/tools as advisers which promotes interpretability. All the aforementioned requirements were specially considered in the design and development of REACT. REACT utilizes the principle of model-agnostic explanations and HitAI to support decision-making process in real-time. Thus, REACT can be considered as a step in making explainable and interpretable real-time decision support tool for instructors.

3.2.2 Cluster Analysis

The main goal of the cluster analysis or clustering is to organize a collection of data items into clusters such that the data items within a cluster are more similar to each other compared to the items in other clusters [111]. Clustering is considered as unsupervised machine learning because it doesn't have the predefined labels of data items. The process of clustering consists of four steps - Feature extraction & selection, Clustering algorithm design, Result evaluation, and Result explanation [112]. There are several taxonomies suggested by Brian Everitt [113], Rousseeuw and Kaufman [114], Jain et al. [115], and Xu et al. [116]. The simplified taxonomy is given by Jain et al. [115], which divides clustering into two groups – Partitional and Hierarchical. The Partitional clustering can be defined as a division of data objects into non-overlapping clusters such that each data object is in exactly one cluster, while Hierarchical clustering is organized as a tree which permits clusters to have sub-clusters [117]. It is noted that Hierarchical clustering provides good results for small datasets [118], as is the expected number of students in a class. It is suitable for the datasets with arbitrary shape, type and hierarchical relationships among the clusters can be easily detected using it [112]. Further, the entire clustering process can be visualized by plotting a dendrogram, which shows the cluster-subcluster relationship and the order in which they are merged [117]. This results in informative descriptions and visualization for the potential data clustering structures [112], fulfilling the goal of explainability. Thus, because of these reasons, Hierarchical clustering makes a suitable choice for clustering on REACT.

There are two basic approaches to Hierarchical clustering – Agglomerative and Divisive [119]. The Divisive method is computationally very expensive [120] and not commonly used in practice [112]. The Agglomerative hierarchical clustering is a bottom-up approach that starts with the points as individual clusters and merges the closest pair of clusters [117]. The algorithm of Agglomerative hierarchical clustering provided provided by Ryan Adams[121] is presented below. There are different proximity methods such as single linkage, complete

Algorithm 1 Agglomerative hierarchical clustering[121]

Input: Data vectors $\{x_n\}_{n=1}^N$, group-wise distance $DIST(G, G')$

$A \leftarrow \phi$ ▷ Active set starts out empty.

for $n \leftarrow 1...N$ **do** ▷ Loop over the data.

  $A \leftarrow A \cup \{\{x_n\}\}$ ▷ Add each datum as its own cluster.

end for

$\tau \leftarrow A$ ▷ Store the tree as a sequence of merges.

while $|A| > 1$ **do** ▷ Loop until the active set only has one item.

  $G_1^*, G_2^* \leftarrow \underset{G_1, G_2 \in A; G_1, G_2 \in A}{\operatorname{argmin}} DIST(G_1, G_2)$ ▷ Choose pair in A with best distance.

  $A \leftarrow (A \backslash \{G_1^*\}) \backslash \{G_2^*\}$ ▷ Remove each from active set.

  $A \leftarrow A \cup \{G_1^* \cup G_2^*\}$ ▷ Add union to active set.

  $\tau \leftarrow \tau \cup \{G_1^* \cup G_2^*\}$ ▷ Add union to tree.

end while

Return: Tree τ.

linkage, average linkage, etc. that are used in Agglomerative hierarchical clustering, but the single linkage and complete linkage are the most popular methods [122]. The formulas and descriptions of the proximity methods used in REACT are given below, which are provided by Tan et al.[117]. The following formulas are given for measuring proximity between C_1 and C_2. The observations from C_1 are represented as $X_1, X_2, ..., X_k$, and the observations

from C_2 are represented as $Y_1, Y_2, ..., Y_l$. Also, $d(x, y)$ indicates the distance between a point from a vector X and a point from vector Y.

Single linkage is the proximity between the closest two points in different clusters.

$$d_{12} = \min_{i,j} \ d(X_i, Y_j)$$

Complete linkage is the proximity between the farthest two points in different clusters.

$$d_{12} = \max_{i,j} \ d(X_i, Y_j)$$

Average linkage is the average pairwise proximities of all pairs of points from different clusters.

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} D(X_i, Y_j)$$

Ward's method minimizes the total within-cluster variance and combines the clusters based on the minimum information loss.

$$d_{12} = \sqrt{\frac{2 * |k||l|}{|k| + |l|}} * ||\bar{x} - \bar{y}||$$

It is also essential to evaluate the result to select the best possible method of clustering, and for that purpose, the Agglomerative Coefficient (AC) is useful. The AC describes the strength of the clustering structure, which is dimensionless and always lies between 0 to 1 [123]. The formula for calculating AC is given below.

$$AC = \frac{1}{n} \sum_{i=1}^{n} 1 - m(i)$$

In the above formula $m(i)$ denotes the dissimilarity of each observation to the first cluster it is merged with divided by the dissimilarity of the merger in the final step. So, AC is average of all $1 - m(i)$ values. Thus, the AC close to 1 reflects the clarity of the clustering structure i.e., higher the AC clear the clustering structure and vice versa. In this way, the four steps of the clustering process noted by Xu and Tian [116] using hierarchical clustering are implemented in REACT.

3.3 Architecture and Features of REACT

3.3.1 Architecture

REACT is a web-based interactive LAD developed using R & Shiny framework[1]. It incorporates the principles of reactive programming [124] that are suitable for interactive applications. REACT is developed based on a mantra provided by Ben Shneiderman – "Overview first, zoom and filter, then details-on-demand" [125]. REACT first gives instructors an overview; then they can zoom in/out or filter the data. REACT also allows them to reveal the details as needed using tooltips and downloads. The architecture of REACT is shown in Figure 3.1. REACT consists of five main components and are described below.

[1]https://shiny.rstudio.com

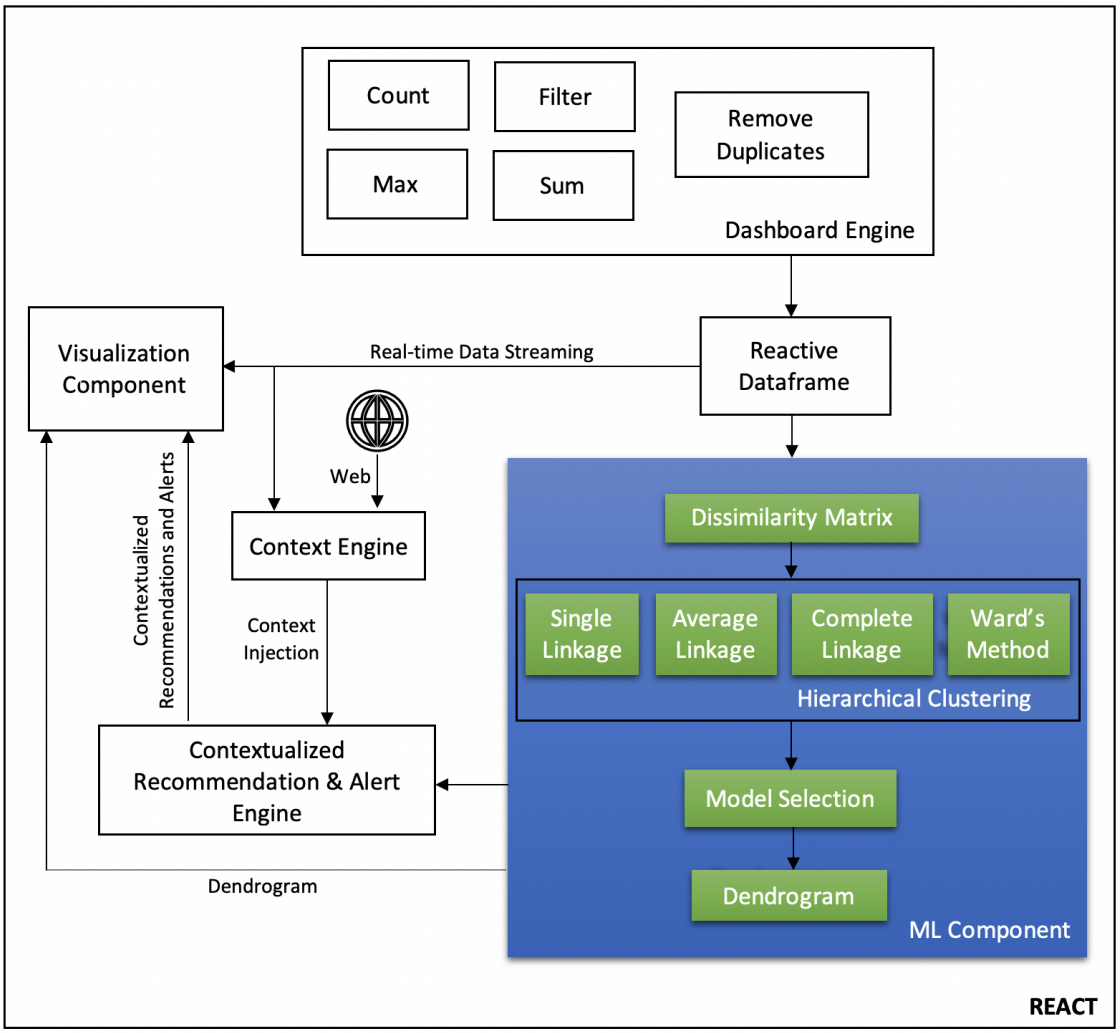


Figure 3.1: Architecture of REACT.

- The **Dashboard Engine** is responsible for periodical access of the data from the database or LMS. It also performs data cleaning and preprocessing. The output produced from the Dashboard Engine is merged and moved into a **Reactive DataFrame** [126]. The contents of this tabular data structure are updated automatically with any update of the database or LMS. This reactive DataFrame acts as an input to the AI Component, the Context Engine, and the Visualization Component.

- The **ML component** receives input from a reactive data frame that contains the input features of each student and initiates the clustering process by first calculating all the pairwise distances (i.e., dissimilarities) of students. We used the Gower distance [127] as the dissimilarity metric for the clustering. The Gower distance uses separate distance metric for both qualitative and quantitative data[128]. For qualitative data, the distance between two categories is 1 if the categories have the same value and 0 otherwise as shown below.

$$d(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ 1 & \text{if } x_i \neq x_j \end{cases}$$

For quantitative data the distance between two points x_i and x_j can be calculated using the formula given below.

$$d(x_i, x_j) = 1 - \frac{|x_i - x_j|}{abs(x_{max} - x_{min})}$$

In the above formula, R_x is the range of the vector. In this way, the distance measure is computed and the average distance is used as the overall distance. Thus, due to this main advantage, Gower distance can be applied to the mixed data (i.e., a mix of numerical and categorical variables) [123]. The details about different modules in the ML component are as follows.

  - The **Dissimilarity Matrix** sub-component calculates the pairwise distances between all n observations (i.e., students) in the data set organized in an $n \times n$ matrix, using the `daisy()` function in `R`. This dissimilarity matrix then becomes the input of the Hierarchical Clustering sub-component.
  - The **Hierarchical Clustering** sub-component trains four different hierarchical clustering models using the same dissimilarity matrix. These models are based

on four different linkage methods: Single linkage, Average linkage, Complete linkage, and Ward's method. The `R` function `agnes()` is used for building these models and computing their ACs.

- The **Model Selection** sub-component ensures robustness and acts as an internal index for evaluations. It compares the four clustering results based on their ACs. Their values lie between 0 to 1, and describe the strength of the corresponding clustering structure [129]. This sub-component selects the model with the highest AC.
- The **Dendrogram** sub-component creates a visualization of the hierarchy of clusters and sub-clusters that are the result of the selected model. This visualized hierarchy is called a *dendrogram*. The dendrogram provides a diagrammatic representation of the hierarchical cluster analysis. Dendrorgram also is an *explainable* diagram that assist in understanding how the clustering algorithm is forming clusters (i.e., groups) of learners. This approach can help to understand the clustering mechanism which may help to incorporate explainability.

- The **Context Engine** is responsible for injecting context for recommendations and alerts. The context can be derived from the student's information such as attendance, previous activity submissions, time stamp of submissions, etc. Additionally, more context about the other events such as infection rates due to COVID-19 or internet speed in surrounding counties where most students live can also be injected by fetching real-time updates from the web. This may assist instructors in deciding about deadlines for the assignments or in-class activities. This information can enable REACT to provide contextualized recommendations and alerts.

The majority of students are using hints for [Feature 1] and [Feature 2] topics. The majority of incorrect responses are observed for [Feature 3] and [Feature 4] topics.
The majority of students are using hints for Mean and Circle Graph topics. The majority of incorrect responses are observed for Venn Diagram and Mean topics.

Figure 3.2: A template (top) and an example (bottom) of a message based on the textual template-based approach. The placeholders of the template (shown in brackets) will get replaced based on the output of the AI Component, as shown in the example, in blue color.

- The **Contextualized Recommendations & Alerts** play an essential role in the efficiency, readability, and interpretability of REACT. REACT utilizes a textual template-based approach as shown in Figure 3.2 to provide interpretations in terms that are understandable to humans. In this approach the designed templates are filled with appropriate words and numbers, based on the real-time updates from the AI Component and Context Engine.
- The **Visualization Component** receives data from the reactive DataFrame and the output of the AI Component. It helps to track students' progress in real-time using interactive visualizations such as dot plot, bar plots, histogram, etc. while the dendrogram from the AI Component shows the clusters of learners in real-time.

3.3.2 Features

REACT support seven features, which are as follows.

1. **Interactive visualizations** – The visualisations included on REACT are interactive. The interactive visualizations can help educators to increase the quality and broadens the variety of angles of analysis to serve other curiosities [130]. This may enhance

educators' perception to decide what to focus on and reflect on their teaching practices. These visualizations can be downloaded in Portable Network Graphics (PNG) format, for record-keeping purposes. Educators can interact with them by zooming in, zooming out, selecting different components of the visualizations, viewing additional information using mouse hover, etc.

2. **Dynamic tables** – REACT includes a DataTable[2] which is an interactive dynamic table and update in real-time. In particular, the scorecard of the class is displayed in the form of a DataTable whose contents gets updated in real-time as new students' responses are submitted. This table can be downloaded in CSV format for record keeping purposes.

3. **Portability** – REACT is a cross-platform tool. It is configurable and portable in the sense that it can be connected to Learning Management Systems (LMSs) like Moodle, Blackboard, as well as different databases, including MySQL, Oracle, Salesforce, etc. This can be achieved by using different packages such as DBI (for databases), bRush and rcanvas (for canvas) which are available in R. Additionally, many other LMSs offer REST APIs which can be connected with REACT using httr and jsonlite packages. Further, it can also be deployed on local servers or on the cloud, such as Amazon AWS.

4. **Real-time** – Updates on REACT will get triggered by changes in the database/LMS, which will capture the students' in-class activities in real-time. A feature of pausing and resuming of the real-time streaming is also offered.

5. **Explainability and interpretability** – The visualisation component creates a dendrorgram that are *explainable* diagrams which will assist in understanding how the clustering algorithm is forming clusters of students. It can bring interpretability and can help answer questions such as: "Which students are selected in each cluster and

[2] `https://rstudio.github.io/DT/`

at which stage?", "Are there more clusters in the data?", etc. The textual template-based approach also will add *interpretability* to the tool. It give REACT an ability to provide the meaning in understandable terms to a human.

Figure 3.3: Responsive design of REACT.

6. **Data confidentiality** – REACT will be made available only to the educators. Whether REACT is hosted on a local server, Shiny server, or on the cloud, it is possible to provide user authentication. Thus, students information will only be available to their educators, adhering to the Family Educational Rights and Privacy Act (FERPA) [131].

7. **Responsive design -** REACT can be used on smartphones, tablets and laptops/desktops. It supports *responsive design*, i.e. it adapts to the user's behavior and environment based on screen size, orientation, and platform as shown in Figure 3.3.

3.4 Design and Demo

3.4.1 Design elements of REACT

REACT has a user-friendly GUI that may allow instructors to (i) instantly get an overview of the overall performance of their class in real time, (ii) gain deeper insights on the individual learners' performance, (iii) understand the circumstances that may affect the learning process, and (iv) assist in making decisions that improve their learning experience. The design of REACT is based on the design recommendations provided by Few [90] which includes three aspects - (i) incorporate "eloquence through simplicity" i.e., simple design but clearly communicates the objectives. (ii) provide an instant high-level overview of the state of things to the viewer i.e. make information available at a glance. (iii) should fit on a single computer screen. Additionally, the selection of visualizations on REACT are based on the recommendations by Abela [132] and a literature review provided by Schwendimann et al. [85]. We also have followed 'golden rules' (strive for consistency, permit easy reversal of actions, keep users in control, and reduce short-term memory load) of User Interface (UI) design [133] and Gestalt principles [103] (similarity, enclosure, closure, and connection) to improve the usability of REACT. REACT has five tabs: *Overview, Quick Analysis, Scorecard, AI* and *Public Health*, presented in Figures 3.4-3.8, respectively.

- The **Overview** tab is shown in Figure 3.4 which is the first tab instructors see when they open the application. It includes four *Key Performance Indicators* (KPIs): minimum, maximum, mean score, and the number of students who have completed the assignment thus far. It also includes an interactive dot plot for monitoring students' performance. This tab also provides alerts and recommendations to assist instructors. The alerts indicate the students who need the most attention, while the recommendations suggest specific actions, based on the used hints and the incorrect responses.

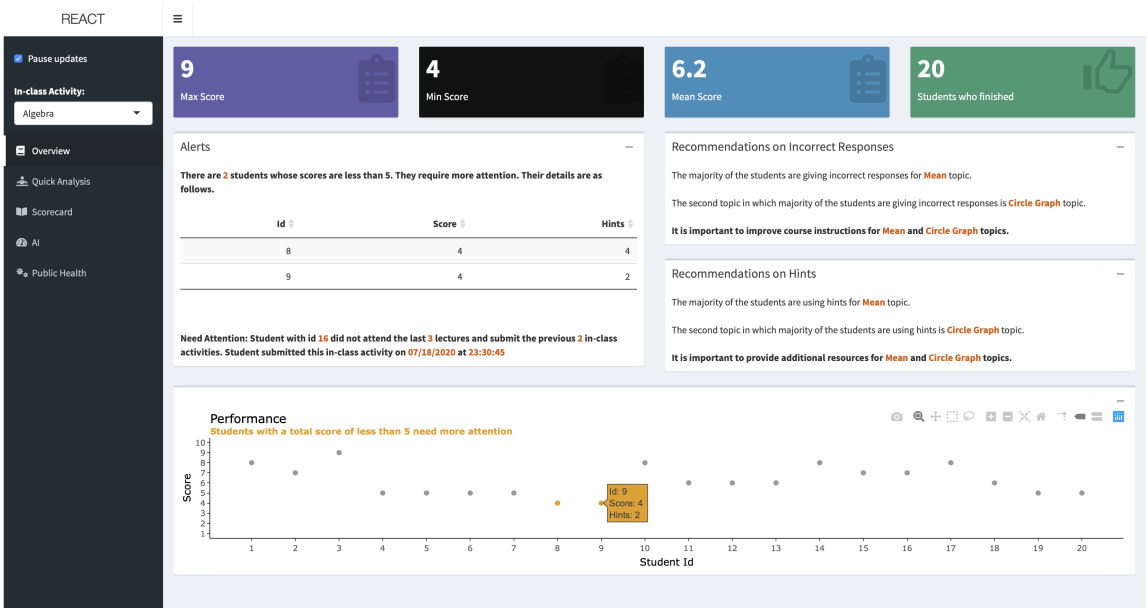


Figure 3.4: The overview tab includes Key Performance Indicators (KPIs), an interactive dot plot for understanding learners' performance, alerts, and recommendations for supporting the instructor's decision-making process.

- The **Quick Analysis** tab is shown in Figure 3.5 which displays students' responses in real-time. It also shows the four aforementioned KPIs at the top of the screen. It further includes an interactive dot plot for monitoring students' individual responses, and two bar charts that indicate the concept-wise percentage of incorrect answers and hints, respectively. On this tab, the dot plot in the top panel indicates correct responses by green color while incorrect responses by red color. The interactive legend can also be used as a filter to show only correct or incorrect responses. This tab provides the instructors with valuable insights about the whole class activity, at a glance. It is utilised for tracking students' responses in real-time and it becomes available on REACT from the beginning of the class activity. In contrast, the overview tab is most useful after the activity, while making decisions. Precise information about specific responses or concept-wise percentages is shown when hovering the cursor over a plot.

Figure 3.5: Quick Analysis tab includes KPIs, interactive dot plot, and bar charts, which can be utilized for tracking learners' responses in real-time.

- The **Scorecard** tab is shown in Figure 3.6. It gives an overview of the scores using an interactive histogram and an interactive density plot. The exact count or density can be seen by hovering the cursor over the plot. Further, this tab also includes a dynamic table that provides individual student information on the number of questions attempted, final score, total hints used, and feedback on whether the student needs attention. The contents of this table are automatically updated with changes in the database (i.e., as new students' responses are submitted) in real-time. Instructors can save the scorecard in CSV format by clicking on the download button, which is a functionality not offered by other LADs.

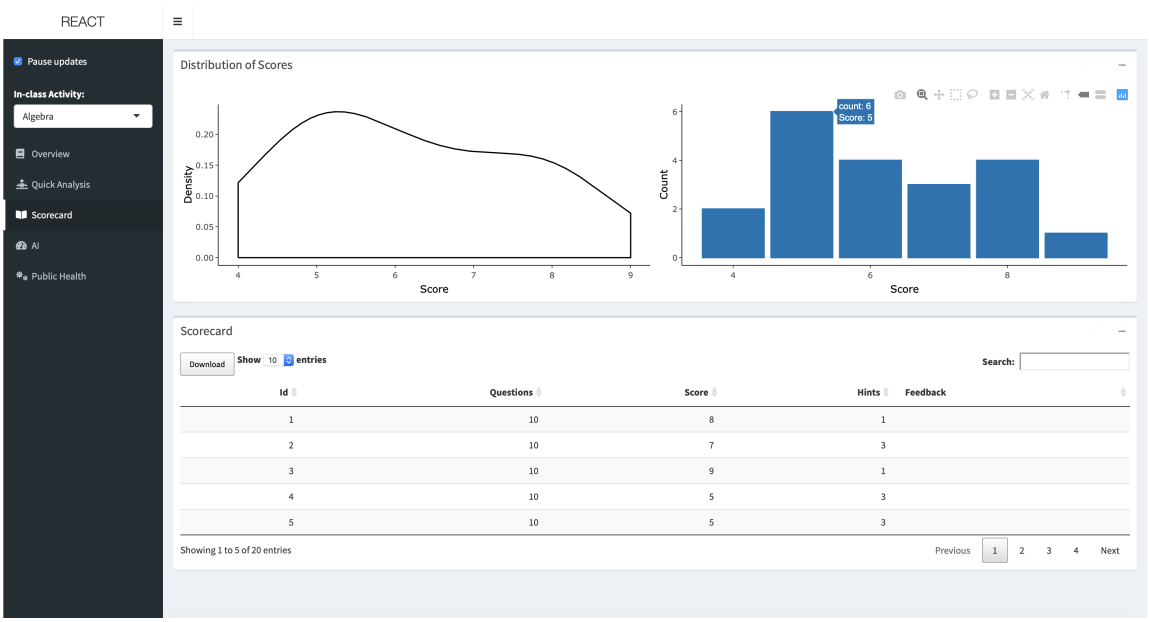


Figure 3.6: Scorecard gives an overview of the scores using interactive histogram, density plot and also includes a dynamic table that provides individual learner's information.

- The **AI** tab is shown in Figure 3.7. It distinguishes REACT from other LADs as it provides real-time insights of clustering to instructors. The objects of the cluster analysis are the students in the class. The features used by the clustering algorithm are the responses for each question and hints used in each question. The purpose of this analysis is to form groups (i.e., clusters) of students with similar performance. The rationale for identifying these groups is to give the instructor insights on how to, for instance, form study groups, recommend specific additional study materials to clusters of students, or have them form groups for in-class discussions focusing on those concepts that they need more help with. Note that the result of hierarchical clustering is a hierarchy of clusters and sub-clusters. This hierarchy is depicted on the *dendrogram*. Choosing an appropriate distance threshold as a cut-off point results in a specific set of non-overlapping clusters. Thus, the dendrogram might help to incorporate the transparency and explainability of the clustering process. In the use

case of Figure 3.7, it shows three distinct clusters in the data which are enclosed in the three boxes. This helps to understand which student is member of which cluster and how this membership was formed. REACT also provides the average scores and hints used for every cluster along with cluster interpretations. This helps the instructor understand what are the main characteristics of each cluster, such as "high-performing students", or "students who struggle with a specific concept", etc. Additionally, there are recommendations for every cluster based on performance. All these insights might make REACT easy to interpret. Instructors can also see concept-wise details concerning hints and responses for each cluster, using the interactive dot plots provided in the bottom panel.

- The **Public Health** tab is a unique feature of this tool and shown in Figure 3.8. It has been introduced due to the current COVID-19 pandemic. It includes a bar plot that indicates the infection rate in the surrounding counties from where students travel the most. Using this information, REACT provides information about the average infection rate in the area and useful tips concerning the current situation. The infection data gets updated from CovidActNow[3].

[3] `https://covidactnow.org`

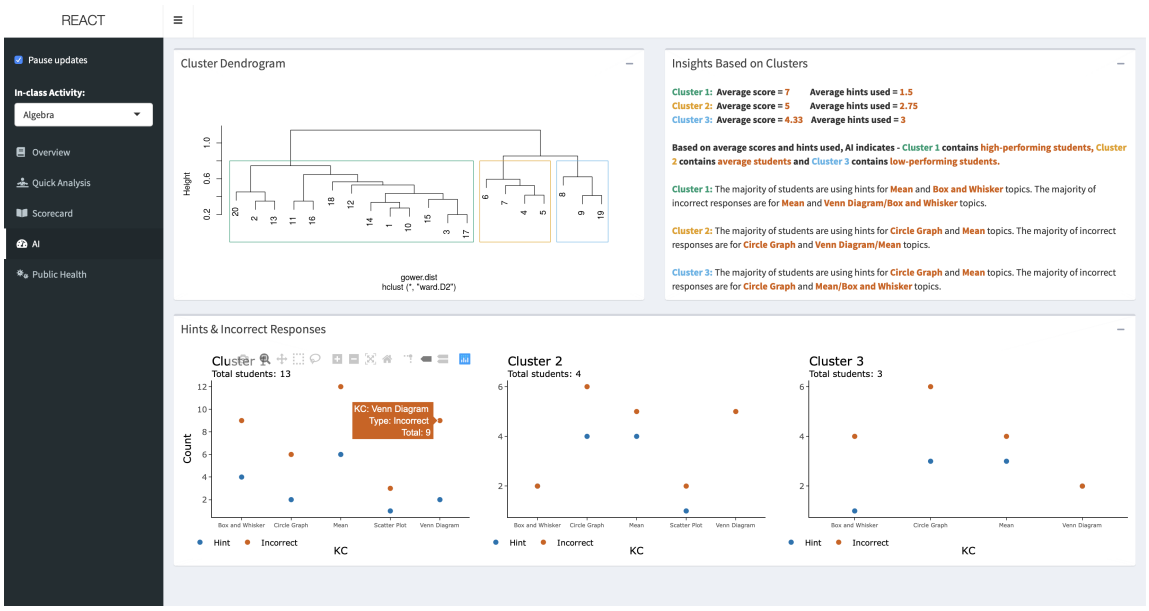


Figure 3.7: AI tab provides real-time insights of clustering to instructors with dendrogram and textual-template based recommendations. Dendrogram may help to incorporate transparency and explainability, while easy to read insights on clusters may provide interpretability.

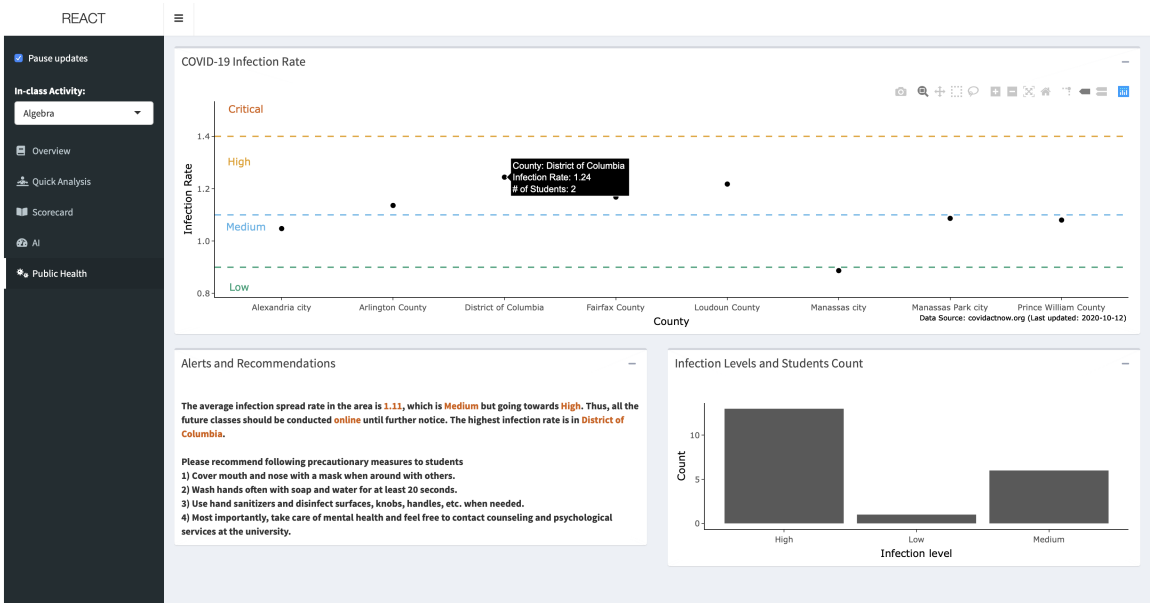


Figure 3.8: Public Health tab provides information on the current COVID-19 infection rate in the surrounding counties.

3.4.2 Demo

Holstein et al. [38] noted the importance of using real-world datasets to understand the behavior of LA tools. Thus, we used the 2009-2010 Skill-builder ASSISTments dataset [134] for demonstration and evaluation of REACT. The raw data consists of more than 100,000 rows representing details of 4217 students and 111 Knowledge Components (KCs). To achieve the objective of demonstration, we randomly selected a sample of 20 students. Due to privacy, this data set includes only pseudo-ids. In real-word application of REACT, authenticated instructors will be able to see students' names, as memorising their ids would be troublesome. Our approach to create a demonstration of REACT is shown in Figure 3.9 and can be summarized in the following steps:

Step 1 (Filter): We selected 20 students and two questions from five KCs from the topic of Algebra (Mean, Circle Graph, Venn Diagram, Box and Whisker Plot, and Scatter Plot). This filtered data set is first stored in a spreadsheet on a local hard disk.

Step 2 (Stream): The filtered data from *Step 1* are then streamed on a Google sheet that acts as a database for this demonstration. It is connected to REACT using the `googlesheets4`[4] package.

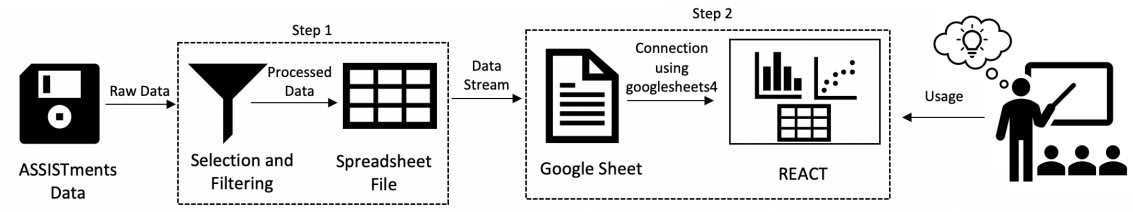


Figure 3.9: The procedure used for creating a real-time demo of REACT [4].

[4] `https://googlesheets4.tidyverse.org`

In this way, these two steps are used for replicating a use-case scenario of REACT in classrooms. This demonstration also shows how the live updates are processed on the fly and used to update the visualizations, alerts, and recommendations displayed on the user interface. A live version of REACT[5] is deployed on `Shiny Server` which can be accessed using a web browser on any desktop, laptop, tablet, or smartphone. Further, this deployed version of REACT along with the recorded demonstration[6] are used for understanding educators perceptions on impact and usability of REACT.

3.5 Discussion

REACT's main objective is to provide insights about learners at a glance to help instructors make decisions. The **Overview tab** provides highlights of the class using KPIs and interactive plots. From this tab, it can be observed which students require more attention, and where exactly students are struggling. In dashboards, KPIs play an important role because, at a glance, they can help decision-makers to understand the performance or the deviation from the set target [135]. Further, the recommendations provided based on hints and incorrect responses can help the instructor understand more about the class. Overall, the recommendations shown in Figure 3.4 make one thing clear; two topics (i.e., *Mean* and *Circle Graph*) need major attention. REACT recommends providing additional materials on these topics to the students. The performance plot, which is at the bottom panel, highlights the students with poor performance in orange color. In this way, this tab helps in interpreting the class requirements by reflecting the teaching practices and needs of the students for making decisions.

The **Quick Analysis tab** shown in Figure 3.5 assists in tracking learners' responses in real-time. The bar plots at the bottom panel show the percentage of incorrect responses or hints used per concept. Based on the bar plot, it can be easily noted that the highest number of incorrect responses were given for questions related to the Mean (28%) and the

[5] `https://tinyurl.com/y7cbbbej`
[6] `https://www.youtube.com/watch?v=gEBuqvgEsqM`

lowest for the Scatter Plot (6.67%). Further, the students have used the largest number of hints for the Mean (37.14%) and the lowest for the Scatter Plot (5.71%). The KPIs that are presented at the top get updated in near real-time. This can help instructors estimate the pace of the class.

The **Scorecard tab** is useful for analyzing the distribution of the score of the class using interactive histogram & density plots. For example, Figure 3.6 indicates that four students scored 6 points, while most of the students scored 5 points. Additionally, instructors can download the plots in PNG format and the scorecard in CSV format, which can be very helpful in keeping records of the learners.

On the **AI tab** as shown in Figure 3.7, we can observe three main clusters on the dendrogram that are enclosed in yellow, blue, and green boxes. This helps the instructors to understand which student is member of which cluster, and how this membership was formed. The insights that are provided about the cluster analysis indicate what are the average scores and the hints used in each clusters. These insights may assist in the interpretation of each cluster. In this use case, the clusters represent high-performing students, average students, and low-performing students. This interpretation may enable instructors to understand the decisions made by the AI Component in non-technical terms. Further, the concept-wise details for every cluster are included in textual and visual formats as well. These additional details may help instructors while forming the study groups or groups for class activities. These results may also help instructors in self-reflection and sense-making of their teaching practices.

The context of instructors' assistance tools may not be limited to variables corresponding to quiz responses and hints; it could be potentially unbounded [136]. Context plays an important role as it adds more relevance in the instructors' decision-making process [137]. It is also possible to use the contextual variables within a dataset as features [138]. The context that we used as an experiment in REACT includes the students' attendance, previous activity submissions, timestamps of submissions, and residential areas. In the use

case, an alert for student #16 is provided on the Overview tab, based on these parameters, informing the instructor that the student was absent for three lectures and did not participate in the past two activities. Furthermore, on the **Public Health tab** shown in Figure 3.8, the residential areas of learners are used as context for showing important information about COVID-19 exposure risk. This can be a valuable input for decision-making on the format of the class, such as temporarily switching to online mode. Therefore, by considering all these elements, REACT may be described as a real-time decision support tool that incorporates explainability, interpretability, along with portability for showing different indicators of students, learning processes, and recommendations which may increase efficiency in decision-making.

Chapter 4: Understanding Educators Perceptions on Experience and Usability of REACT

4.1 Introduction

The previous chapter discussed REACT's objectives, its utilization and focused on the design and development aspects. This chapter presents a next step that emphasizes experiment design for evaluations and insights obtained from these evaluations. The use of Artificial Intelligence (AI) in education (AIED) brings innovation in instructional design, technology development, and educational research beyond traditional educational modes [7, 139]. AI shows better performance in computing [140] and also enhances human productivity in educational settings by facilitating teaching, learning, and decision-making [7, 141]. Recently, Zhang and Aslan [17] explored six applications of AIED – chatbot, expert systems, intelligent tutors or agents, machine learning, personalized learning systems, and visualizations – and noted that it is essential to seek inputs from the educational communities/educators for making advances in AIED. Additionally, a need for smart learning analytics tools has been explored, which can help educators to improve and adapt their teaching by monitoring, understanding student's progress, and identifying students who are struggling in a particular topic [141]. One of the possible ways to achieve these objectives is by integrating two applications of AIED - visualizations and Machine Learning (ML) – on a Learning Analytics Dashboard (LAD). Chen et al. [142] noted that visualizations are one of the most direct and effective ways to transfer and process knowledge by gaining insights into complex data. Furthermore, cluster analysis – an unsupervised machine learning technique - can help to find clusters of students in a classroom based on various student characteristics

(learning style preferences, academic performance, behavioral interaction, etc.) that identify collaborative learning opportunities and at-risk students at an early stage [18]. In this way, visualizations and Machine Learning (ML) can help achieve a better understanding of teaching and learning.

To achieve this goal it is imperative that dashboards are evaluated comprehensively using a human-centered process. Recent work though shows that evaluations are usually limited to usability of a LAD [85, 143–145] and many LADs are not even evaluated [145], making it an under-explored research area [107]. To bring the promise of LADs to fore, it is imperative that they are evaluated primarily on their ability to fulfil the goal of providing an understanding of teaching and learning and then their usability [145]. With regards to AIED applications, which are mainly concerned with supporting human learning and teaching [146], this implies that it is important to conduct evaluations from the perspective of interpretability, and explainability [147]. In other words, the evaluation approach has to human-centered. The human-centered AI is defined as "a perspective on AI and ML that intelligent systems must be designed with an awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity" [148]. It combines visualizations and Human-Computer Interaction (HCI) to enable experts to solve complex analysis tasks and brings attention to human factors like trust as well as efforts during evaluations [149]. Further, Doshi-Velez and Kim [108] recommends using an application-grounded approach of evaluation with domain experts because it is the most realistic way of evaluation to understand a tool's effect, impact, and usability from the stakeholder's role.

Therefore, by considering these factors REACT is evaluated using an application-grounded approach with domain experts. Thus, this chapter answers two question: *i) how can a visual LA tool - REACT (Real-time Educational AI-powered Classroom Tool) - that incorporates Visual Analytics (VA) and AI be helpful in classrooms?*, and *ii) how do educators from different domains perceive the integration of VA and AI in real-time on REACT?*

4.2 Background

4.2.1 Prototypes

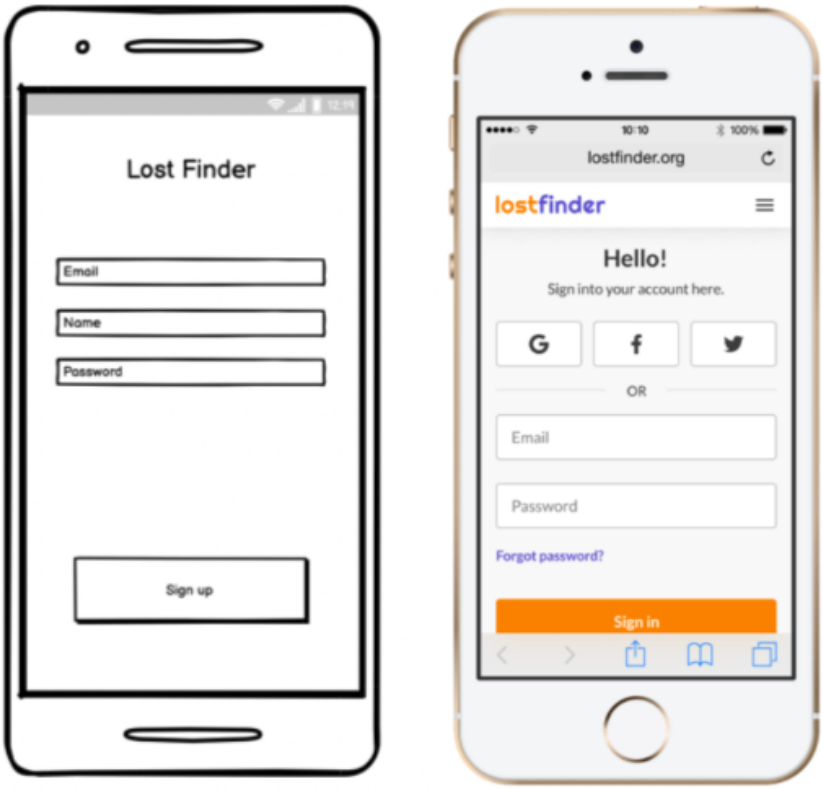


Figure 4.1: An example of low-fidelity (left) and high-fidelity (right) prototypes of an application (Source: `https://tinyurl.com/2whz42mp`).

Prototypes are incomplete designs of a tool or product which are cheap but fast to develop for experimentation purposes [150]. Design and development of a prototype is one of the effective ways to communicate, discuss and evaluate ideas with stakeholders [45]. It is also an integral part of iterative user-centered design approach [150]. Prototyping can be defined as "an activity with the purpose of creating a manifestation that, in its simplest form, filters the qualities in which designers are interested, without distorting the understanding

of the whole" [151]. Prototyping can be done by considering two different philosophies – evolutionary and throwaway [45]. Evolutionary prototyping utilizes engineering principals for making a final product while throwaway prototyping uses prototypes as iterative process which leads to the final design [45]. Further, prototypes broadly classified as low-fidelity and high-fidelity as shown in Figure 4.1. High-fidelity prototype looks like a final product which helps to get valuable feedback in real contexts from the stakeholders [152]. High-fidelity prototypes can be used for exploration and test because they are fully interactive, user-driven, and includes almost complete functionalities which can't be done with low-fidelity prototyping [152]. For High-fidelity vertical or horizontal prototyping can be done. Vertical prototyping provides complete details only for a few functions while horizontal prototyping includes a wide range functionality for only a few functions [45]. Additionally, prototypes can help users to get a better impression of the user experience compared to textual descriptions. Therefore, by considering above reasons a high-fidelity prototype of REACT using horizontal prototyping is used for this study.

4.2.2 Usability and User Experience

The principal approach for understanding the quality of interaction in the field of Human-Computer Interaction (HCI) is through the concept of usability [153]. Usability can be defined as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [154]. Overall usability covers six goals - effectiveness, efficiency, safety, utility, learnability & memorability [45]. Thus, it is important to ensure that the designed tool is fulfilling these goals and satisfactory to users via usability testing [155]. Mortem Hertzum [156] explored six images of usability – (i) Universal usability (embracing the challenge of making products for everybody to use), (ii) Situational usability (the quality-in-use of a product in a specified situation with its users, tasks, and context of use), (iii) Perceived usability (the user's subjective experience of a product based on his or her interaction with it), (iv) Hedonic usability (joy of use rather than ease of use, task accomplishment, and freedom

of discomfort), (v) Organizational usability (group of people collaborating in an organizational setting), and (vi) Cultural usability (takes on different meanings depending on the users' cultural background) - and advised to understand usability by applying an alternative usability image to challenge the dominate image. There are three widely used methods to preform usability testing – Think Aloud (TA), Heuristic Evaluation (HE) and Cognitive Walkthrough (CW) [155]. TA method is a standard method in which participants are asked to "think aloud" about their experience using the tool while an evaluator observes the users and listens their thoughts [157]. Additionally, John Brooke [43] proposed a quick way to measure usability based on the System Usability Scale (SUS) which can be interpreted based on the scale provided by Bangor et al. [158]. SUS is a questionnaire-based approach that can help to evaluate a tool for ensuring the satisfaction of goals of usability defined by [45]. The other important consideration while performing usability testing is number of participants. Dumas and Redish [159] suggested that 5 to 12 participants are acceptable numbers for usability testing. At the same time, Nielsen and Landauer [160] showed that five users could find as many usability problems as possible that can be found with more participants. While Caine [161] advised conducting usability testing with 15 participants. Thus, by considering all these factors, we performed evaluations with 33 participants using a mixed TA method with System Usability Scale (SUS). In the SUS questionnaire, the 5-point Likert response format from "Strongly agree" to "Strongly disagree" is used for collecting responses. Further, during TA studies, Universal, Situational, and Perceived usability are explored by explaining a situation, context and assigning tasks to different domain experts while interacting with REACT.

Along with usability, it is also crucial to understand factors such as users' feelings, their motivation, expectations, satisfaction when using, looking, and opening or closing a tool [45, 162]. This can be achieved by performing User Experience (UX) testing which is defined as "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" [154]. UX testing is a key factor for understanding the quality of a service or a product [163]. Usually, UX goals can be broadly classified as desirable

and undesirable [45]. The desirable goal includes properties such as satisfying, helpful, motivation, cognitive stimulating etc. while undesirable goal consist of properties such boring, unpleasant, frustrating, annoying etc. Further, Marc Hassenzahl [164] suggested to consider pragmatic and hedonic aspects while performing UX testing. Pragmatic aspect deals with how simple, practical and obvious it for the user to achieve their goals while hedonic aspect deals with how evocative and stimulating the interaction is to the users [45]. UX testing can be conducted using Surveys, Expert Evaluation and Mixed methods [165]. Surveys are cheap and less time-consuming that can help to get feedback from users in a short amount of time. On the other hand, expert evaluations are expensive and time consuming which provides subjective insights of experts based on their expertise. For this study we have used an approach of conducting survey from experts. Further we evaluated REACT using a framework [166] based on Kirkpatrick's four level evaluation model [42]. This framework includes four criteria – 1) Reaction (goal orientation, information usefulness, visual effectiveness, user-friendliness, and appropriation of visual representation), 2) Learning (understanding and reflection), 3) Behavior (increase in motivation and change in behavior), and 4) Result (performance improvement and competency development). Additionally, a fifth criterion which deals with understanding interpretability and explainibilty is developed due to use of unsupervised learning (clustering). For understanding users' attitudes on these five criteria 5-point Likert response format from "Strongly agree" to "Strongly disagree" is used.

4.2.3 Likert Scale

A Likert scale can be defined as "a set of statements (items) offered for a real or hypothetical situation under study" and are used to rate the degree to which respondents agree or disagree with a statement [167, 168]. Likert scale is developed by Rensis Likert which typically is a 5 or 7 point ordinal scale [167]. In Human Computer Interaction (HCI) Likert scales are commonly used to determine a person's attitude or opinion on a topic [169]. The studies [168, 170–173] suggested to use four to six points response format along with a neutral midpoint for maintaining the reliability of a Likert scale. Further scores from single

items are less valid, less accurate and less reliable [174, 175] while summated scores i.e., composite scores calculated from the multiple items for a participant on a Likert-type scale are more reliable [49,50]. Therefore, it is recommended to use composites scores for including complete and a reliable representation of a participant's attitude [175, 176].

Further, multi-item scales are "suitable for measuring latent characteristics with many facets" [177]. Additionally, the Likert items within a Likert scale should be clear, concise and measure the same idea [178,179]. Thus, it is utmost important to ensure the reliability of the scale i.e., a scale should provide repeatable results from the same participants [169] which in a broad sense refers to the consistency or precision of measurements [180]. There are many methods for measuring reliability such as consistency over time, test-retest reliability, coefficient of equivalence etc. but the most common method is Cronbach's alpha [180]. Cronbach's alpha is a number between 0 and 1. It describes inter-relatedness of the items within an instrument which is the extent to which all the items in an instrument measures the same concept [181]. A useful rule of thumb is - Cronbach's alpha should be at least 0.7 or higher to ensure the reliability [182]. For this study we have followed the above guidelines by keeping a neutral midpoint with five-point response format, using composite scores for analyzing data and measuring the reliability of the instrument using Cronbach's alpha.

4.3 Experiment Design

4.3.1 Participant Selection

After ethics review board approval, this study was conducted in online settings with thirty-three (N = 33) educators from 9 public universities and one community college. Initially, we used convenience sampling to solicit educators (n = 5) teaching a class/classes in the Spring 2021 semester. We contacted these educators via the campus e-mail system. We identified additional educators (n = 28) through snowball sampling. Most of the educators (29 out of 33) hold Ph.D. and based on their expertise, were divided into three domains – Engineering, Science, and Humanities & Social Sciences. The Engineering domain included

10 educators (M = 8, F = 2), Science domain included 12 educators (M = 10, F = 2), and Humanities & Social Sciences included 11 educators (M = 4, F = 7) in this study. Overall fewer female domain experts participated in this study compared to male domain experts, which matches the observation noted by Kelly Caine [161]. Further, complete representation of the educators via our sample is impossible, given the nature of convenience and snowball sampling [48], an effort was made to find a diverse group of educators, from various universities and disciplines, with varying academic experience as well as ethnicity.

4.3.2 Procedure

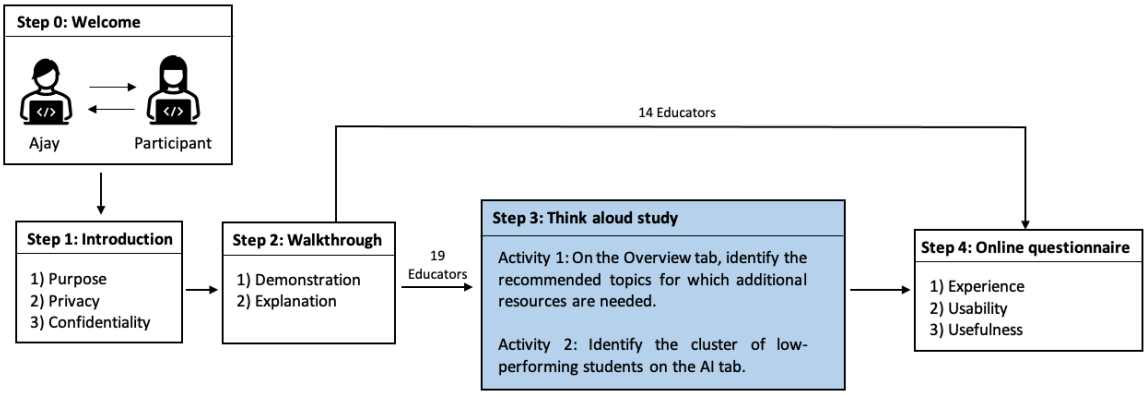


Figure 4.2: The summarised study procedure that was used to understand educators' perceptions on experience and usability of REACT.

This main objective of this study was to understand educators perceptions on experience and usability of REACT. This objective is achieved by using a hybrid approach of think-aloud study and questionnaire as shown in Figure 4.2. Due to the current pandemic, it was not possible to meet in a lab physically. Therefore, this study was conducted meeting virtually in a private meeting room using Zoom. At the start of each meeting, the purpose of the study, privacy, and confidentiality about the collected data was explained to the

participants. In the next step, we introduced REACT to participants using the walkthrough method, i.e., we explained and demonstrated different options and features they can use. The purpose behind the walkthrough is for participants to get comfortable with the tool. After that, a scenario was explained in which we requested to assume they are teaching a class of Algebra and they have initiated an in-class activity. Next, we showed a short simulation video[1] of REACT for which we used publicly available 2009-2010 Skill-builder ASSISTments data [134]. The motivation behind the simulation was to show participants a demonstration of a practical use case scenario for the evaluation. Next, we provided access to a web-based version of REACT and asked participants to freely interact with it. We also requested them to share their thought with us while interacting and analyzing information on REACT. Before ending the session, we requested participants to perform the following two activities.

Activity 1: On the Overview tab identify the recommended topics for which additional resources are needed.

Activity 2: Identify the cluster of low-performing students on the AI tab.

In the last step, we provided a link to an online questionnaire hosted on a server using Qualtrics. In the end, the participants thanked for their participation and asked if they have any questions. The details on comments and the overall interactions of the participants are noted in a logbook.

4.3.3 Instrument

The survey instrument used for this study is motivated by Yoo et al. [166] and John Brooke [43]. Yoo et al. [166] provided a framework for the assessment of visual educational tools based on four criteria: Reaction (goal orientation, information usefulness, visual effectiveness, user-friendliness, and appropriation of visual representation), Learning (understanding and reflection), Behavior (increase in motivation and change in behavior), and Result (performance improvement and competency development). Effectiveness, the fifth

[1] `https://www.youtube.com/watch?v=gEBuqvgEsqM`

criterion, is developed to get insights on the interpretability, and explainability of REACT. Further, the System Usability Scale (SUS) proposed by John Brooke [43] is also combined with these five criteria. To understand more about the usefulness of REACT, we also asked participants to give points (out of 5) to every tab based on their perception of utilization. In the end, we included three open-ended questions to know – which other features participants would like to see on REACT, which features they don't want to see on REACT and any additional comments about REACT. In this way total of 51 questions are used for evaluations.

4.3.4 Limitations

There are several limitations to the research design that impact the study conclusions. First, the number of participants from different domains might not be sufficient to gain in-depth and comprehensive evaluations. However, there are lessons to be learned from this study, highlighted in the Discussion section of this chapter. Second, results from this study represent experiences and perspectives of educators at one moment in time. We acknowledge that perceptions can change over time. Third, we used real-world data but created a fictitious scenario for the evaluation of REACT. Further, we used COVID-19 infection rates as an example context on REACT. To generalize our findings, we should test different scenarios from different domains and with different contexts. However, despite these limitations, we believe that this study is an interesting first step to understand the usability of model-agnostic explanations and the integration of context to support educators' decision-making process. The results from this study may raise the awareness of utilizing AI and context on a real-time tool.

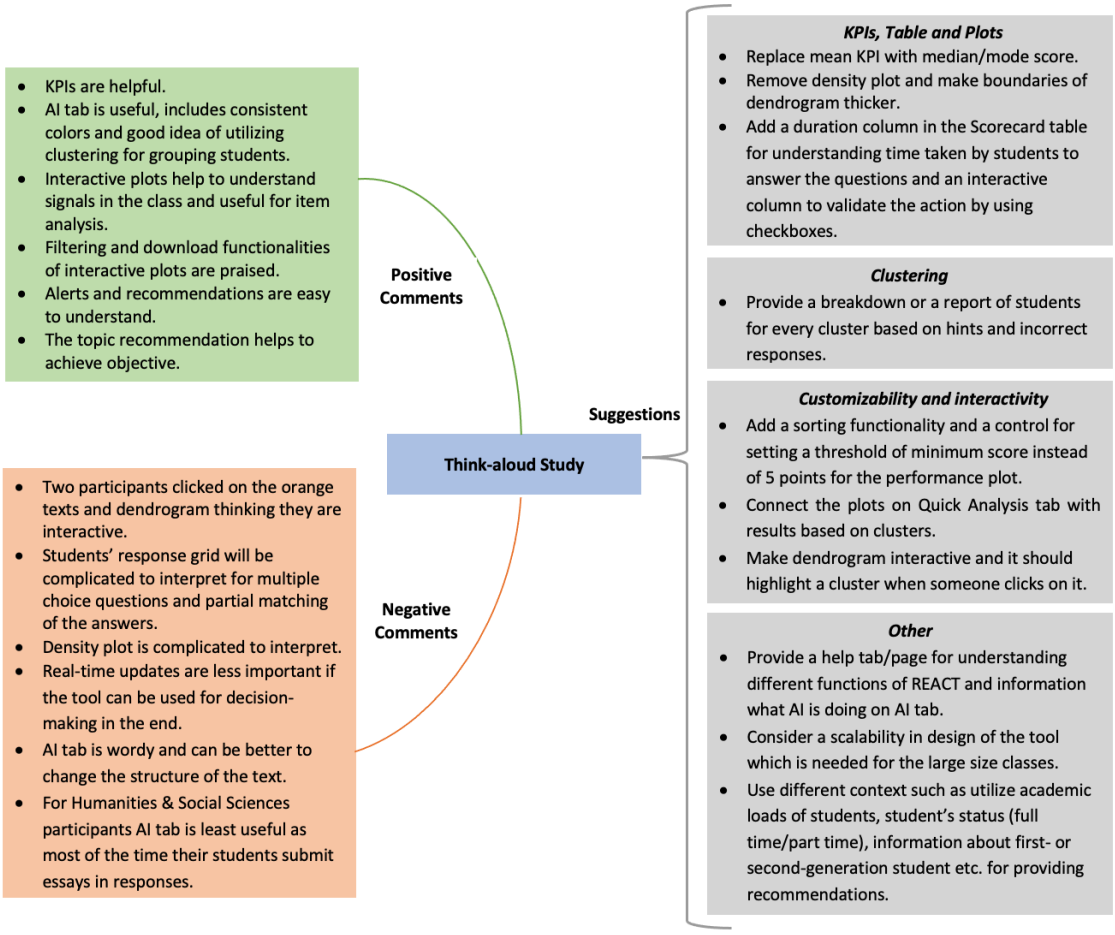


Figure 4.3: Summarised comments from think-aloud experiment.

4.4 Data Analysis

The assessment instrument (questionnaire) used for this study is developed by combining 6 Likert scales. Thus, it is essential to measure the reliability of collected data for understanding the consistency or precision of measurements using Cronbach's alpha [180]. Based on the rule of thumb provided by Fraenkel & Wallen [182], the assessment instrument used for this study satisfied the requirements for reliability. The calculated Cronbach's Alpha values for Reaction, Learning, Behavior, Result, Effectiveness, and SUS are 0.9, 0.7, 0.8, 0.8, 0.8, and 0.9, respectively. In the following sub-sections, results from the think-aloud studies and the questionnaire data are summarized. The summarized results of the think-aloud studies are also visualized and shown in Figure 4.3.

4.4.1 Qualitative Data Analysis

This sub-section summaries the results from the think-aloud studies (n = 19). The total time allocated to finish the study (think-aloud & questionnaire) was one hour. However, during the think-aloud study, we observed that the comments given by the participants started repeating after the 15th interview. Thus, we decided to stop conducting think-aloud studies after nineteen interviews. Overall, the think-aloud study was conducted with eight educators from the Science domain, four educators from the Engineering domain, and seven educators from the Humanities & Social Sciences domain. For the rest of the participants, i.e., 14 educators, the study was conducted with the same format, but we only asked them to submit questionnaire responses. The summary of the results conducted from this study is as follows.

General comments and activity patterns

In the think-aloud study, 14 participants out of 19 performed both the activities successfully, while five participants answered the second activity correctly but incorrectly answered the first activity. In general, all the participants took more time to answer the first activity, and participants who answered incorrectly also got confused about finding the correct

recommended topics. During this study, almost all the participants analyzed REACT by moving from the Quick Analysis tab to the Public Health tab. However, two participants first went on the AI tab, read all the details, and then started from the Quick Analysis tab. Further, six participants viewed the Public Health tab but didn't comment anything about it. The other interesting observation during this study was that only two participants clicked on the checkbox for pausing updates. Also, only one participant asked the purpose of the dropdown menu for selecting in-class activities. All the other participants didn't make any comments about those aspects of REACT. Almost all the participants first hover over the text, read recommendations based on incorrect responses and hints. After that, the participants checked the Performance plot, and then they checked the alerts. The Key Performance Indicators (KPIs) were observed more on the Quick Analysis tab than the Overview tab.

After getting access to REACT, one participant from Engineering domain compared it with iClicker and said, "it provides better interactive functionality," while in the other interview, one participant from the Humanities & Social Sciences immediately said, "it mimics blackboard". Two participants from Humanities & Social Sciences mentioned, "this tool is useful for formative assessment compared to summative assessment." One of them also asked a question – "how can this be used for open-ended questions?". In contrast, one other participant from Humanities & Social Sciences, said, "it is good for linguistic courses but less useful for classes like English composition". Another participant from Humanities & Social Sciences asked a question, "how to interpret count and score?" by hovering over the bar plot on the Scorecard tab. In one interview, a participant from the Science domain asked three questions – "what if there are no hints for the activity or assignment? bins are they automatically chosen? 3 is a good number, but what about 2 clusters?". Additionally, one participant from the Engineering domain asked, "what if points are not equally weighted? such as different questions weigh different points" while interacting with the performance plot. At the end of the study, most (n = 8) participants appreciated introducing context in a tool and provided some suggestions to use different contexts.

Positive Responses

KPIs - All the participants found KPIs very useful, and one of the participants from Humanities and Social Sciences mentioned, "the minimum KPI is useful for identifying weaker students".

Interactivity and Plots - The usability experiment indicated a positive response for interactive visualizations. All the participants appreciated the interactivity of the tool, especially the filtering and download features. Most of the participants interacted with the Student Responses grid on the Quick Analysis tab and filtered responses using the interactive legend. None of the participants tried the feature of downloading the plots or table, but all of them appreciated having this functionality. One of the Humanities and Social Sciences participants found the performance plot especially useful for identifying students' mastery of a topic and the Student Responses grid for performing item analysis.

AI and Model-agnostic Explanations - Most of the participants found the AI tab helpful and liked the utilization of clustering for grouping students. All the Participants appreciated the consistency of colors and the utilization of text for providing information about the dendrogram. All the participants but especially participants from the Humanities & Social Sciences found the text information on the AI tab easy to read and understand. After looking at the AI tab, one participant from the Humanities & Social Sciences commented, "It can also be useful for K-12 classes".

Textual Recommendations and Alerts – Most of the participants found that the recommendations on the Overview tab can help them achieve their objectives, and the alerts make it easier to identify red flags about students.

Negative Responses

Interactivity and Plots - One participant from the Humanities and Social Sciences found it challenging to interpret the quick analysis tab and mentioned, "it is hard

[to interpret] without training". Overall, the participant was slightly uncomfortable while using the tool due to a lack of technical expertise. Another participant from the same domain was a qualitative scholar, and she didn't find the Student Responses grid useful compared to other plots. One participant from the Science domain got confused while exploring the Scorecard tab as he was searching for an option to show all the records, but then he found it. The most interesting observation was that on the scorecard tab density plot was difficult to understand for most participants, and almost all the participants from Humanities & Social Sciences found it complicated to interpret.

AI and Model-agnostic Explanations - One participant from the Engineering domain didn't find the bottom panel of the AI tab (hints and incorrect responses for every cluster) helpful. In comparison, one participant from the Humanities & Social Sciences domain found it useful but difficult to interpret. A participant from Humanities and Social Sciences found the AI tab least useful as most of the time their students submit essays in responses. Two participants clicked on the dendrogram, thinking it is interactive, but later they understood it is not. One participant from the Engineering domain found that reading the details about every cluster is time-consuming in text format and suggested using other techniques to provide details on clusters. One participant from the Science domain found real-time updates are less important if the tool can be used for decision-making in the end. Two participants from the Science domain informed that the information on the AI tab is wordy and can be better to change the text structure.

Textual Recommendations and Alerts - One participant from the Engineering domain found texts on the Overview tab confusing, especially the orange parts he thought were clickable. Two participants thought the orange texts, which highlights topics in alerts and recommendations, are clickable. They clicked their multiple times to see if it shows anything.

Suggestions

KPIs

a. Replace mean KPI with the median/mode score KPI as it is more robust to outliers.

Plots

a. Use counts instead of percentages in the bar plot on the quick analysis tab for ease of understanding.

b. Remove density plot or use it with histogram for increasing usability and readability.

Customizability and Interactivity

a. Add a sorting functionality and control for setting a minimum score threshold instead of 5 points for the performance plot.

b. Change default hover to a finger or a hand icon.

c. Make the dendrogram interactive, and it should highlight a cluster when someone clicks on it.

d. Connect the plots on the Quick Analysis tab with results based on clusters.

e. Provide control to move students' data which can help do custom grouping on the Quick Analysis tab.

f. Add a duration column in the Scorecard table to understand the time taken by students to answer the questions and an interactive column to validate the action by using checkboxes.

AI and Model-agnostic Explanations

a. Provide a breakdown or a report of students for every cluster based on hints and incorrect responses.

b. One participant from the Humanities & Social Sciences suggested including additional information like English-speaking students, non-English speakers, or disabled students in the clustering process.

c. One participant from the Humanities and Social Science indicated caution while interpreting low-performance cluster and suggested replacing "low-performing students" with "making progress towards mastery".

d. Make the boundaries of the dendrogram thicker to increase its visibility.

Textual Recommendations and Alerts

a. Change the text in alerts and recommendations to "The top 2 categories or concepts are ...." which can help to add more information about topics if needed.

Other Suggestions

a. A detailed report on students making more mistakes may help in course improvement and remove complicated questions.

b. It can be a good idea to provide a help tab/page for understanding different functions of REACT and provide information on what AI is doing on the AI tab.

c. Consider the scalability aspect in the design, which is needed for the large size classes.

d. A participant from Humanities & Social Sciences suggested adding reflective questions such as what it tells me about instructions or lessons I teach? Or what to do differently?

e. Some suggestions about context are noted, such as utilizing academic loads of students, student's status (full time/part time) for providing recommendations, and a menu for selecting different contexts. One participant also suggested providing alerts and other information published by the university as context. While another participant suggested using information about the first or second-generation student as context.

	Presentation	Deision-making	Personalization
Science	Educators positively responded and appreciated the interactive functionalities such as filter, download, and tooltips.	Educators found information on REACT clear and concise to understand for making decisions.	Educators suggested several customizations such as - i) setting a manual threshold for a minimum score, ii) presenting responses from randomized questions, iii) an option for setting bins for histogram, iv) representing scores with different weights.
Engineering	Educators positively responded but expected an interactive dendrogram and noted that real-time feature might not be useful.	Educators liked the idea of utilizing clustering on the AI tab and appreciated the text-based approach for providing insights from clustering.	Educators suggested to add - i)more interactivity by connecting the dendrogram with students' responses grid, ii) an option to see more clusters instead of default clusters.
Humanities & Social Sciences	Educators were uncomfortable interpreting the plots, especially the density plot, and asked questions such as – i) "how to interpret count and scores?" ii) "how can this be used for open-ended questions?"	Educators liked the text-based information but found REACT least useful as most students submit essay responses.	Educators suggested to – i) add more text, ii)remove density plot, iii) provide insights for the summative assessment questions, which includes essay-based responses, iv) add reflective questions for educators for their sensemaking.

Table 4.1: Summarised educators' responses on Presentation, Decision-making, and Personalization aspects of REACT.

The collected comments are also categorized based on presentation, decision-making, and personalization, shown in Table 4.1. The presentation category summarizes the comments on interactive plots, dynamic table, dendrogram and other interactive features. The decision-making category summarizes comments about clustering, alerts, recommendations, and KPIs. The personalization category summarizes the comments on additional needs and improvements on REACT based on educators' specific fields.

4.4.2 System Usability Scale (SUS) Scores and Usefulness

The SUS scores for all the participants are displayed in Figure 4.4. The average SUS score for REACT based on all the responses (N =33) was calculated as 75.37 points. Based on the interpretation REACT is Acceptable and the adjective rating is between Good & Excellent as shown in Figure 4.5. The lowest (35 points) SUS score is noted by a participant from Humanities & Social Sciences while highest (97.5 points) SUS score is noted by two participants from the Science domain. The average SUS scores are 77 points, 78.86 points, and 70.83 points for Engineering, Science and Humanities & Social Sciences domains respectively. Overall, REACT seems to be Acceptable based on the responses from Engineering and Science domain participants. It also indicates adjective rating between Good & Excellent. On the other hand, REACT seems to Marginal (High) based on the responses from the Humanities & Social Sciences domain which indicates adjective rating between OK & Good as shown in Figure 4.6.

Additionally, we also requested all the participants to score different tabs of REACT between 0 to 5. Here, 0 is the lowest score and 5 is the highest score. From Table 4.2, it can be observed that overall the highest mean score (4.51 points) is noted for the Quick Analysis tab, while the lowest score (3.21 points) is noted for the Public Health tab. This observation is true for all the domains except Humanities & Social Sciences for which Overview and Quick Analysis tabs both scored equal points. The participants from the Science domain allocated on average more points to AI tab compared to other domains. On the other hand, on average participants from Humanities & Social Sciences allocated equal scores to

Overview and Quick Analysis tabs.

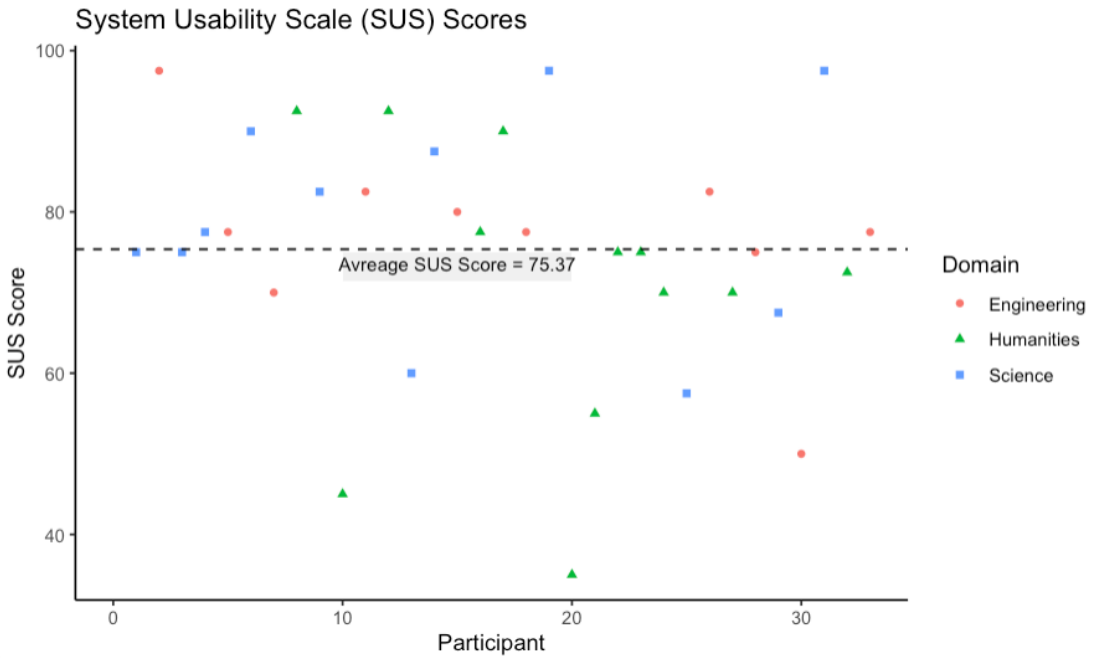


Figure 4.4: The average SUS score for REACT is 75.37 points. The highest SUS score is calculated for the Science domain while the lowest SUS score is calculated for Humanities & Social Sciences.

Table 4.2: The Quick Analysis tab scored the highest mean score (4.51 points), while the Public Health tab scored the lowest mean score (3.21 points).

Tab	**Overall**	**Science**	**Engineering**	**Humanities & Social Sciences**
Overview	4.45	4.54	4.3	**4.5**
Quick Analysis	**4.51**	**4.63**	**4.4**	**4.5**
Scorecard	3.57	4	3.4	3.33
AI	4.03	4.18	4.1	3.83
Public Health	3.21	3.18	3.7	2.83

Figure 4.5: Overall, REACT is Acceptable, and the adjective rating is between Good & Excellent.

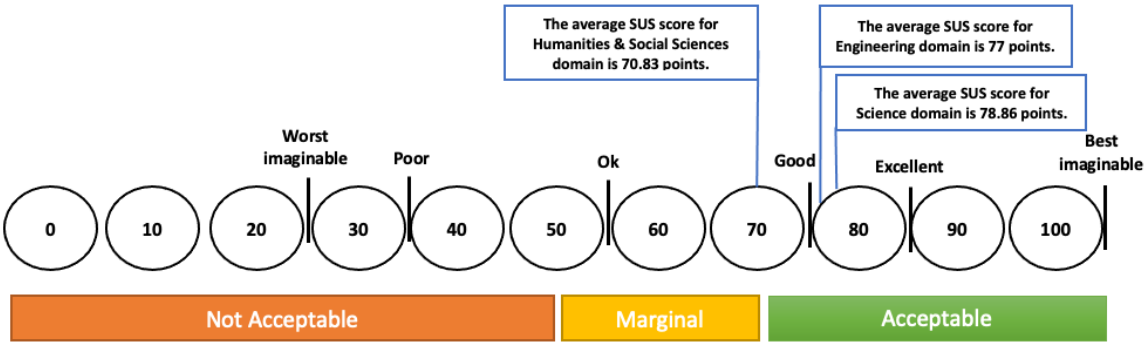


Figure 4.6: The adjective rating of REACT for Science and Engineering domains is between good & Excellent. On the other hand, for the Humanities & Social Sciences domain, the adjective rating is between OK & Good.

4.4.3 Quantiative Data Analysis

Studies have [174,175] indicated that scores from single likert items are less valid, less accurate and less reliable while composite scores are more reliable [49, 50, 175, 176]. Therefore, the results in this section are represented in terms of mean composites scores. The information on questions and domain-wise responses are included in the form of Likert plots in the Appendix.

Table 4.3: The overall mean composite scores from the questionnaire indicate REACT fulfills all (Reaction, Learning, Behaviour, Result, and Effectiveness) criteria.

Criteria	**Domain**	**Mean Composite Score**	**Mean Acceptance Score**
Reaction	All	**68.42**	64
	Science	**69.02**	
	Engineering	**68.2**	
	Humanities Social Sciences	**68**	
Learning	All	**17.09**	16
	Science	**17.45**	
	Engineering	**17.2**	
	Humanities Social Sciences	**16.67**	
Behavior	All	**16.52**	16
	Science	**17.18**	
	Engineering	**16.9**	
	Humanities Social Sciences	15.58	
Result	All	**12**	12
	Science	**12.27**	
	Engineering	11.5	
	Humanities Social Sciences	**12.08**	
Effectiveness	All	**24**	24
	Science	**25**	
	Engineering	**24.6**	
	Humanities Social Sciences	22.58	

Composite Scores

Table 4.3 shows the mean composite scores for the collected data from 33 participants. There were 16, 4, 4, 4, 3, and 6 questions included in the Reaction, Learning, Behavior, Results and Effectiveness criteria respectively. The mean accepted score for every criterion is also included in Table 4.3. These results sheds light on the perception of REACT from different domain experts for understanding its effect on their experience. From the mean composite scores, it can be seen that, overall the mean composite scores are reflecting satisfaction of REACT for all the criteria. The participants from the Science domain scored the highest mean composite scores for all the criteria compared to other domain participants. Further, for Science domain participants the mean composite scores are higher than the mean acceptance score for all the criteria. The mean composite scores for the engineering domain crosses the mean acceptance scores for all the criteria except Result. Similarly, the mean composite scores for the Humanities and Social Sciences domain crosses the mean acceptance scores for the Reaction, Learning and Result criteria.

Correlation Analysis

One of the ways to quantify the relationship between usability and other criteria can be with correlation analysis. The correlation analysis is performed in three steps. First, the composite score of every participant for every criterion is calculated. Next, the Pearson's correlation coefficients are computed between composite scores of criteria and SUS. In the end, the calculated coefficients are tested with 95% level of significance using two tailed hypothesis tests for identifying significant relationship. The null and alternate hypothesis are as follows.

Null Hypothesis: There is not a statistically significant correlation.

Alternate Hypothesis: There is a statistically significant correlation.

Table 4.4 shows the correlation results for all the participants. Based on 95% level of significance if the absolute value of correlation coefficient is above .344 then there is statistically significant relationship. The results indicate significant positive correlation of

SUS with Reaction (r(31) = .57, p < .05), Learning (r(31) = .663, p < .05), Result (r(31) = .420, p < .05), and Effectiveness (r(31) = .427, p < 05). The results for the Science domain are shown in Table 4.5. From the results it can be seen that based on the 95% level of significance there are no criteria which shows statistically significant relationships with the SUS.

Table 4.4: Overall, SUS shows significant positive correlations with Reaction, Learning, Result, and Effectiveness criteria.

	Reaction	**Learning**	**Behavior**	**Result**	**Effectiveness**	**SUS**
Reaction	1	**0.728**	0.302	**0.715**	**0.710**	**0.574**
Learning	**0.728**	1	0.127	**0.420**	**0.645**	**0.663**
Behavior	0.302	0.127	1	0.338	0.106	0.147
Result	**0.715**	**0.420**	0.338	1	**0.469**	**0.420**
Effectiveness	**0.710**	**0.645**	0.106	**0.469**	1	**0.427**
SUS	**0.575**	**0.663**	0.147	**0.420**	**0.427**	1

Table 4.5: The results for the Science domain indicate that none of the variables shows statistically significant relationships with the SUS.

	Reaction	**Learning**	**Behavior**	**Result**	**Effectiveness**	**SUS**
Reaction	1	**0.746**	**0.616**	**0.797**	**0.719**	0.208
Learning	**0.746**	1	0.524	0.520	0.504	0.549
Behavior	**0.616**	0.524	1	**0.802**	0.364	0.122
Result	**0.797**	0.520	0.802	1	**0.698**	0.235
Effectiveness	**0.719**	0.504	0.364	**0.698**	1	0.265
SUS	0.210	0.549	0.122	0.235	0.265	1

Table 4.6: The Engineering domain shows a significant positive correlation of SUS with Reaction criterion.

	Reaction	**Learning**	**Behavior**	**Result**	**Effectiveness**	**SUS**
Reaction	1	0.544	**0.792**	**0.644**	**0.703**	**0.648**
Learning	0.544	1	**0.781**	0.063	**0.651**	0.413
Behavior	**0.792**	**0.781**	1	0.408	0.621	0.564
Result	**0.644**	0.063	0.408	1	0.132	0.380
Effectiveness	**0.703**	**0.651**	0.621	0.132	1	0.373
SUS	**0.648**	0.413	0.564	0.380	0.373	1

Table 4.7: The Humanities & Social Sciences domain shows a significant positive correlation of SUS with Reaction, Learning, and Result criteria.

	Reaction	**Learning**	**Behavior**	**Result**	**Effectiveness**	**SUS**
Reaction	1	**0.833**	0.106	**0.726**	**0.769**	**0.778**
Learning	**0.833**	1	-0.093	**0.636**	**0.675**	**0.772**
Behavior	0.106	-0.093	1	0.220	-0.152	0.026
Result	**0.726**	**0.636**	0.220	1	**0.678**	**0.657**
Effectiveness	**0.769**	**0.675**	-0.152	**0.678**	1	0.439
SUS	**0.778**	**0.772**	0.026	**0.657**	0.439	1

The results for the Engineering domain are displayed in Table 4.6. Based on 95% level of significance if the absolute value of correlation coefficient is above .632 then there is statistically significant relationship. Based on the 95% level of significance Reaction ($r(8) = .648$, $p < .05$), shows statistically significant relationships with the SUS.

For the Humanities and Social Sciences domain based on 95% level of significance if the absolute value of correlation coefficient is above .602 then there is a statistically significant relationship. Based on results displayed in Table 4.7 it can be seen that the Learning ($r(9) = .833$, $p < .05$), Result ($r(9) = .726$, $p < .05$), and Effectiveness ($r(9) = .769$, $p < .05$) shows statistically significant relationships with the SUS.

4.5 Discussion

User Experience

Reaction - The result from the questionnaire on the Reaction criterion indicates educators feel that REACT can help monitor goal-related activities and fulfill teaching goals. Further, there is also no disagreement with a proper visual representation of data on REACT, which enables rapid perception and helps users to check the information at a glance. This may indicate that educators from different domains may share similar views on goal orientation, information usefulness, and appropriation of visual representation. For the other aspects - visual effectiveness and user-friendliness – some disagreements can be observed. However, they are not sufficient to conclude that educators share significantly different views except in the customization of context. This coincides with the results from usability tests.

Learning - The second user experience criterion for evaluation is Learning, which deals with understanding and reflection. It has been observed during usability study that educators from Science and Engineering domains were very comfortable while using and understanding the different features of REACT. In contrast, it took more time for educators from Humanities & Social Sciences to get familiar with the interface and to understand the different features of REACT. One of the reasons for this can be that the evaluation scenario was math-oriented, and some participants from Humanities & Social Sciences found it challenging to understand. A similar pattern is also observed from the questionnaire. The data indicates no disagreement from educators from Science and Engineering domains for Learning. In contrast, some disagreements can be seen from Social Sciences educators on understanding and learning aspects. Thus, this may expose a need to use different scenarios for testing a LAD. This may also indicate that educators' views from the Humanities & Social Sciences domain slightly differ from the Science and Engineering domains educators.

Behavior - The third user experience criterion for evaluation is Behavior which measures motivation and change in behavior. Most of the educators from all three domains find information on REACT useful for student management and planning as well as managing teaching activities. There is no disagreement about it from Humanities & Social Sciences educators. Some disagreements from Science and Humanities & Social Sciences educators are observed for the questions based on a behavior change. From most of the engineering educators' perspectives, they think REACT can help them to motivate and bring changes in their behavior. Thus, for the Behavior criterion, educators from Science and Humanities & Social Sciences show a similar perspective compared to educators from the Engineering domain.

Result - The fourth user experience criterion for evaluation is Result, and it measures performance improvement and competency development. For this criterion, there is no disagreement from the Humanities & Social Sciences educators. Most of them agree that REACT can help them increase self-management skills, achieve instructional goals, enhance teaching performance, and improve their teaching skills. In contrast, Science and Engineering educators disagree with increasing their self-management skills, enhancing teaching performance, and improving their teaching skills. Thus, this may indicate that the perspective of Science and Engineering educators may be similar for the Result criterion.

Effectiveness - The final user experience criterion is Effectiveness which deals with interpretability, and explainability. The educators from Science and Engineering domains find alerts and recommendations generated by REACT are easy to understand, interpretable, and trustworthy. There is no disagreement for these aspects from Science and Engineering Educators. In contrast, some disagreement from the Humanities & Social Sciences educators is noted for the same aspects. Additionally, some level of disagreement is also observed for understanding the process of clustering with visual and textual explanations from all three domains. Thus, Science and Engineering educators may think similarly about interpretability compared to Humanities & Social

Sciences educators.

Therefore, by considering overall responses, we may say that educators from the Science and Engineering domains perceive user experience similarly. In contrast, the educators from the Humanities & Social Sciences domain have a slightly different perception.

Usability & Usefulness

The perception of usability can also be analyzed from the data collected from think-aloud experiments and questionnaires. During the think-aloud study, it was observed that 5 participants incorrectly answered the first activity. Also, most of the participants were confused while answering the first activity. These patterns may indicate a problem with identifying the recommendations on REACT and may also reflect a need to change the wording or structure to increase usability. Further, most of the Humanities & Social Sciences participants found interpreting plots difficult, but they were very comfortable with textual information. On the other hand, all the educators from Science and Engineering domains were comfortable with analyzing plots and interpreting insights from them. The interesting observation was with a density plot. All the Humanities & Social Sciences domain participants found it very difficult to understand and interpret a density plot. In contrast, most of the participants from the Science and Engineering domains found it easy to understand but not helpful. Similar patterns can also be observed from the responses of the SUS questionnaires. The responses to the SUS questionnaire are included in Appendix. When we asked the educators about their opinion on using REACT frequently, half of the educators from the Humanities & Social Sciences domain were neutral, and some disagreed using REACT frequently. In contrast, all the educators from Science and Engineering domains would like to use REACT frequently. This may indicate that educators from Humanities & Social Sciences have different perceptions about the effectiveness and utility of REACT compared to educators from the Science and Engineering domains.

Next, we asked opinions about their confidence while using REACT and learning to use REACT very quickly. We found that some of the educators from the Humanities &

Social Sciences domain were not confident enough to use REACT and thought that they couldn't learn to use REACT very quickly. In contrast, all the educators from the Engineering domain think that they are confident and learn to use REACT very quickly. Similar patterns are also observed from the Science educators, but a few educators were neutral when asked about their confidence in using REACT. These patterns may expose the learnability aspect of REACT, and Science and Engineering educators indicate similar patterns on the learnability compared to Humanities & Social Sciences educators. The results don't indicate much difference in opinions for the remaining usability aspects – memorability and efficiency. These patterns of usability can also be verified from the average SUS scores. It can be seen that the average SUS scores for Science (78.86) and Engineering (77) domains are very close compared to Social Sciences (70.83) domain. Thus, this may indicate that usability for Science and Engineering educators is similar to Humanities & Social Sciences educators. Additionally, the perception of effectiveness, utility, and learnability may be significantly different for educators from Humanities & Social Sciences than the Science and Engineering educators.

Finally, the data collected about usefulness indicates that the Quick Analysis tab is most useful, and all the domain experts indicate agreement via collected data. The reason can be that this tab is available at the beginning of the in-class activity, which provides sufficient insights about students to educators. The second most useful tab on REACT is the Overview tab. This can be because the Overview tab provides a high-level summary of the class and alerts & recommendations. This information may help educators in the end while making decisions. This information may also indicate successful integration of interactive visualizations, KPIs, textual-template recommendations, and alerts. The third useful tab is the AI tab, which scored higher than 4 points from Science and Engineering participants. In contrast, participants from Humanities & Social Sciences on average scored it less than 4. This data may indicate that the AI tab needs some improvement if it needs to be used by educators from Humanities & Social Sciences. In the end, the Scorecard and Public Health tabs scored on average less than 4 points. This may be because of the density

plot on the Scorecard tab and the utilization of the Covid-19 infection rate as a context on the Public Health tab. Thus, it may be essential to increase the usefulness by removing the density plot and utilizing more relevant context in REACT.

Relation between usability and other criteria

It has been seen that SUS scores show significant positive correlations with Reaction, Learning, Result, and Effectiveness. This may indicate that these four criteria might significantly affect the usability of a LAD. The correlation analysis also indicates significant positive correlations of Reaction with Learning, Result, and Effectiveness. Similar patterns of correlations can also be seen for the Learning, Result, and Effectiveness. Thus, these patterns may indicate that focusing on one criterion from Reaction, Learning, Result, or Effectiveness during the development of a LAD might positively affect the other three, which again may affect usability.

Similarly, the usability can also be quantified using correlation analysis for Science, Engineering, and Humanities & Social Sciences domains. There is no significant correlation of any of the criteria with the SUS score for the Science domain. However, Reaction indicates a significant positive correlation with Learning, Behavior, Result, and Effectiveness. In contrast, the other variables show some significant positive correlations with some criteria but not with all. Thus, we may say that these criteria may or may not affect SUS score and thus usability. For the Engineering domain, the SUS score shows a significant positive correlation with Reaction. The Reaction also indicates a significant positive correlation with Behavior, Result, and Effectiveness, while Learning shows significant positive correlations with Behavior and Effectiveness. This result may be interpreted as the Reaction criterion might positively affect the SUS score and thus usability. Lastly, the Humanities & Social Sciences domain educators show significant positive correlations of SUS score with Reaction, Learning, and Result criteria. Further, the Reaction criterion also indicates significant positive correlations with Learning, Result, and Effectiveness, while Behavior doesn't show any significant relationship with any other criterion. Therefore, considering all the results,

we may say that the useability of a tool is quantified chiefly using the Reaction criterion. Also, if a LAD needs to be developed only for educators in the Science domain, it may be essential to provide more attention to Reaction and Learning criteria, but it may not affect usability.

Chapter 5: Conclusion and Future Work

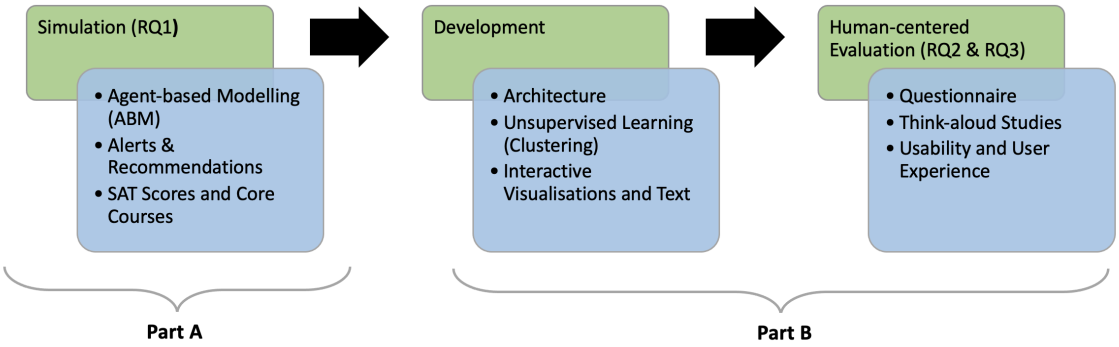


Figure 5.1: Three research questions in this work are divided into two parts - A and B. Part A of this work focus on simulation while Part B focuses on development, and evaluations. This diagram also highlights essential concepts utilized to answer research questions.

This work contributes to the LA domain, and the primary focus of this work is on understanding the effects of a visual LA tool on educators. This objective is achieved by developing a visual LA tool to support data-driven decision-making and understanding its impact on educators. To achieve this objective, three research questions are proposed and answered. These three research questions are mainly divided into two parts - Part A: (i) simulation, and Part B: (i) development, & (ii) evaluation - as indicated in Figure 5.1. The Part A of this work that focuses on simulation is essential because the ABM approach is not commonly used in education. Further, the results from Part B help to improve visual LA tools and provide a better understanding of educators' perceptions. The following sections present the conclusions derived from the research questions, lessons learned during evaluations, and future research directions.

5.1 Conclusions

The first research question - If educators used a LA tool for advising what effect will it have on graduation rates? - focused on the simulation approach and helped to understand the effects of a LA tool on graduation rates. This research question is based on the hypothesis - *If educators used a LA tool with alert and recommendation components, then it can help to increase the college graduation rate.* To test this hypothesis, four different situations are simulated by developing an ABM. The rules in the ABM were based on the average SAT scores and core courses for the Department of Physics and Astronomy at a large public university. The results based on 100 simulations using 100 agents indicated that the average graduation rate for the baseline situation is 61.38%. This result is then compared with other situations using paired t-tests. The paired t-tests indicated that the alerts on a LA tool might increase the average graduation rate up to 0.71%, while recommendations may not always be helpful. Further, the results also indicated that a combination of alerts recommendations could gradually increase the graduation rate up to 0.73%, suggesting that it may provide positive results. These results reflect an encouraging positive effect and shows that alert & recommendation components might help to increase the college graduation rate. Thus, in conclusion, the results from the RQ1 confirm H_1.

The positive results from the simulation experiment motivated the design and development of a Learning Analytic Dashboard (LAD) - a visual LA tool - for educators. Part B, as shown in Figure 5.1, focused on the development and evaluation elements. The second research question - How can a visual LA tool - REACT (Real-time Educational AI-powered Classroom Tool) - that incorporates Visual Analytics (VA) and AI be helpful in classrooms? - is answered by conducting think-aloud experiments with 19 educators. First, a high-fidelity prototype of REACT is developed, which shows one way to integrate VA and AI. It also indicates a possible way to combine model-agnostic explanations for clustering and context to support educators' decision-making process. To understand the usefulness of REACT in the classrooms, a demonstration of a use-case scenario based on real-world data - ASSISTments dataset - is recorded. This demonstration and the deployed version of

REACT are used in a think-aloud experiment - a qualitative research technique. The results conclude that REACT is more useful and suitable for Science and Engineering educators. These experiments also exposed that educators from Humanities & Social Sciences prefer text over visualizations and numbers.

The third research question - How do educators from different domains perceive the integration of VA and AI in real-time on REACT? - is answered by using the survey research technique. This research question is based on the hypothesis - *If educators from different domains use REACT, then they show a similar perception based on Kirkpatrick's four-level evaluation model.* To test this hypothesis, responses were collected from 33 educators. The results indicated that the educators from Science and Engineering noted similar perceptions for four criteria - Reaction, Learning, Behaviour, and Effectiveness. Also, for the usability of REACT, the educators from Science and Engineering domains showed very close SUS scores. Thus, based on Kirkpatrick's four-level evaluation model, the results reflect that Science and Engineering educators show similar perceptions of Reaction, Learning, and Behaviour compared to Humanities & Social Sciences educators. In conclusion, the collected data doesn't provide sufficient evidence to accept H_2.

Further, to verify H_3 - *The participants' scores on usability have significant correlations with Reaction, Learning, Behavior, Result, and Effectiveness criteria* - the descriptive correlation research technique is used. The goal was to understand strength of correlations between SUS scores, Kirkpatrick's four-level evaluation model, and effectiveness for a visual LA tool. Overall, the results from the correlation tests indicate significant positive correlation of SUS scores with Reaction ($r(31) = .57$, $p < .05$), Learning ($r(31) = .663$, $p < .05$), Result ($r(31) = .420$, $p < .05$), and Effectiveness ($r(31) = .427$, $p < 05$). Thus, in conclusion, these results don't provide sufficient evidence on the relationship of Behavior with SUS, which leads to the rejection of H_3. Furthermore, the results from the think-aloud and questionnaire experiments verifies that there is no one-size fit solution for a visual LA tool.

Thus, following three keys takeaways can be highlighted based on think-aloud and questionnaire experiments: 1) Educators from the Science and Engineering domains perceive user experience similarly. In contrast, the educators from the Humanities & Social Sciences domain have a slightly different perception. 2) For usability, the perception of effectiveness, utility, and learnability may be significantly different for the Humanities & Social Sciences educators than the Science and Engineering educators. Further, the data doesn't indicate much difference in opinions for memorability and efficiency. 3) Based on the correlation analysis, the results from this study shows that the usability of a visual LA tool might be described mainly using the Reaction criterion.

5.2 Lessons Learned

The analyzed results indicated educators' positive experience while using REACT and understanding cluster analysis with the help of Model-agnostic explanations. Therefore, six important lessons from this study are given below, which may help future researchers to design visual LA tools.

Lesson 1: Interactivity was well appreciated on REACT by all the educators. Thus, it is helpful to include interactivity to increase the usability of a LAD. The minimum interactivity on a LAD can be provided by incorporating interactive visualizations, filters, and download features. These features can be integrated using plotly, highcharter, etc.

Lesson 2: The results indicated that educators like to review KPIs, and it might have shown a positive effect on the Reaction criterion. Thus, KPIs should be included in a LAD.

Lesson 3: It was observed that the educators from Science and Engineering domains prefer plots for interpretation while educators from Humanities & Social Sciences prefer texts. Therefore, a balance between visualization and text needs to be maintained based on the users. Further, the technological knowledge gap needs to be considered while developing technology-focused tools for the Humanities & Social Sciences educators.

Lesson 4: It is essential for AI-based LADs to include model-agnostic explanations via

visualizations and texts. This may help to provide better interpretability (predictions, recommendations, alerts, etc.).

Lesson 5: Contexts and scenarios can affect users' experience and usability of a tool. Thus, it may be better to use different contexts and user-based scenarios while testing the tool.

Lesson 6: The density plots are challenging to interpret and may not be usable. Thus, the results suggest avoiding using density plots on LADs.

5.3 Future Directions

In the future, the research could go in two directions. The first direction could focus on ABM - improvement, validation, and verification - which may help to make the results generalizable. The second direction could concentrate on the improvement and deployment of REACT. The details on these aspects are provided below.

5.3.1 Agent-based Model (ABM)

The ABM is developed for the Science domain for the simulation experiment based on the Department of Physics and Astronomy's core courses. The results from this ABM are not generalizable to other disciplines and departments. Thus, this ABM could be improved by utilizing data from different domains and departments in the future. Further, student-student interactions and other student dynamics, such as student enrollment in the spring semester, continuing education, etc., could be used for the model enhancement. This may also help in the validation and verification of the model. Additionally, the average college graduation rate is currently used as a metric for model evaluation. In the future, alternative metrics such as the number of semesters left before graduation or a similar kind of metric for judging the system's effectiveness could be used in addition to the average graduation rate.

5.3.2 Improvement and deployment of REACT

In this work, the evaluation experiments were performed with a small sample size which may not be sufficient to gain in-depth and comprehensive insights, especially from the Humanities & Social Sciences domain. Also, the results showed that a visual LA tool is not a one-size-fits-all for the different domains. Thus, it will be crucial to collect more data from Humanities & Social Sciences educators and implement the feedback. This can help to develop a second version of REACT, specialized for the Humanities & Social Sciences domain. Next, this work focuses on Kirkpatrick's four-level evaluation model, but the literature also indicated the Learning Analytics Process Model (LAPM) framework for performing evaluations. Thus, it will be insightful to use the LAPM framework in the assessment in the future that could shed light on the relationship between LAPM, Kirkpatrick's four-level evaluation model, and SUS. Finally, it is also essential to give attention to the scalability aspect, the accuracy of clustering structure with high-dimensional data, and validating its user experience. This could help to deploy REACT that can assist educators in data-driven making in classrooms.

Appendix A: Questionnaire

The following questions are based on reaction criterion which measure goal-orientation, information usefulness, visual effectiveness, appropriation of visual representation, and user friendliness.

1. REACT helps to fulfill your teaching goals by presenting the specific information.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

2. REACT helps the user monitor goal-related activities.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

3. REACT displays the information that the user wants to know.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

4. The overview tab on REACT include all or most of the essential information.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

5. REACT fits on a single computer screen.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

6. REACT presents visual information that the user can scan at a glance.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

7. Visual elements on REACT are arranged in a way for that enables rapid perception.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

☐ Strongly disagree

8. REACT includes proper graphical representations of the data.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

9. Plots on REACT appropriately represent the scales and units.

   ☐ Strongly agree

   ☐ Somewhat agree

   ☐ Neither agree nor disagree

   ☐ Somewhat disagree

   ☐ Strongly disagree

10. REACT delivers information in a concise, direct and clear manner.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

11. REACT uses appropriate pre-attentive attributes such as form and color.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

12. REACT displays information correctly on both desktop computers and mobile devices.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

13. REACT is easy to access.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

14. REACT is customized based on the instructor's context.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

15. REACT has intuitive interface and user-friendly menus.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

16. REACT allows the user to explore more information that is embedded or hidden on a single page (such as the information that appears when hovering over an interactive plot).

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

The following questions are based on learning criterion which measure understanding and reflection.

17. A user understands what the visual information on REACT implies.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

18. A user is able to compare the student's status or position in relation to the overall activity patterns.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

19. A user can monitor the student's learning process consistently based on the information present on REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

20. A user can compile the information on REACT that is related to his/her teaching activity.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

The following questions are based on behavior criterion which measure increase in motivation and change in behavior.

21. A user is motivated to be engaged in studying his/her teaching approach as he/she reviews REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

22. A user can make plans for his/her teaching and students' management (forming groups in class etc.) based on the information shown on REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

23. A user can manage his/her teaching activities (giving extra questions, reading material etc.) based on REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

24. A user can make changes in teaching interventions as he/she monitors the information on REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

The following questions are based on result criterion which measure performance improvement and competency development.

25. REACT can help the user to achieve their instructional goal.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

26. REACT can enhance the user's teaching performance and improve their teaching skills.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

27. REACT enhances the user's self-management skill.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

The following questions measure the effectiveness of REACT from interpretability, and explainability point of view.

28. Textual information on REACT such as recommendations, alerts, as well as other information is easy to understand.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

29. REACT provides interpretable insights about the students in the classroom.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

30. Visual and textual explanations are interpretable enough to understand the process of clustering.

    ☐ Strongly agree

    ☐ Somewhat agree

    ☐ Neither agree nor disagree

    ☐ Somewhat disagree

    ☐ Strongly disagree

31. I can trust the recommendations provided by REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

32. I can trust the alerts provided by REACT.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

33. The number clusters of students generated by REACT are trustworthy for making important decisions.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

The following questions are based on the System Usability Scale (SUS), which helps to measure the REACT's usability.

34. I think that I would like to use this application frequently.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

35. I found the application unnecessarily complex.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

36. I thought the application was easy to use.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

37. I think that I would need the support of a technical person to be able to use this application

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

38. I found the various functions in this application were well integrated.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

39. I thought there was too much inconsistency in this application.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

40. I would imagine that most people would learn to use this application very quickly.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

41. I found the application very cumbersome to use.

☐ Strongly agree

☐ Somewhat agree

☐ Neither agree nor disagree

☐ Somewhat disagree

☐ Strongly disagree

42. I felt very confident using the application.

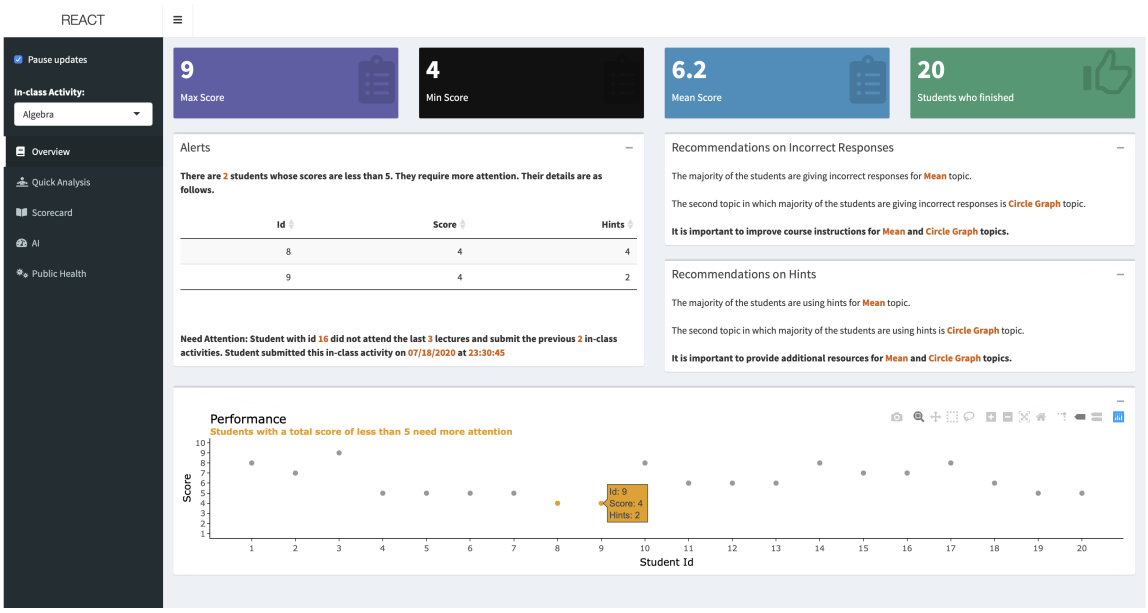
- ☐ Strongly agree
- ☐ Somewhat agree
- ☐ Neither agree nor disagree
- ☐ Somewhat disagree
- ☐ Strongly disagree

43. I needed to learn a lot of things before I could get going with this application.

- ☐ Strongly agree
- ☐ Somewhat agree
- ☐ Neither agree nor disagree
- ☐ Somewhat disagree
- ☐ Strongly disagree

Please rate the usefulness of different tabs of REACT out of 5 points, where 5 means most useful, and 1 means least useful.

44. Tab 1 - Overview



☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

45. Tab 2 - Quick Analysis

☐ 1

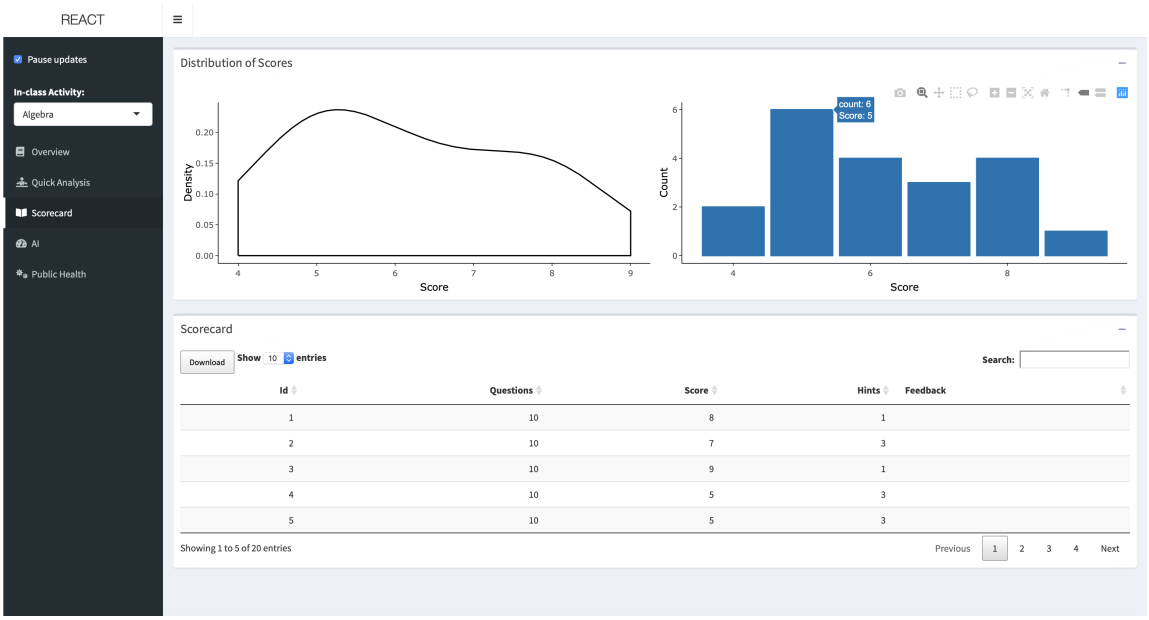
☐ 2

☐ 3

☐ 4

☐ 5

46. Tab 3 - Scorecard



☐ 1

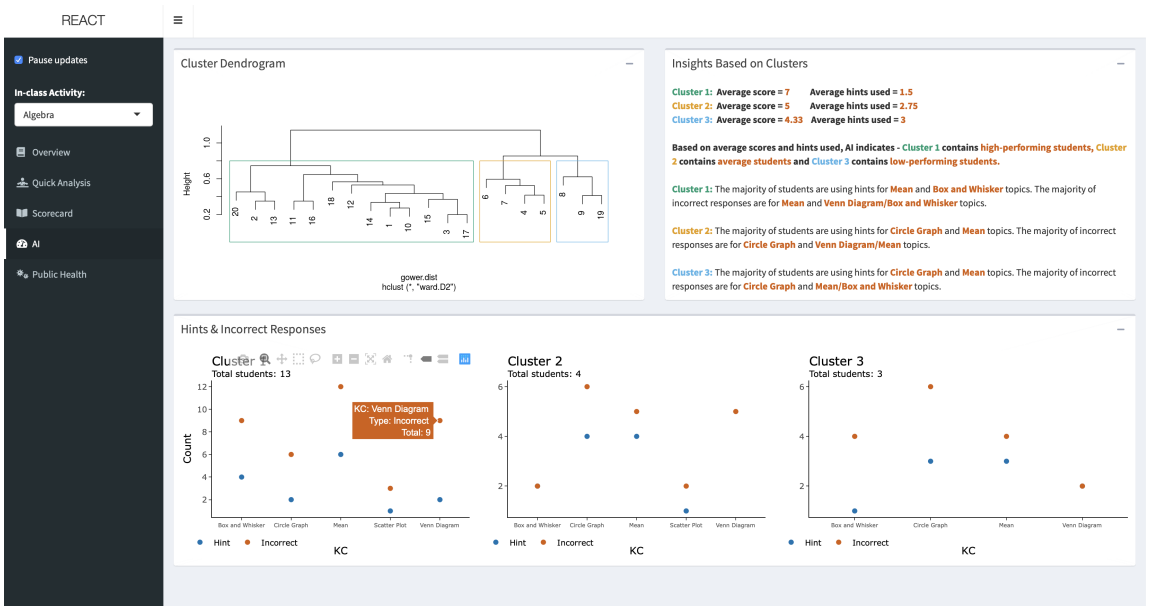
☐ 2

☐ 3

☐ 4

☐ 5

47. Tab 4 - AI



☐ 1

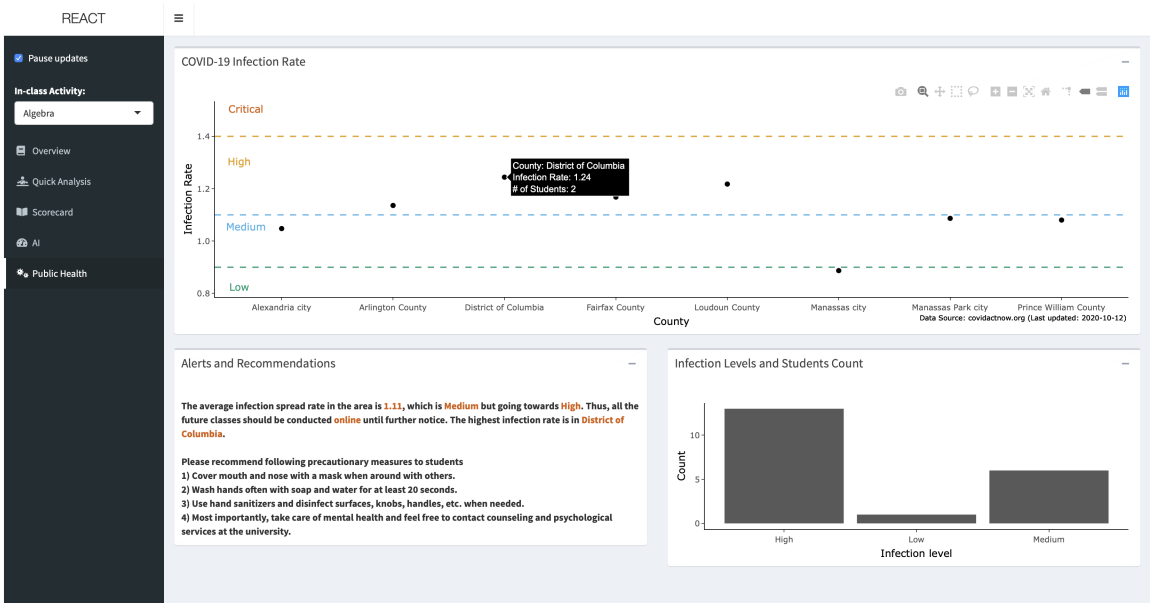
☐ 2

☐ 3

☐ 4

☐ 5

48. Tab 5 - Public Health



- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5

Open ended questions.

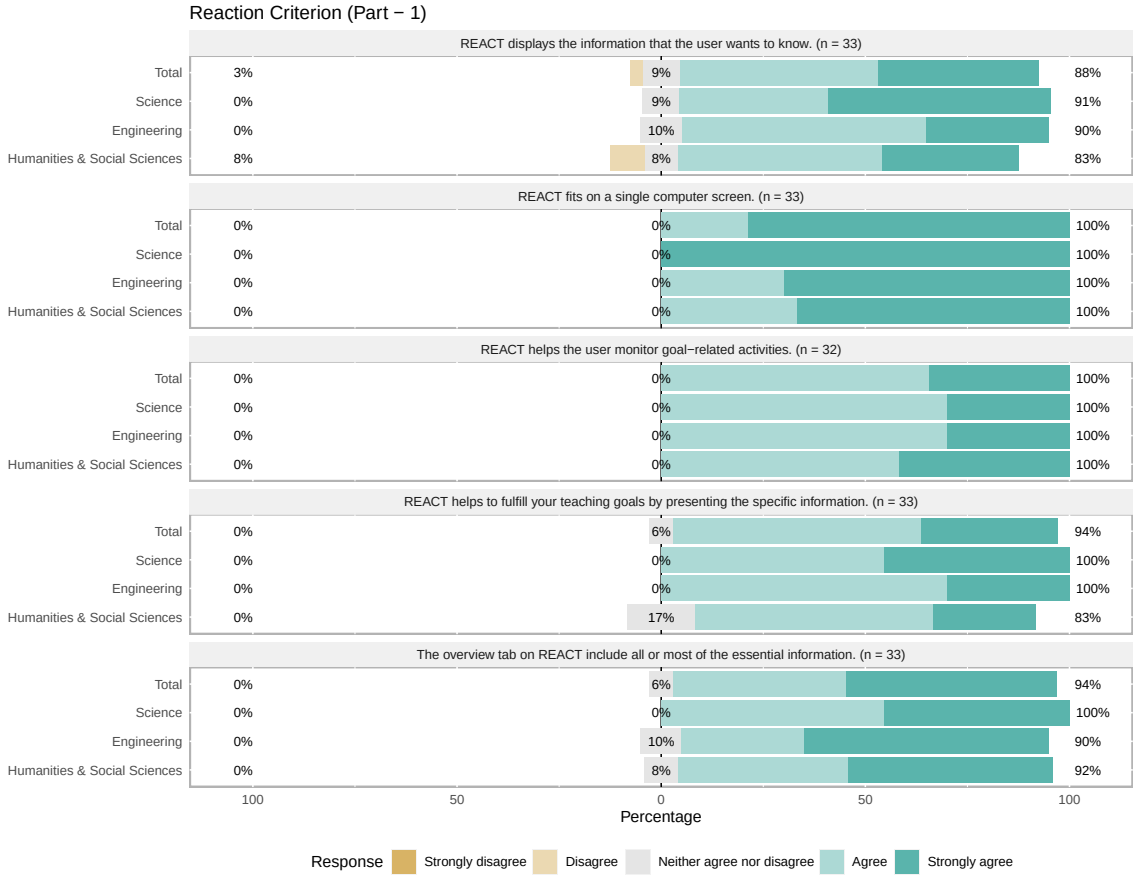
49. Which other information would you like to see in REACT?

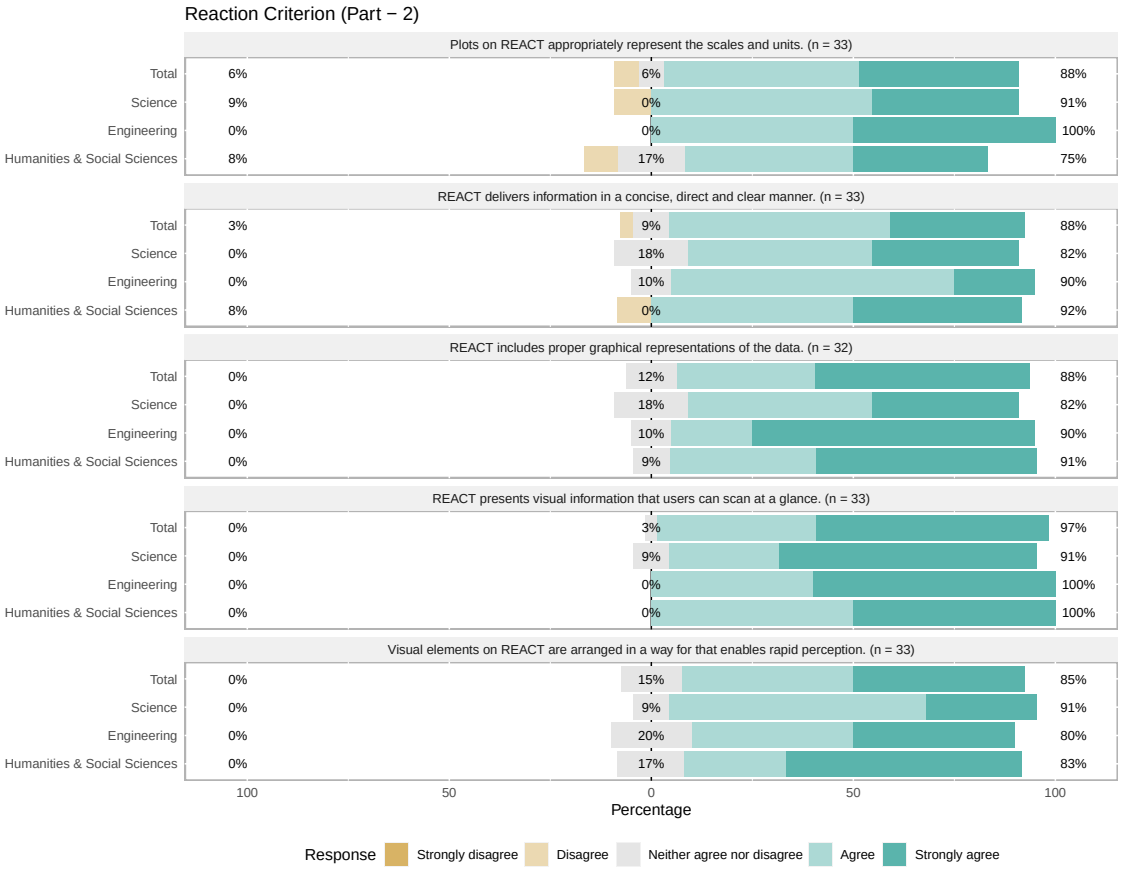
50. Which information would you **not** like to see in REACT?

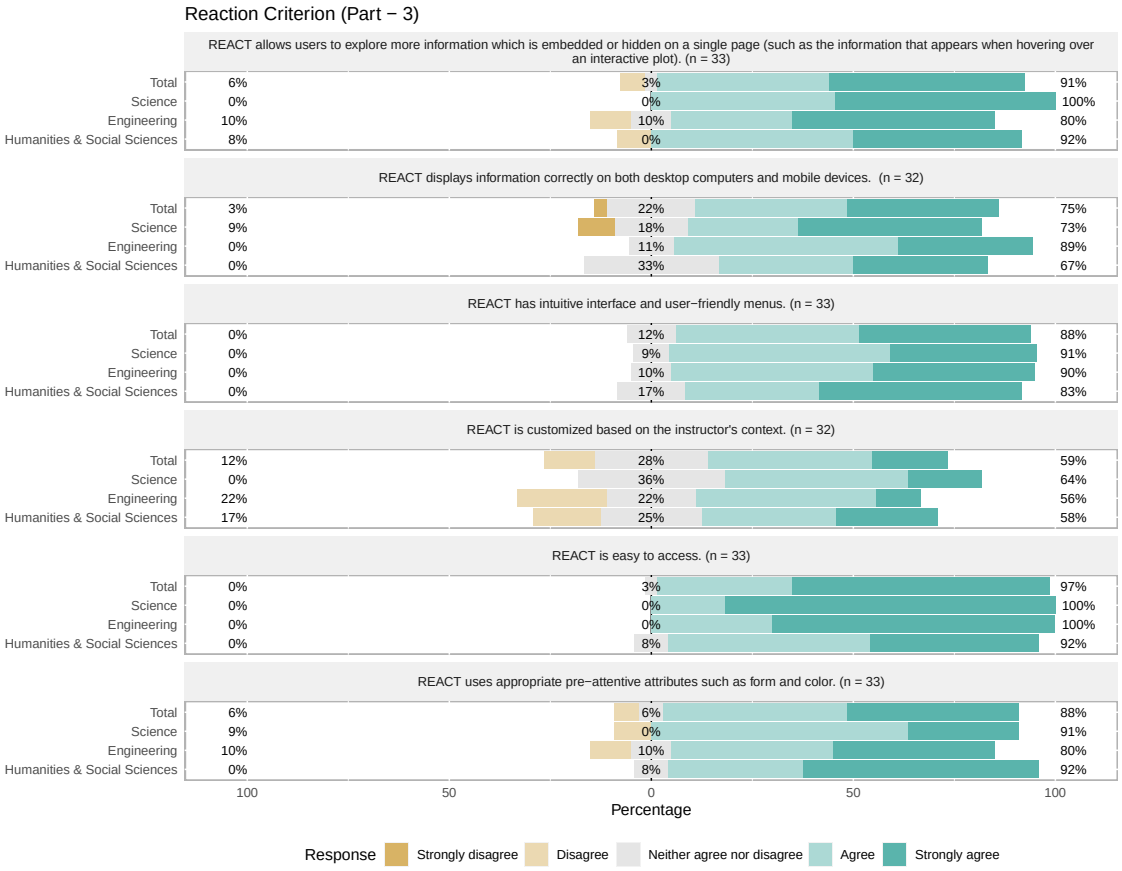
51. Any additional comments on any aspect of REACT

Appendix B: Questionnaire Responses

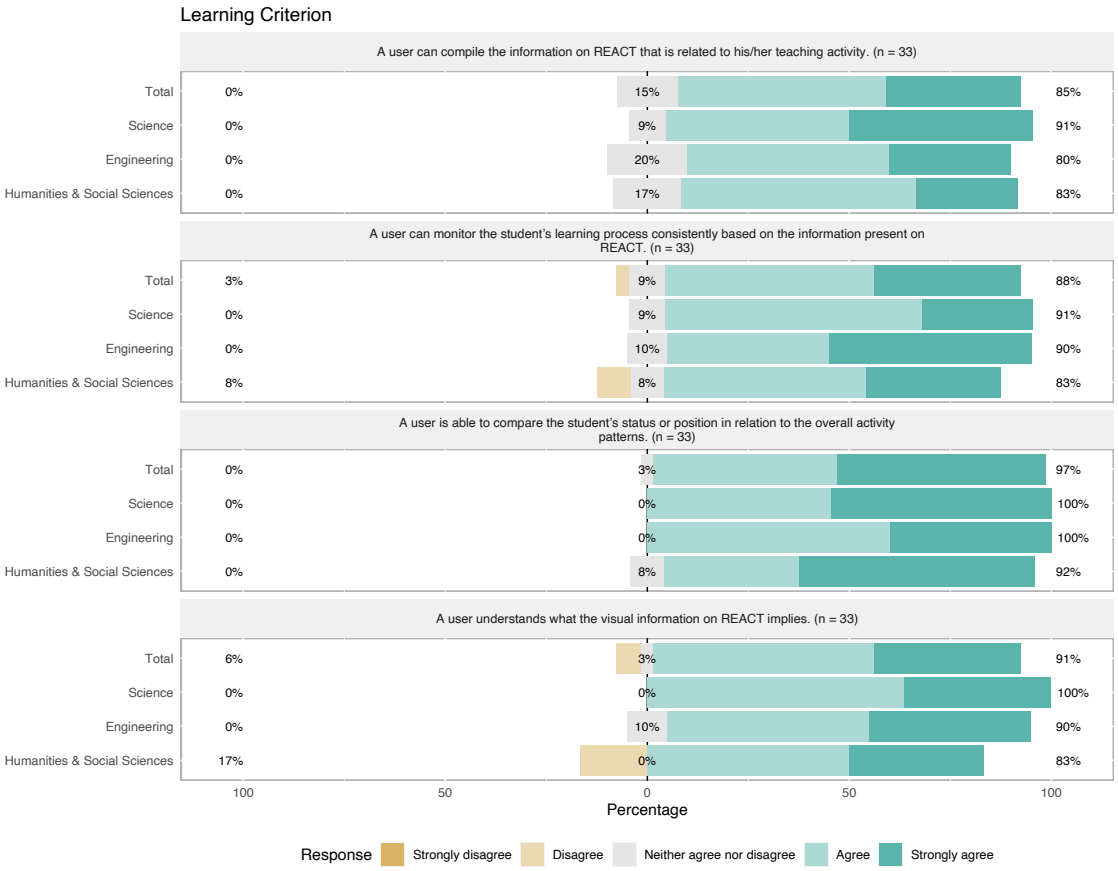
B.1 Responses on Reaction Criterion



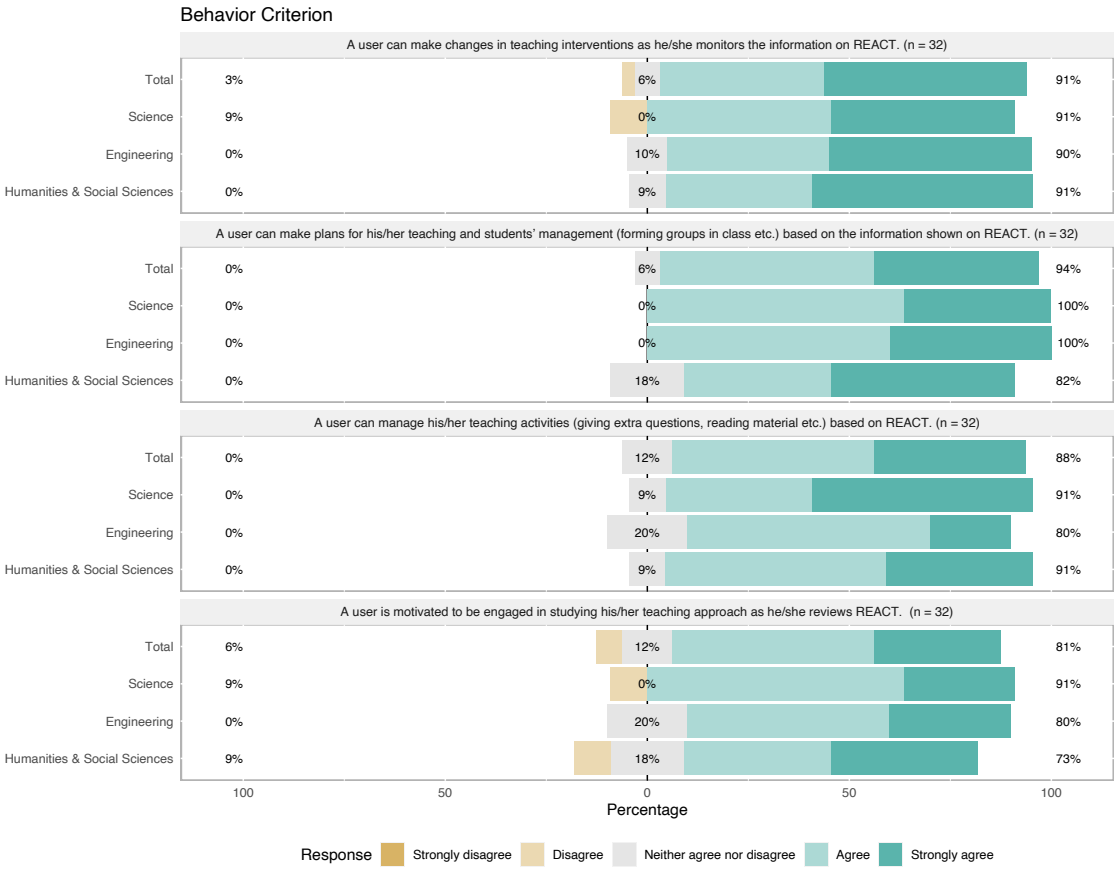

Reaction Criterion (Part − 2)
Plots on REACT appropriately represent the scales and units. (n = 33)
Total 6% 6% 88%
Science 9% 0% 91%
Engineering 0% 0% 100%
Humanities & Social Sciences 8% 17% 75%
REACT delivers information in a concise, direct and clear manner. (n = 33)
Total 3% 9% 88%
Science 0% 18% 82%
Engineering 0% 10% 90%
Humanities & Social Sciences 8% 0% 92%
REACT includes proper graphical representations of the data. (n = 32)
Total 0% 12% 88%
Science 0% 18% 82%
Engineering 0% 10% 90%
Humanities & Social Sciences 0% 9% 91%
REACT presents visual information that users can scan at a glance. (n = 33)
Total 0% 3% 97%
Science 0% 9% 91%
Engineering 0% 0% 100%
Humanities & Social Sciences 0% 0% 100%
Visual elements on REACT are arranged in a way for that enables rapid perception. (n = 33)
Total 0% 15% 85%
Science 0% 9% 91%
Engineering 0% 20% 80%
Humanities & Social Sciences 0% 17% 83%
100
50
0
50
100
Percentage
Response
Strongly disagree
Disagree
Neither agree nor disagree
Agree
Strongly agree


Reaction Criterion (Part − 3)
REACT allows users to explore more information which is embedded or hidden on a single page (such as the information that appears when hovering over an interactive plot). (n = 33)
Total 6% 3% 91%
Science 0% 0% 100%
Engineering 10% 10% 80%
Humanities & Social Sciences 8% 0% 92%
REACT displays information correctly on both desktop computers and mobile devices. (n = 32)
Total 3% 22% 75%
Science 9% 18% 73%
Engineering 0% 11% 89%
Humanities & Social Sciences 0% 33% 67%
REACT has intuitive interface and user−friendly menus. (n = 33)
Total 0% 12% 88%
Science 0% 9% 91%
Engineering 0% 10% 90%
Humanities & Social Sciences 0% 17% 83%
REACT is customized based on the instructor's context. (n = 32)
Total 12% 28% 59%
Science 0% 36% 64%
Engineering 22% 22% 56%
Humanities & Social Sciences 17% 25% 58%
REACT is easy to access. (n = 33)
Total 0% 3% 97%
Science 0% 0% 100%
Engineering 0% 0% 100%
Humanities & Social Sciences 0% 8% 92%
REACT uses appropriate pre−attentive attributes such as form and color. (n = 33)
Total 6% 6% 88%
Science 9% 0% 91%
Engineering 10% 10% 80%
Humanities & Social Sciences 0% 8% 92%
100
50
0
50
100
Percentage
Response
Strongly disagree
Disagree
Neither agree nor disagree
Agree
Strongly agree

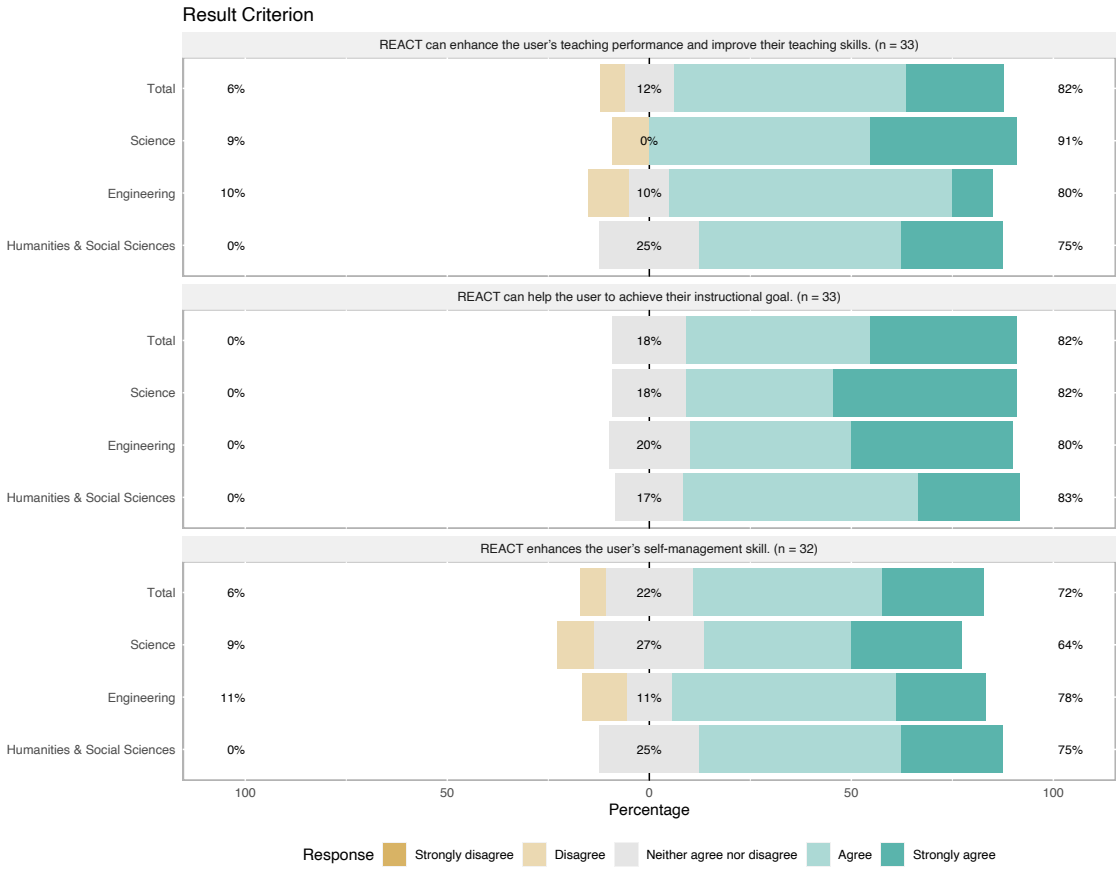
B.2 Responses on Learning Criterion



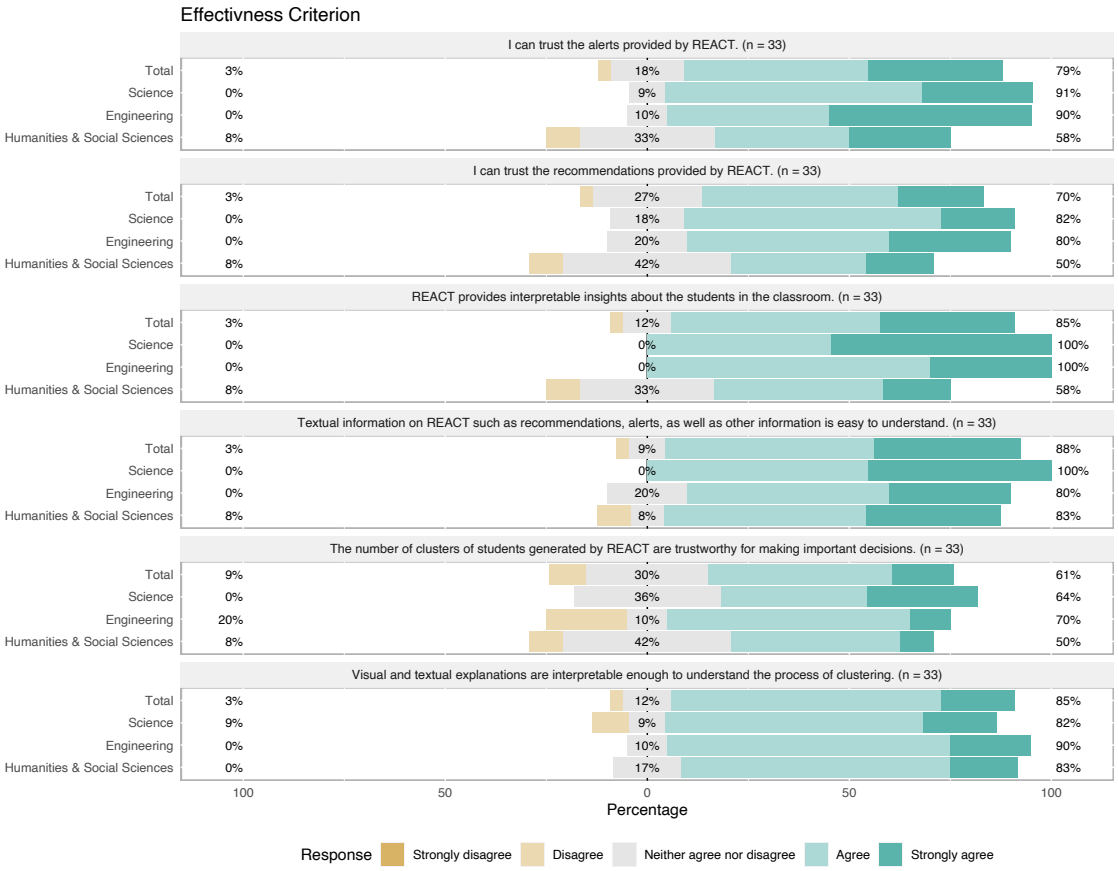
B.3 Responses on Behavior Criterion



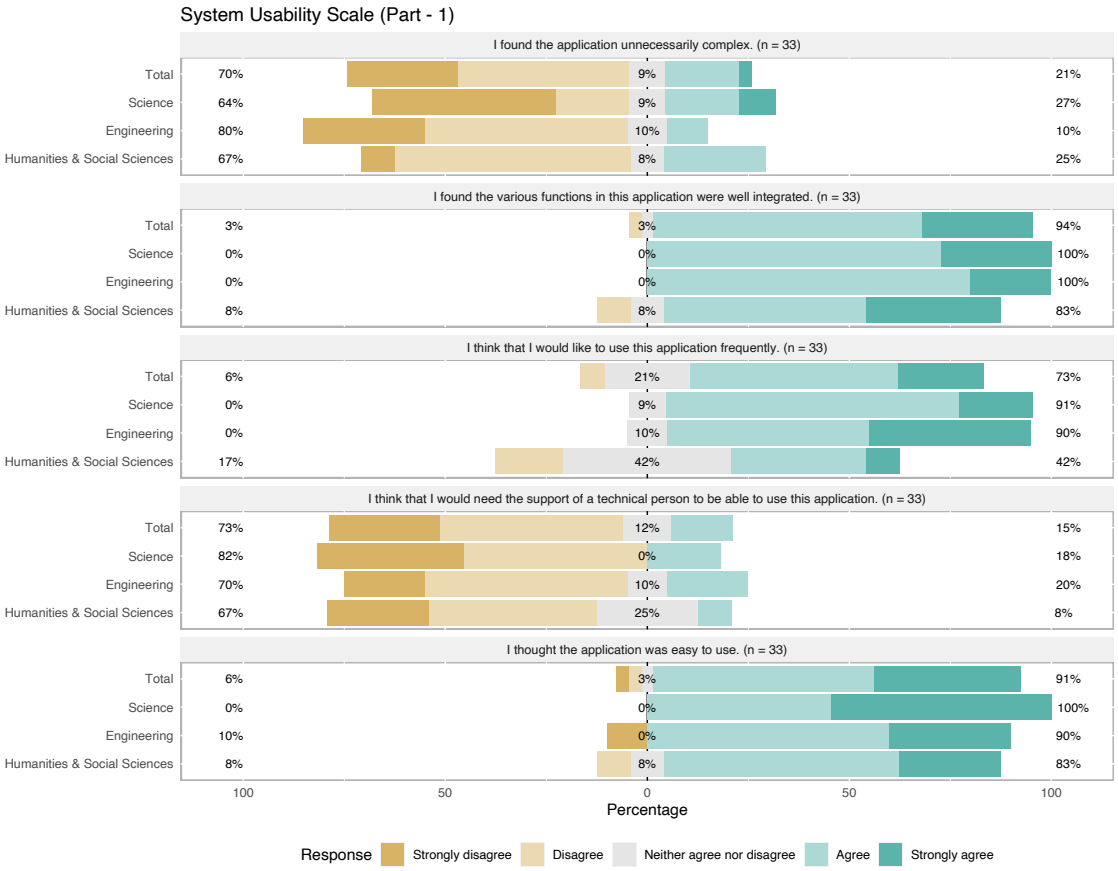
B.4 Responses on Result Criterion

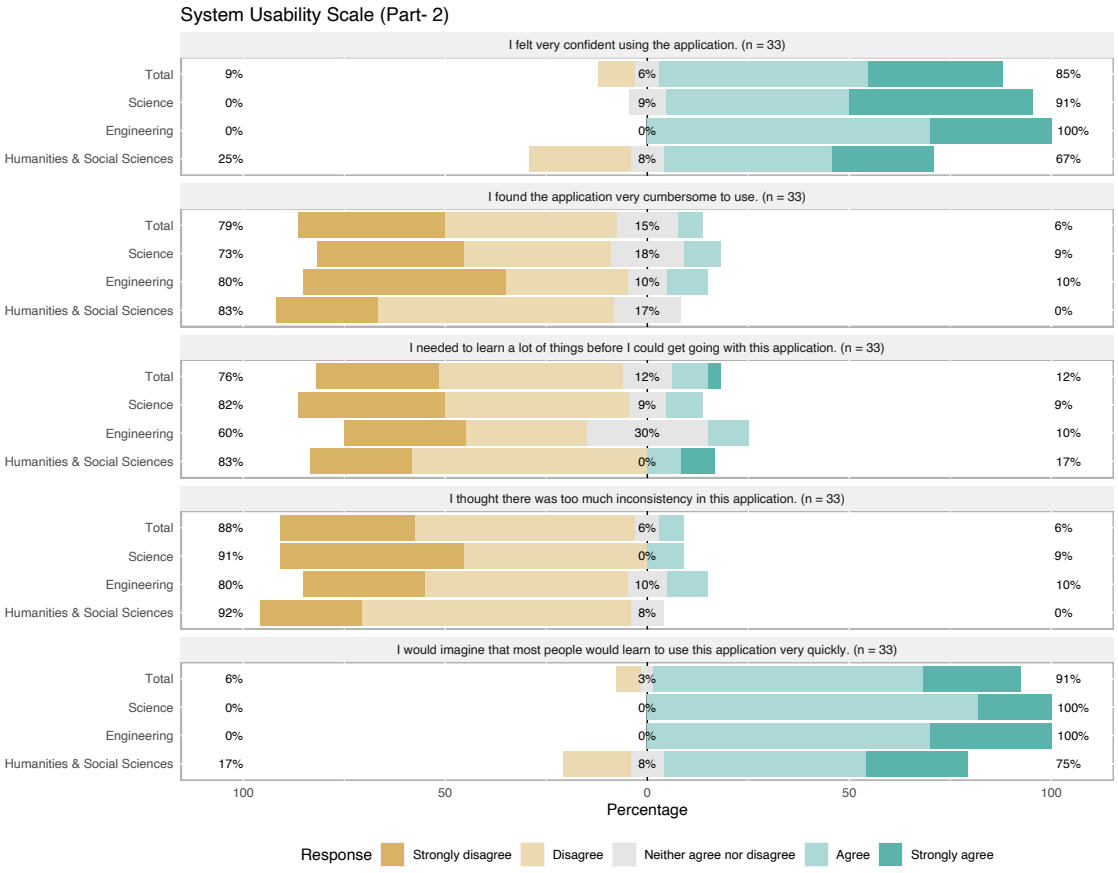


B.5 Responses on Effectiveness Criterion



B.6 Responses on System Usability Scale (SUS)




System Usability Scale (Part- 2)
I felt very confident using the application. (n = 33)
Total 9% 6% 85%
Science 0% 9% 91%
Engineering 0% 0% 100%
Humanities & Social Sciences 25% 8% 67%
I found the application very cumbersome to use. (n = 33)
Total 79% 15% 6%
Science 73% 18% 9%
Engineering 80% 10% 10%
Humanities & Social Sciences 83% 17% 0%
I needed to learn a lot of things before I could get going with this application. (n = 33)
Total 76% 12% 12%
Science 82% 9% 9%
Engineering 60% 30% 10%
Humanities & Social Sciences 83% 0% 17%
I thought there was too much inconsistency in this application. (n = 33)
Total 88% 6% 6%
Science 91% 0% 9%
Engineering 80% 10% 10%
Humanities & Social Sciences 92% 8% 0%
I would imagine that most people would learn to use this application very quickly. (n = 33)
Total 6% 3% 91%
Science 0% 0% 100%
Engineering 0% 0% 100%
Humanities & Social Sciences 17% 8% 75%
100
50
0
50
100
Percentage
Response
Strongly disagree
Disagree
Neither agree nor disagree
Agree
Strongly agree

Bibliography

Bibliography

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[2] A. Crooks, A. Heppenstall, N. Malleson, and E. Manley, "Agent-based modeling and the city: a gallery of applications," in *Urban Informatics*. Springer, 2021, pp. 885–910.

[3] W. Van Der Aalst, "Data science in action," in *Process mining*. Springer, 2016, pp. 3–23.

[4] A. Kulkarni and O. Gkountouna, "Demonstrating react: a real-time educational ai-powered classroom tool," *arXiv preprint arXiv:2108.07693*, 2021.

[5] R. Baker and E. Wang, "Big data and education," *New York: Teachers College, Columbia University*, 2015.

[6] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Education and Information Technologies*, vol. 23, no. 1, pp. 537–553, 2018.

[7] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, "Vision, challenges, roles and research issues of artificial intelligence in education," 2020.

[8] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda," *International Journal of Information Management*, vol. 48, pp. 63–71, 2019.

[9] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[10] A. Weller, "Challenges for transparency," *arXiv preprint arXiv:1708.01870*, 2017.

[11] C. Lang, G. Siemens, A. Wise, and D. Gasevic, *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research, 2017.

[12] J. A. Larusson and B. White, *Learning analytics: From research to practice*. Springer, 2014, vol. 13.

[13] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with java implementations," *Acm Sigmod Record*, vol. 31, no. 1, pp. 76–77, 2002.

[14] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM— Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[15] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.

[16] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[17] K. Zhang and A. B. Aslan, "Ai technologies for education: Recent research & future directions," *Computers and Education: Artificial Intelligence*, p. 100025, 2021.

[18] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019.

[19] W. A. MENNER, "Introduction to modeling and simulation," *Johns Hopkins APL Technical Digest*, vol. 16, no. 1, pp. 6–17, 1995.

[20] A. Maria, "Introduction to modeling and simulation," in *Proceedings of the 29th conference on Winter simulation*, 1997, pp. 7–13.

[21] H. Bossel, *Modeling and simulation*. AK Peters/CRC Press, 2018.

[22] O. B. Oyediran, "Mathematical modelling and simulation and applications," in *Workshop on COMSOL Software held at NMC, 31st July–6th August*, 2016.

[23] L. G. Birta and G. Arbez, *Modelling and simulation*. Springer, 2013.

[24] N. Gilbert and K. Troitzsch, *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[25] A. B. Shiflet and G. W. Shiflet, *Introduction to computational science: modeling and simulation for the sciences*. Princeton University Press, 2014.

[26] D. Helbing, *Agent-Based Modeling*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 25–70. [Online]. Available: https://doi.org/10.1007/978-3-642-24004-1_2

[27] R. J. Allan *et al.*, *Survey of agent based modelling and simulation tools*. Science & Technology Facilities Council New York, 2010.

[28] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," *EDUCAUSE review*, vol. 42, no. 4, p. 40, 2007.

[29] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education." *EDUCAUSE review*, vol. 46, no. 5, p. 30, 2011.

[30] L. P. Prieto, K. Sharma, P. Dillenbourg, and M. Jesús, "Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors," in *Proceedings of the sixth international conference on learning analytics & knowledge*, 2016, pp. 148–157.

[31] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the national academy of sciences*, vol. 99, no. suppl 3, pp. 7280–7287, 2002.

[32] C. M. Macal and M. J. North, "Tutorial on agent-based modeling and simulation," in *Proceedings of the Winter Simulation Conference, 2005.* IEEE, 2005, pp. 14–pp.

[33] G. Triulzi, R. Scholz, and A. Pyka, "R&d and knowledge dynamics in university-industry relationships in biotech and pharmaceuticals: an agent-based model," FZID Discussion Paper, Tech. Rep., 2011.

[34] M. Kramer, "Best practices in systems development lifecycle: An analyses based on the waterfall model," *Review of Business & Finance Studies*, vol. 9, no. 1, pp. 77–84, 2018.

[35] S. Buckingham Shum, R. Ferguson, and R. Martinez-Maldonado, "Human-centred learning analytics," *Journal of Learning Analytics*, vol. 6, no. 2, pp. 1–9, 2019.

[36] J. Ahn, F. Campos, M. Hays, and D. DiGiacomo, "Designing in context: Reaching beyond usability in learning analytics dashboard design." *Journal of Learning Analytics*, vol. 6, no. 2, pp. 70–85, 2019.

[37] A. F. Wise and Y. Jung, "Teaching with analytics: Towards a situated model of instructional decision-making," *Journal of Learning Analytics*, vol. 6, no. 2, pp. 53–69, 2019.

[38] K. Holstein, B. M. McLaren, and V. Aleven, "Co-designing a real-time classroom orchestration tool to support teacher–ai complementarity," *Journal of Learning Analytics*, vol. 6, no. 2, pp. 27–52, 2019.

[39] K. N. Kelley and M. J. Lynch, "Factors students use when evaluating advisors," *NACADA Journal*, vol. 11, no. 1, pp. 26–33, 1991.

[40] R. B. WINSTON Jr and J. A. Sandor, "Developmental academic advising: What do students want?" *NACADA journal*, vol. 4, no. 1, pp. 5–13, 1984.

[41] C. L. Smith and J. M. Allen, "Essential functions of academic advising: What students want and get," *Nacada Journal*, vol. 26, no. 1, pp. 56–66, 2006.

[42] D. Kirkpatrick and J. Kirkpatrick, *Evaluating training programs: The four levels.* Berrett-Koehler Publishers, 2006.

[43] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[44] N. K. Denzin and Y. S. Lincoln, *The Sage handbook of qualitative research.* sage, 2011.

[45] H. Sharp, *Interaction design.* John Wiley & Sons, 2003.

[46] H. R. Bernard, *Research methods in anthropology: Qualitative and quantitative approaches.* Rowman & Littlefield, 2017.

[47] U. Goswami, "Inductive and deductive reasoning." 2011.

[48] J. W. Creswell and C. N. Poth, *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[49] J. P. Rushton, C. J. Brainerd, and M. Pressley, "Behavioral development and construct validity: The principle of aggregation." *Psychological bulletin*, vol. 94, no. 1, p. 18, 1983.

[50] M. J. Strube, "Reliability and generalizability theory." 2000.

[51] E. Seeram, "An overview of correlational research," *Radiologic technology*, vol. 91, no. 2, pp. 176–179, 2019.

[52] N. Burns and S. K. Grove, *Understanding nursing research-eBook: Building an evidence-based practice*. Elsevier Health Sciences, 2010.

[53] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, vol. 1, no. 1, pp. 51–59, 2013.

[54] X. Gu and K. Blackmore, "A systematic review of agent-based modelling and simulation applications in the higher education domain," *Higher Education Research & Development*, vol. 34, no. 5, pp. 883–898, 2015.

[55] C. M. Macal and M. J. North, "Agent-based modeling and simulation," in *Proceedings of the 2009 Winter Simulation Conference (WSC)*. IEEE, 2009, pp. 86–98.

[56] G. De Bakker, J. Van Bruggen, W. Jochems, and P. Sloep, "Introducing the saps system and a corresponding allocation mechanism for synchronous online reciprocal peer support activities," 2011.

[57] Y. Cai, "Graduate employability: A conceptual framework for understanding employers' perceptions," *Higher Education*, vol. 65, no. 4, pp. 457–469, 2013.

[58] K. Mori and S. Kurahashi, "Optimising of support plans for new graduate employment market using reinforcement learning," *International Journal of Computer Applications in Technology*, vol. 40, no. 4, pp. 254–264, 2011.

[59] P. A. Murtaugh, L. D. Burns, and J. Schuster, "Predicting the retention of university students," *Research in higher education*, vol. 40, no. 3, pp. 355–371, 1999.

[60] K. M. Murphy and F. Welch, "Inequality and relative wages," *The American Economic Review*, vol. 83, no. 2, pp. 104–109, 1993.

[61] V. J. Baldridge, F. R. Kemerer, K. C. Green *et al.*, "Enrollments in the eighties: factors, actors, and impacts," *AAHE-ERIC/higher education research report; 3*, 1982.

[62] W. R. Habley, "Academic advisement: The critical link in student retention," *Naspa Journal*, vol. 18, no. 4, pp. 45–50, 1981.

[63] W. R. Habley and R. McClanahan, "What works in student retention? four-year public colleges." *ACT, Inc.*, 2004.

[64] B. B. Crookston, "A developmental view of academic advising as teaching," *NACADA journal*, vol. 29, no. 1, pp. 78–82, 2009.

[65] T. O'Banion, "An academic advising model," *NaCADA Journal*, vol. 14, no. 2, pp. 10–16, 1994.

[66] H. K. Swecker, M. Fifolt, and L. Searby, "Academic advising and first-generation college students: A quantitative study on student retention," *NACADA Journal*, vol. 33, no. 1, pp. 46–53, 2013.

[67] R. B. Winston, S. C. Ender, and T. K. Miller, *Developmental approaches to academic advising*. Jossey-Bass, 1982.

[68] M. Kirk-Kuwaye and D. Nishida, "Effect of low and high advisor involvement on the academic performances of probation students," *NACADA journal*, vol. 21, no. 1-2, pp. 40–45, 2001.

[69] A. D. Young-Jones, T. D. Burt, S. Dixon, and M. J. Hawthorne, "Academic advising: does it really impact student success?" *Quality Assurance in Education*, vol. 21, no. 1, pp. 7–19, 2013.

[70] F. C. Kot, "The impact of centralized advising on first-year academic performance and second-year enrollment behavior," *Research in higher education*, vol. 55, no. 6, pp. 527–563, 2014.

[71] J. L. Kobrin, B. F. Patterson, E. J. Shaw, K. D. Mattern, and S. M. Barbuti, "Validity of the sat® for predicting first-year college grade point average. research report no. 2008-5." *College Board*, 2008.

[72] S. A. Hezlett, N. Kuncel, M. Vey, A. Ahart, D. Ones, J. Campbell, and W. Camara, "The effectiveness of the sat in predicting success early and late in college: A comprehensive meta-analysis," in *annual meeting of the National Council on Measurement in Education, Seattle, WA*, 2001.

[73] J. L. Kobrin and R. S. Michel, "The sat® as a predictor of different levels of college performance. research report no. 2006-3." *College Board*, 2006.

[74] Business insider. (2014) Here's the average sat score for every college major. [Online]. Available: https://www.businessinsider.com/heres-the-average-sat-score-for-every-college-major-2014-10

[75] A. Kulkarni and M. Eagle, "Estimating effects of the decision support system on educational agents with simulations," in *2020 Spring Simulation Conference (SpringSim)*. IEEE, 2020, pp. 1–12.

[76] M. Aparicio and C. J. Costa, "Data visualization," *Communication design quarterly review*, vol. 3, no. 1, pp. 7–11, 2015.

[77] B. chul Kwon, B. Fisher, and J. S. Yi, "Visual analytic roadblocks for novice investigators," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2011, pp. 3–11.

[78] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Computers & Education*, vol. 122, pp. 119–135, 2018.

[79] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," in *2009 IEEE Symposium on visual analytics science and technology*. IEEE, 2009, pp. 3–10.

[80] J.-D. Fekete, J. J. Van Wijk, J. T. Stasko, and C. North, "The value of information visualization," in *Information visualization*. Springer, 2008, pp. 1–18.

[81] J. L. Wesson, "Applying visualisation techniques in novel domains," in *Ninth International Conference on Information Visualisation (IV'05)*. IEEE, 2005, pp. 619–625.

[82] A.-M. Tervakari, K. Silius, J. Koro, J. Paukkeri, and O. Pirttilä, "Usefulness of information visualizations based on educational data," in *2014 IEEE global engineering education conference (EDUCON)*. IEEE, 2014, pp. 142–151.

[83] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: principles, policy & practice*, vol. 5, no. 1, pp. 7–74, 1998.

[84] A. Poulos and M. J. Mahony, "Effectiveness of feedback: The students' perspective," *Assessment & Evaluation in Higher Education*, vol. 33, no. 2, pp. 143–154, 2008.

[85] B. A. Schwendimann, M. J. Rodriguez-Triana, A. Vozniuk, L. P. Prieto, M. S. Boroujeni, A. Holzer, D. Gillet, and P. Dillenbourg, "Perceiving learning at a glance: A systematic literature review of learning dashboard research," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 30–41, 2016.

[86] J. Klerkx, K. Verbert, and E. Duval, "Learning Analytics Dashboards," in *The Handbook of Learning Analytics*, 1st ed., C. Lang, G. Siemens, A. F. Wise, and D. Gaševic, Eds. Alberta, Canada: Society for Learning Analytics Research (SoLAR), 2017, pp. 143–150. [Online]. Available: http://solaresearch.org/hla-17/hla17-chapter1

[87] S. Shemwell, "Futuristic decision-making," *Executive Briefing Business Value from*, 2005.

[88] R. S. Baker, D. Gašević, and S. Karumbaiah, "Four paradigms in learning analytics: Why paradigm convergence matters," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100021, 2021.

[89] A. Van Leeuwen, J. Janssen, G. Erkens, and M. Brekelmans, "Supporting teachers in guiding collaborating students: Effects of learning analytics in cscl," *Computers & Education*, vol. 79, pp. 28–39, 2014.

[90] S. Few, *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc., 2006.

[91] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[92] Y. Park and I.-H. Jo, "Factors that affect the success of learning analytics dashboards," *Educational Technology Research and Development*, vol. 67, no. 6, pp. 1547–1571, 2019.

[93] A. Kulkarni, "Towards understanding the impact of real-time ai-powered educational dashboards (raed) on providing guidance to instructors," *arXiv preprint arXiv:2107.14414*, 2021.

[94] R. Martinez-Maldonado, J. Kay, K. Yacef, M.-T. Edbauer, and Y. Dimitriadis, "Mt-classroom and mtdashboard: supporting analysis of teacher attention in an orchestrated multi-tabletop classroom"," in *International Conference on Computer Supported Collaborative Learning (CSCL2013)*, vol. 1, 2013, pp. 320–327.

[95] M. Tissenbaum, C. Matuk, M. Berland, L. Lyons, F. Cocco, M. Linn, J. Plass, N. Hajny, A. Olsen, B. Schwendimann, M. Boroujeni, J. Slotta, J. Vitale, L. Gerard, and P. Dillenbourg, "Real-time visualization of student activities to support classroom orchestration," in *12th International Conference of the Learning Sciences, ICLS 2016*, vol. 2, 2016, pp. 1120–1127.

[96] N. Valle, P. Antonenko, D. Valle, K. Dawson, A. C. Huggins-Manley, and B. Baiser, "The influence of task-value scaffolding in a predictive learning analytics dashboard on learners' statistics anxiety, motivation, and performance," *Computers & Education*, p. 104288, 2021.

[97] K. Holstein, B. M. McLaren, and V. Aleven, "Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms," in *Proceedings of the seventh international learning analytics & knowledge conference*, 2017, pp. 257–266.

[98] S. Charleer, A. V. Moere, J. Klerkx, K. Verbert, and T. De Laet, "Learning analytics dashboards to support adviser-student dialogue," *IEEE Transactions on Learning Technologies*, vol. 11, no. 3, pp. 389–399, 2017.

[99] M. Ez-Zaouia and E. Lavoué, "Emoda: a tutor oriented multimodal and contextual emotional dashboard," in *Proceedings of the seventh international learning analytics & knowledge conference*, 2017, pp. 429–438.

[100] F. Gutiérrez, K. Seipp, X. Ochoa, K. Chiluiza, T. De Laet, and K. Verbert, "Lada: A learning analytics dashboard for academic advising," *Computers in Human Behavior*, vol. 107, p. 105826, 2020.

[101] J. Han, K. H. Kim, W. Rhee, and Y. H. Cho, "Learning analytics dashboards for adaptive support in face-to-face collaborative argumentation," *Computers & Education*, vol. 163, p. 104041, 2021.

[102] B. B. Schwarz, O. Swidan, N. Prusak, and A. Palatnik, "Collaborative learning in mathematics classrooms: Can teachers understand progress of concurrent collaborating groups?" *Computers & Education*, vol. 165, p. 104151, 2021.

[103] K. Koffka, *Principles of Gestalt psychology*. Routledge, 2013.

[104] A. Rai, "Explainable ai: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, 2020.

[105] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[106] J. Zou and L. Schiebinger, "Ai can be sexist and racist—it's time to make it fair," pp. 324–326, 2018.

[107] K. Verbert, X. Ochoa, R. De Croon, R. A. Dourado, and T. De Laet, "Learning analytics dashboards: the past, the present and the future," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, pp. 35–40.

[108] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[109] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[110] F. M. Zanzotto, "Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, vol. 64, pp. 243–252, 2019.

[111] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," *A review of machine learning techniques for processing multimedia content*, vol. 1, pp. 9–16, 2004.

[112] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[113] B. Everitt, "Cluster analysis . new york: Halsted," 1993.

[114] K. L. R. PJ, "Finding groups in data: an introduction to cluster analysis," *Hoboken NJ John Wiley & Sons Inc*, 1990.

[115] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[116] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[117] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining.* Pearson Education India, 2016.

[118] O. A. Abbas, "Comparisons between data clustering algorithms." *International Arab Journal of Information Technology (IAJIT)*, vol. 5, no. 3, 2008.

[119] A. Nagpal, A. Jatain, and D. Gaur, "Review based on data clustering algorithms," in *2013 IEEE Conference on Information & Communication Technologies.* IEEE, 2013, pp. 298–303.

[120] B. S. Everitt, S. Landau, and M. Leese, "Cluster analysis arnold," *A member of the Hodder Headline Group, London*, pp. 429–438, 2001.

[121] R. P. Adams, "Hierarchical agglomerative clustering," in *Proc. Ninth SIAM Data Mining Conf.(SDM'09),* $\Sigma\varepsilon\lambda$, 2009, pp. 510–516.

[122] T. SVVorensen, "A new method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on danish commons," *K. Dan. Videns. Selsk*, vol. 5, no. 4, p. 1, 1948.

[123] P. J. Rousseeuw and L. Kaufman, "Finding groups in data," *Hoboken: Wiley Online Library*, vol. 1, 1990.

[124] E. Bainomugisha, A. L. Carreton, T. v. Cutsem, S. Mostinckx, and W. d. Meuter, "A survey on reactive programming," *ACM Computing Surveys (CSUR)*, vol. 45, no. 4, pp. 1–34, 2013.

[125] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The craft of information visualization.* Elsevier, 2003, pp. 364–371.

[126] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[127] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[128] M. Kuhn and K. Johnson, *Feature engineering and selection: A practical approach for predictive models.* CRC Press, 2019.

[129] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program pam)," *Finding groups in data: an introduction to cluster analysis*, vol. 344, pp. 68–125, 1990.

[130] A. Kirk, *Data visualisation: A handbook for data driven design.* Sage, 2016.

[131] J. Feder, "The family educational rights and privacy act (ferpa): a legal overview." Library of Congress, Congressional Research Service, 2013.

[132] A. Abela, "Chart suggestions-a thought starter," *Revisado el*, vol. 20, 2006.

[133] B. Shneiderman, C. Plaisant, M. S. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, *Designing the user interface: strategies for effective human-computer interaction.* Pearson, 2016.

[134] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Modeling and User-Adapted Interaction*, vol. 19, no. 3, pp. 243–266, 2009.

[135] V. Podgorelec and S. Kuhar, "Taking advantage of education data: Advanced data analysis and reporting in virtual learning environments," *Elektronika ir Elektrotechnika*, vol. 114, no. 8, pp. 111–116, 2011.

[136] M. Bazire and P. Brézillon, "Understanding context before using it," in *International and Interdisciplinary Conference on Modeling and Using Context*. Springer, 2005, pp. 29–40.

[137] F. Batarseh and A. Kulkarni, "Context-driven data mining through bias removal and incompleteness mitigation." in *EDML@ SDM*, 2019, pp. 24–30.

[138] P. D. Turney, "The management of context-sensitive features: A review of strategies," *arXiv preprint cs/0212037*, 2002.

[139] W. Holmes, M. Bialik, and C. Fadel, "Artificial intelligence in education," *Boston: Center for Curriculum Redesign*, 2019.

[140] S. Banerjee, P. K. Singh, and J. Bajpai, "A comparative study on decision-making capability between human and artificial intelligence," in *Nature Inspired Computing*. Springer, 2018, pp. 203–210.

[141] S. J. Yang, H. Ogata, T. Matsui, and N.-S. Chen, "Human-centered artificial intelligence in education: Seeing the invisible through the visible," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100008, 2021.

[142] C.-M. Chen, J.-Y. Wang, and L.-C. Hsu, "An interactive test dashboard with diagnosis and feedback mechanisms to facilitate learning performance," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100015, 2021.

[143] K. Verbert, E. Duval, J. Klerkx, S. Govaerts, and J. L. Santos, "Learning analytics dashboard applications," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1500–1509, 2013.

[144] M. Ginda, N. Suri, A. Bueckle, and K. Börner, "Empowering instructors in learning management systems: Interactive heat map analytics dashboard," in *Learning Analytics and Knowledge Conference, Vancouver, BC*, 2017.

[145] I. Jivet, M. Scheffel, M. Specht, and H. Drachsler, "License to evaluate: Preparing learning analytics dashboards for educational practice," in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 2018, pp. 31–40.

[146] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100020, 2021.

[147] A. Weller, *Transparency: Motivations and Challenges*. Cham: Springer International Publishing, 2019, pp. 23–40. [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6_2

[148] M. O. Riedl, "Human-centered artificial intelligence and machine learning," *Human Behavior and Emerging Technologies*, vol. 1, no. 1, pp. 33–36, 2019.

[149] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim, "A survey of human-centered evaluations in human-centered machine learning," in *Computer Graphics Forum*, vol. 40, no. 3. Wiley Online Library, 2021, pp. 543–567.

[150] J. Preece *et al.*, "Extract-chapter 22: Envisioning design," 1994.

[151] Y.-K. Lim, E. Stolterman, and J. Tenenberg, "The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 15, no. 2, pp. 1–27, 2008.

[152] J. Rudd, K. Stern, and S. Isensee, "Low vs. high-fidelity prototyping debate," *interactions*, vol. 3, no. 1, pp. 76–85, 1996.

[153] P. C. Wright, J. C. McCarthy, and T. Marsh, "From usability to user experience," in *UNIVERSITY OF YORK, UK*. Citeseer, 2001.

[154] B. ISO and B. STANDARD, "Ergonomics of human-system interaction," 2010.

[155] W. Hwang and G. Salvendy, "Number of people required for usability evaluation: the 10±2 rule," *Communications of the ACM*, vol. 53, no. 5, pp. 130–133, 2010.

[156] M. Hertzum, "Usability testing: A practitioner's guide to evaluating the user experience," *Synthesis Lectures on Human-Centered Informatics*, vol. 13, no. 1, pp. i–105, 2020.

[157] J. E. Fox, "The science of usability testing," in *Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference, Washington, DC, USA*, 2015, pp. 1–3.

[158] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[159] J. S. Dumas, J. S. Dumas, and J. Redish, *A practical guide to usability testing*. Intellect books, 1999.

[160] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 1993, pp. 206–213.

[161] K. Caine, "Local standards for sample size at chi," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 981–992.

[162] H. Tokkonen and P. Saariluoma, "How user experience is understood?" in *2013 Science and Information Conference*. IEEE, 2013, pp. 791–795.

[163] I. Díaz-Oreiro, G. López, L. Quesada, and L. A. Guerrero, "Standardized questionnaires for user experience evaluation: A systematic literature review," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 31, no. 1, 2019, p. 14.

[164] M. Hassenzahl, S. Diefenbach, and A. Göritz, "Needs, affect, and interactive products–facets of user experience," *Interacting with computers*, vol. 22, no. 5, pp. 353–362, 2010.

[165] V. Roto, M. Obrist, and K. Väänänen-Vainio-Mattila, "User experience evaluation methods in academic and industrial contexts," in *Proceedings of the Workshop UXEM*, vol. 9. Citeseer, 2009, pp. 1–5.

[166] Y. Yoo, H. Lee, I.-H. Jo, and Y. Park, "Educational dashboards for smart learning: Review of case studies," *Emerging issues in smart learning*, pp. 145–155, 2015.

[167] G. M. Sullivan and A. R. Artino Jr, "Analyzing and interpreting data from likert-type scales," *Journal of graduate medical education*, vol. 5, no. 4, pp. 541–542, 2013.

[168] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, p. 396, 2015.

[169] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of likert scales in human-robot interaction studies," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 43–52.

[170] R. F. Guy and M. Norvell, "The neutral point on a likert scale," *The Journal of Psychology*, vol. 95, no. 2, pp. 199–204, 1977.

[171] B. C. Courtenay and C. Weidemann, "The effects of a "don't know" response on palmore's facts on aging quizzes," *The Gerontologist*, vol. 25, no. 2, pp. 177–181, 1985.

[172] T. M. Madden and F. J. Klopfer, "The" cannot decide" option in thurstone-type attitude scales," *Educational and Psychological Measurement*, vol. 38, no. 2, pp. 259–264, 1978.

[173] J. Lee and I. Paek, "In search of the optimal number of response categories in a rating scale," *Journal of psychoeducational assessment*, vol. 32, no. 7, pp. 663–673, 2014.

[174] J. McIver and E. G. Carmines, *Unidimensional scaling*. Sage, 1981, vol. 24.

[175] J. C. Nunnally, *Psychometric theory 3E*. Tata McGraw-hill education, 1994.

[176] J. A. Gliem and R. R. Gliem, "Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales." Midwest Research-to-Practice Conference in Adult, Continuing, and Community ..., 2003.

[177] J. R. Rossiter, "The c-oar-se procedure for scale development in marketing," *International journal of research in marketing*, vol. 19, no. 4, pp. 305–335, 2002.

[178] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[179] T. Nemoto and D. Beglar, "Likert-scale questionnaires," in *JALT 2013 conference proceedings*, 2014, pp. 1–8.

[180] J. Barbera, N. Naibert, R. Komperda, and T. C. Pentecost, "Clarity on cronbach's alpha use," *Journal of Chemical Education*, vol. 98, no. 2, pp. 257–258, 2020.

[181] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International journal of medical education*, vol. 2, p. 53, 2011.

[182] J. Fraenkel and N. Wallen, "How to design and evaluate research in education ne mcgraw-hill higher education," 2000.

Curriculum Vitae

Ajay Kulkarni graduated with a B.E. (Bachelor of Engineering) degree in Computer Engineering in 2013 from the University of Pune, India. This was followed by an M.Tech (Master of Technology) in Modelling & Simulation in 2015 from the University of Pune, India. In Fall 2016, he moved to George Mason University, where he obtained his second Masters's degree in Computational Science in 2018. He continued on to his Ph.D. at George Mason University from Fall 2018 to Spring 2022 in Computational Sciences & Informatics.

List of Publications

1. **Kulkarni, A.** and Gkountouna, O., 2021. Demonstrating REACT: a Real- time Educational AI-powered Classroom Tool. In Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, pp. 708-712.

2. **Kulkarni, A.** and Eagle, M., 2020. Towards Understanding the Impact of Real-Time AI-Powered Educational Dashboards (RAED) on Providing Guidance to Instructors. In International Conference on Educational Data Mining (EDM 2020) pp. 781 - 784.

3. **Kulkarni, A.** and Eagle, M., 2020, May. Estimating effects of the decision support system on educational agents with simulations. In 2020 Spring Simulation Conference (SpringSim) (pp. 1-12). IEEE.

4. Batarseh, F. and **Kulkarni, A.**, 2019. Context-Driven Data Mining Through Bias Removal and Incompleteness Mitigation. In EDML@ SDM (pp. 24-30).