LEVERAGING DATA FOR INCLUSIVE AND EQUITABLE EDUCATION: A MULTI-FACETED STUDY OF EDUCATOR PERCEPTIONS AND PRACTICES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kimberly Williamson

May 2025

LEVERAGING DATA FOR INCLUSIVE AND EQUITABLE EDUCATION: A MULTI-FACETED STUDY OF EDUCATOR PERCEPTIONS AND PRACTICES

Kimberly Williamson, Ph.D.

Cornell University 2025

The integration of data, machine learning, and artificial intelligence (AI) in education has ushered in a new era of data-driven decision-making, particularly in efforts to enhance equity and inclusion within educational institutions. However, despite the significant increase in using data for these purposes, there remains limited empirical research exploring the effectiveness and potential unintended consequences of such practices. This dissertation aims to bridge this gap by presenting a series of studies that provide empirical evidence on the backlash to using data for Equity, Diversity, and Inclusion (EDI). Moreover, it offers recommendations on effectively using data for EDI while minimizing future backlash.

The dissertation explores several research questions, including how learning analytics dashboard research is being used to improve justice, equity, diversity, and inclusion (JEDI) in higher education, as well as the potential maintenance or exacerbation of inequitable outcomes in this context. It also delves into scalable measures for equality and inclusion in courses and examines how educators might use equality and inclusion metrics to make decisions regarding EDI. Additionally, the dissertation investigates the impact of explaining complex algorithms on educators' attitudes and intent to use AI-powered tools, shedding light on the differences in educator perceptions of tools using simple versus complex algorithms. By addressing these questions, the dissertation aims to

contribute to the understanding of using data for EDI and to provide practical insights for educators and educational institutions.

Through these studies, the dissertation seeks to provide valuable insights into the challenges and opportunities associated with using data for EDI in education. By examining the implications of data-driven decision-making for equity and inclusion, it aims to offer actionable recommendations for educators and educational organizations to navigate this complex landscape effectively.

BIOGRAPHICAL SKETCH

Kimberly Williamson's research employs innovative data science methodologies to collect and present insights to enhance educational outcomes. Over the course of a decade-long career, they have engaged in various roles directed toward this objective. At Cornell University, they conducted mixed-methods studies and utilized data modeling skills to generate insights for educators. Additionally, they held the position of Lead Instructor for Break Through Tech, where they enhanced student support and academic performance by coordinating teaching assistants and organizing small breakout sessions.

In their prior positions, they served as Software Engineers/Data Architects at the Utah Education Network, where they led the evaluation and selection of data warehouse components and acted as the Solutions Architect for the statewide longitudinal data system. Moreover, they worked as Senior Reporting Analysts at Kent State University, spearheading data extraction and business intelligence processes for student data.

They possess a robust skill set in data analysis, data science, data structures and algorithms, development operations, and various programming languages. Their contributions have been recognized through numerous awards, fellowships, and grants, including the 2024 Innovation Science for Education Analytics (ISEA) Data Science Fellow and the 2021-2023 Senior Data Science Fellow at Cornell University. Furthermore, their educational background comprises a Master's in Information Science from Cornell University, a Master's in Education from Iowa State University, and a Bachelor's in Industrial Engineering from Iowa State University.

To the ancestors who dared to dream: I am the realization of their wildest visions.

ACKNOWLEDGEMENTS

I want to take a moment to express my heartfelt gratitude to everyone who has played a part in my journey to this point. A special shoutout goes to my Iowa State University mentors, who believed that a two-year master's program was just the right fit for me. I also appreciate the invaluable knowledge I gained from my colleagues at UNLV, Kent State, and the Utah Education Network. Truly, my journey wouldn't have been possible without all the wonderful individuals I've met along the way. At Cornell, I want to especially thank Rene Kizilcec for believing in a student with an unconventional background. Even when we didn't see eye to eye, I've undoubtedly grown as a researcher and methodologist, all thanks to your insightful guidance. Whenever a tricky survey use case popped up, I knew I could count on Sean Fath for smart solutions. Sean, your encouragement to pursue research that tackles our cultural challenges, no matter how small the racial effect size might be, has resonated deeply with me. To Jeff Rzeszotarski, I'm grateful for your unwavering support of my ideas and feelings throughout this doctoral journey. You've been such a reliable and reassuring foundation for me and many other students in our department.

I also want to shine a light on my amazing peers. We didn't just get through a pandemic together; we also completed a doctoral program. Sarah Riley and Jen Liu, I truly appreciate our much-needed Zoom chats over the years. JiYong Cho and Mina Chen, the unique experiences we've shared mean so much to me, and I'm thankful for your steady presence in my life, which allowed me to offer my support as well. To Scott Allen, my personal devil's advocate, thanks for pushing me to stay sharp and on my toes. I can't forget Barbara Woske, whose invaluable administrative support and knack for managing countless deadlines over the years have kept us all on track.

Reflecting on the challenges brought by the pandemic and the absence of a School of Education at Cornell, I feel incredibly lucky to have had Christine Bae and Kamil Hankour as my steadfast anchors in education. Collaborating with you and sharing supportive conversations over the past couple of years has truly brightened my Thursdays, making our meetings something I always look forward to. Jeremy Franklin and Tanya Lelanuja, although distance may have separated us, your friendship and support have lifted me in ways I can hardly express.

Lastly, to my family, I can't thank you enough for all the support you've shown me throughout my life. Mom and Dad, I deeply appreciate the advantages you provided, knowing that I had to stay two steps ahead as a Black child. To my sister Erica, who inspires me every day, I'm so thankful for your trailblazing spirit and for nudging us both to excel. At one time, being Erica's little sister felt like a burden, but now I see it as the greatest gift, allowing me to witness your greatness while realizing my limitless possibilities. And to my wife, Valerie, I must admit I thought you were a little out of your mind when you first suggested I go for a PhD. I still think you're a bit crazy, but you've proven me wrong again. You recognized the potential in me that I hadn't seen myself and supported me until I could finally see it. Learning alongside you has been a true honor, and I now concede that you might actually be right about interest convergence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

> The point isn't to get people to accept that they have biases, but to get them to see [for themselves] that those biases have negative consequences for others.
>
> Theresa McHenry

Integrating data, machine learning, and artificial intelligence (AI) into education has ushered in a new era of data-driven decision-making. One prominent area where data is increasingly employed is in efforts to enhance equity and inclusion within educational institutions [Xie, 2020, Taylor et al., 2023, Elisa Raffaghelli, 2020, Roegman, 2020]. While the use of data for these purposes has grown significantly, empirical research exploring the effectiveness and potential unintended consequences of such practices remains limited [Williamson and Kizilcec, 2022, Taylor et al., 2023].

In 2020, many educational organizations were asked to address issues of inequity with regards to the increased attention to Anti-Black and Anti-Asian racism [Toraif et al., 2023, Grace et al., 2024]. Organizations responded with statements of support and task forces to evaluate organizational climate and provide recommendations. Within many of these recommendations were calls and initiatives to use data to highlight and explore inequities [Coleman et al., 2022]. While on the surface these recommendations were made to help marginalized communities, the deployment of many of these understudied ini-

tiatives have since been followed by 65 anti-DEI bills (with several focusing specifically on data usage) introduced across the US [Blackstock et al., 2024, Harris et al., 2024]. Rather than progressing these initiatives, they have instead been being dismantled, shunned, and made illegal across many US states.

In this dissertation, I will present a series of studies that provide empirical evidence on the backlash to using data for EDI, then detail recommendations on effectively using data for EDI while minimizing future backlash.

1.1 Research Questions

This dissertation explores how educators might use data to improve EDI. The increasing usage of data in education has forced research to understand what data can be used for EDI, along with how to communicate this data effectively. This dissertation explores these issues through the following questions:

- How is the research on learning analytics dashboards making strides in promoting Justice, Equity, Diversity, and Inclusion (JEDI) within higher education? In what ways might unintended outcomes in this area be maintained or even worsened? What exciting opportunities exist to enhance JEDI through learning analytics dashboards? (Chapter 2)
- What effective measures can we use to boost equality and inclusion in our courses? How would metrics for inclusion and equality differ from the DFW index? How can educators leverage these equality and inclusion metrics to make more informed decisions about EDI? (Chapter 3)
- How does offering a clear explanation for a complex algorithm impact

educators' willingness to utilize an AI-powered tool? What differences can we see in educators' perceptions of tools that employ a simple versus a complex algorithm? Additionally, how does clarifying a complex algorithm shape educators' views toward tools using either complex or simpler algorithms? (Chapter 4)

- What factors might indicate when an instructor feels uncomfortable (either verbally or physically) discussing socio-demographic data? Can we see a link between higher levels of discomfort and an increase in avoidant behaviors? (Chapter 5)

1.2 Dissertation Outline

This dissertation is divided into six chapters. Chapter 2 offers a literature review providing an overview of the use of data dashboards to improve EDI. Four themes were identified using a critical literature review methodology on 45 relevant papers to help guide future research and practice for incorporating EDI into learning analytic dashboards. The first theme addresses how researchers' experience and positionality have the potential to affect research. The second theme addresses the more significant surveillance concerns that happen with collecting and analyzing large datasets. The third theme concerns the implicit pedagogies weaved into learning analytic dashboards that typically represent the majority. The last theme addresses issues surrounding the large amount of time and money resources needed to build these data displays. While the review indicated very few studies directly addressing EDI concepts, these themes help identify future research areas that can be addressed to help improve historical inequities.

In Chapter 3, I introduce new metrics for course evaluation. While previous course evaluation metrics have focused on general measures of how students perform in a course, these new metrics seek to quantify levels of equality and inclusion in courses. In this chapter, I define and explain the properties of these new metrics through various simulations and examples from course datasets. Lastly, an experimental study was designed to assess 242 educators' perceptions and use of these new metrics. As predicted, the new metrics influenced educators' decision-making for the inclusive but not instructional design program to prioritize courses scoring low on equality and inclusiveness. The new metrics also increased educators' confidence and trust in their decisions. This work offers two new metrics for course evaluation and demonstrates their value for empowering educators to make decisions related to equity and inclusion.

Chapter 4 delves into educators' attitudes towards AI tools in education. In two randomized experiments involving 570 educators, I compared their preferences between a simple heuristic algorithm and a complex (Bayesian Knowledge Tracing) algorithm. The focus was on how explanations for the complex algorithm could influence attitudes and adoption. Contrary to prior research on algorithm aversion, which suggested educators might be averse to using tools with complex algorithms, the results from both studies indicated that complexity did not deter educators from using AI tools. Moreover, educators expressed similar levels of trust and confidence in AI tools using complex algorithms, with or without additional explanations. These findings challenge the prevailing notion of algorithm aversion and suggest a shift towards algorithm appreciation, at least in the context of widely used technologies like ITS.

Chapter 5 explores the topic of educator discomfort with data displays show-

ing socio-demographic data. I conducted a mixed-methods study analyzing educators' use and avoidance tendencies of various data visualizations. Findings from 12 interviews and 500 survey responses suggest that when there is a racial gap in performance data in which historically underrepresented racial groups perform worse, educators will try to reduce discomfort by avoiding future data visualizations.

Lastly, Chapter 6 synthesizes the findings of the previous chapters and discusses their implications for using data to improve EDI in education. By examining five key barriers identified in existing literature, the chapter highlights the contributions of this research to addressing these challenges. The discussion then concludes by proposing two key recommendations: the development of EDI data literacy and the use of generative AI to provide personalized instructor feedback.

CHAPTER 2

A REVIEW OF LEARNING ANALYTICS DASHBOARD RESEARCH IN HIGHER EDUCATION: IMPLICATIONS FOR JUSTICE, EQUITY, DIVERSITY, AND INCLUSION

2.1 Introduction

Learning analytics dashboards (LADs) are visualization systems that curate and present data about student learning and engagement in educational contexts [Schwendimann et al., 2017]. They are increasingly used in higher education by a variety of stakeholders, including dashboards for students to monitor their progress in a classes [Bodily and Verbert, 2017], dashboards for faculty to monitor student learning and get feedback on their teaching practice [Brown, 2020], and dashboards for university administrators to manage and support students, instructors, and staff [Guerra et al., 2020]. Although LADs are frequently used by faculty and students, many of them have been designed for staff engaged in student support services like academic advising [Hilliger et al., 2020]. In addition, most LADs are designed for scalability across many students, courses, and organizational units to facilitate their deployment at universities to reach growing numbers of students, instructors, and staff [Meyliana et al., 2014, Ahn et al., 2019]. In particular, providers of major learning management systems (LMS), such as Blackboard and Canvas, have added dashboards as a novel feature available to students and instructors [Instructure, 2021b, Blackboard, 2021b]. Given the pervasive use of LMS in colleges and universities around the world, including over 100 million Blackboard users [Blackboard, 2021a] and over 30 million Canvas users [Instructure, 2021a] as of 2020, the dashboard fea-

ture in LMS likely exposed millions of students and instructors to LADs. The sudden widespread availability of LADs in academic environments raises critical questions about how LADs are designed and used, especially considering that many institutions are grappling with issues of diversity, equality, and inclusion.

Recent advances in learning analytics and educational data mining, combined with an increasing appetite for using data in decision making, have inspired significant research and development efforts around LADs [Schwendimann et al., 2017]. Presenting insights from data collected by learning management and student information systems, LADs have been traditionally used to help students monitor their progress in a course and to help faculty monitor their course as a whole. The status quo of LADs is advancing quickly, incorporating new features like predictive analytics and guidance on how to make sense of the available data for those who make data-informed decisions like students and faculty. It is an opportune time to examine the state of LAD research, especially in light of recent calls to address social inequity in learning analytics [Shum, 2020, SoLAR, 2020]. We conducted a critical literature review to understand how to improve justice, equity, diversity, and inclusion through LAD research and to highlight opportunities for future work in this area.

The acronym justice, equity, diversity, and inclusion (JEDI) has recently been proposed as a change from the commonly used terms diversity and inclusion (DI), or diversity, equity, and inclusion (DEI). This change is not just additive; it prioritizes justice and equity in efforts to address inequities. Truong and Martinez Truong and Martinez [2021] discuss this shift with examples to explain the difference between the DEI and JEDI perspective: one example explains that

DEI is "espousing that we value diversity and inclusion," while JEDI is "connecting these values to accountability for ensuring that our goals are met." In light of this shift, we opted to critically examine research on LADs from a JEDI perspective.

The learning analytics research community has identified a need for more critical scholarship about the work it produces [Castañeda and Selwyn, 2018, Selwyn, 2020]. There are several systematic and comprehensive LAD literature reviews focusing on student usage [Bodily and Verbert, 2017], deployment of LAD applications [Verbert et al., 2013a], the use of learning theories in LADs [Jivet et al., 2018, Matcha et al., 2020], and two general reviews of LAD research as a whole [Schwendimann et al., 2017, Vieira et al., 2018]. However, no critical review of LAD research has been conducted thus far. While systematic literature reviews help readers gain a complete view of a field during a period of time, this broad scope is not conducive to highlighting critical issues in the literature [Paré et al., 2015]. Thus, because LAD research shapes the experiences of many people in education today, this shortcoming can have severe consequences for JEDI in higher education.

The year 2020 brought about a significant push to develop initiatives addressing issues of JEDI across all kinds of institutions and research communities, including Learning Analytics [Shum, 2020, SoLAR, 2020]. Nevertheless, there is significant uncertainty about which directions will create meaningful change. Throughout this review, we will examine how issues of JEDI can be addressed in LAD research to help reduce systemic inequities that give rise to socio-demographic achievement gaps and the underrepresentation of historically disadvantaged groups. We aim not only to review the LAD literature for

these challenges but also to highlight areas in dashboard research where researchers are in a strong position to address issues of social inequity. This critical literature review will add depth to the LAD literature by addressing the following research questions:

RQ1. How is learning analytics dashboard research being used to improve JEDI in higher education?

RQ2. How are inequitable outcomes unknowingly maintained or exacerbated in learning analytics dashboard research?

RQ3. What are the opportunities to improve JEDI in learning analytics dashboard research?

2.2 Literature Search

Following Paré and colleague's [Paré et al., 2015] definition of a critical review, we sought to "reveal weaknesses, contradictions, controversies, or inconsistencies" (p. 189) and "to highlight problems, discrepancies or areas in which the existing knowledge about a topic is untrustworthy" (p. 189). Unlike systematic and comprehensive reviews, a critical review uses a sample of papers instead of reviewing all literature in an area. We approached this review from a critical constructionist epistemology, wherein we searched for alternative ways of knowing and expose unrepresentative assumptions that have been embedded into knowledge [Kincheloe, 2005].

Given that we set out to understand how LADs were being used in higher education to improve student learning outcomes, we initially chose the follow-

ing inclusion criteria for articles in our review: papers about dashboards (a) with a student component (includes both student and non-student facing LADs) that are (b) used in higher education (within and outside of the classroom) and (c) used empirical research methods. We next determined the following search keywords by brainstorming keywords related to LADs: education dashboard, learning dashboard, learning analytics dashboard, advising dashboard, student dashboard, and higher education dashboard. We then compared the first few abstracts obtained from a Google Scholar search for each brainstormed keyword. We found that "higher education dashboard" returned the most relevant papers that met our inclusion criteria. We therefore chose "higher education dashboard" as the initial keyword and used Google Scholar and Scopus to record the metadata (title, journal, year, etc.) for the first 20 papers returned by the search to make the papers retrievable for later reading. These 20 papers were merely a starting point to discover relevant papers. One by one, we read the abstracts and sorted the papers into three folders: Criteria Match, Literature Review, and No Criteria Match. Papers matching the inclusion criteria were sorted into the Criteria Match folder. Existing Literature reviews of LADs were placed into the Literature Review folder. All remaining papers were assigned to the No Criteria Match folder. Figure 2.1 provides a visual description of the process we used to arrive at the final set of articles. The literature search was not limited to a specific time frame.

We skimmed each Criteria Match paper, taking notes on the purpose of the study and how the paper did or did not address JEDI issues. We also performed a backward citation search by keeping a running list of papers citepd in the review papers, which appeared to be potential matches for our inclusion criteria. The existing literature reviews were skimmed for a backward citation search

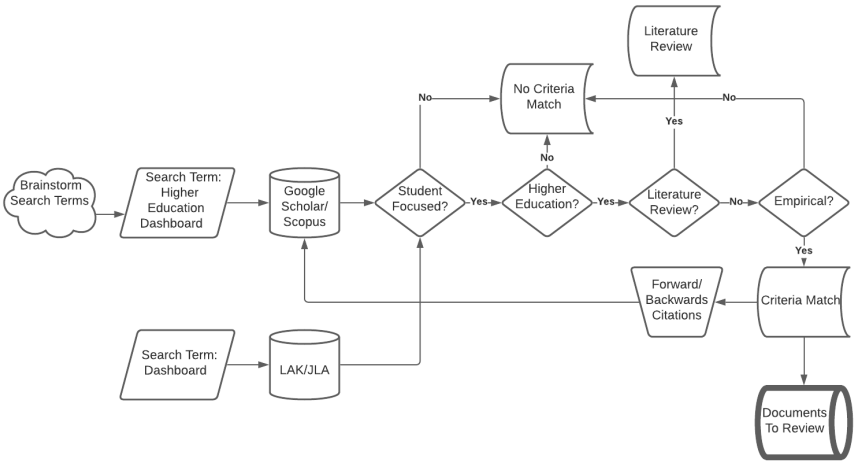


Figure 2.1: Flowchart describing the process used to collect articles for the this critical review sample.

also. Using the new list of papers, we recorded the citations of the papers and sorted them into the appropriate folders. To ensure that we did not miss closely related papers, we conducted a forward citation search on the papers in the Literature Review and Criteria Match folders. This forward citation search was conducted by searching Google Scholar with the paper's title and reviewing the "citepd by" papers. Unlike the previous steps where each paper was returned and sorted, we read the abstracts of each potential new paper and only kept the papers that were a criteria match.

As a final step in building our sample of papers, we searched for the keyword "dashboard" in the conference proceedings of *Learning Analytics and Knowledge* and all issues of the *Journal of Learning Analytics*. These two publication venues were chosen for this final pass because they publish LAD research and represent our targeted audience. We broadened the search term from "higher education dashboard" to "dashboard", because the results returned were small enough to allow us to examine each paper that was returned. After all searches had been completed, The final list of publications was comprised of the initial 20 articles found by searching google scholar and scopus with the

keyword higher education dashboard, and 15 articles from LAK and JLA with the keyword dashboard. Removing duplicates, applying our three inclusion criteria, and conducting a forward and backward citation search we arrived at 45 relevant articles and 4 literature reviews (27 articles were excluded). Table 2.1 displays the publication outlets for the papers included in this review (full list of papers on OSF: `https://osf.io/tg5bn/`).

2.3 Thematic Analysis and Findings

We conducted a thematic analysis over the final sample of 45 relevant papers. To create our themes, we skimmed the papers again and reviewed the notes we had taken for each to develop an initial idea of the overarching themes. At this stage, we developed three broad themes regarding methods, dashboard usage, and software development. We then thoroughly reviewed each paper taking detailed notes on how it relates to the broad themes above. We accomplished this by using a template, where each theme was listed along with room to give an overall summary of how a paper relates to the theme with accompanying quotations. Although we searched for each theme in every paper, not all papers contained information on all themes. Therefore, the themes were continuously modified, redefined, and split out into additional themes during this process to accommodate the evidence provided by the papers.

Four themes emerged from the thematic analysis: Participant Identities and Researcher Positionality, Surveillance Concerns, Implicit Pedagogies, and Software Development Resources. In the following sections, each will be defined, summarized based on the evidence from papers in our sample, and related to

Table 2.1: Number of the articles considered in the literature review by publication venue.

Publication Venue	**Num. of Articles**	**Publication Venue**	**Num. of Articles**
International Conference on Learning Analytics & Knowledge	6	Educational Technology and Society	1
Computers & Education	5	Higher Education	1
Assessment & Evaluation in Higher Education	3	Innovations in Education and Teaching International	1
CHI Conference on Human Factors in Computing Systems	3	International Conference on Information and Communication Technology (ICoICT)	1
Journal of Learning Analytics	3	International Conference on Learning and Collaboration Technologies	1
British Journal of Educational Technology	2	International Journal of Emerging Technologies in Learning (iJET)	1
Computers in Human Behavior	2	Journal of Computing in Higher Education	1
IEEE Transactions on Learning Technologies	2	Journal of Educational Technology Systems	1
Learning @ Scale	2	Journal of Research in Innovative Teaching & Learning	1
Technology, Knowledge and Learning	2	Teaching in Higher Education	1
Asia Pacific Education Review	1	The International Journal of Information and Learning Technology	1
Behaviour & Information Technology	1	The Internet and Higher Education	1
BMC Medical Education	1		

the broader issue of incorporating JEDI into LAD research. We additionally summarize the themes and note related challenges and opportunities in Table 2.2.

2.3.1 Participant Identities and Researcher Positionality

JEDI-informed research needs to understand who was involved in the research (both researchers and participants) and how the inclusion or exclusion of people is reflected in the study findings and general implications. Even when a study does not have a JEDI focus, reporting simple statistics about the population can advance a collective understanding in the field about which groups might not be represented in the research. The socio-technical nature of dashboard research means that different methodologies can generate complementary insights. For this theme, we chose to organize the sampled studies base on their methodological approach to explore how identity information is presented.

We observe a strong methodological skew towards surveys and interviews in LAD research, which has also been noted in prior LAD research [Matcha et al., 2020, Verbert et al., 2013b]. We found many of the sampled studies applied multiple methods, with some studies using both qualitative and quantitative methods [Aguilar et al., 2021, Atif et al., 2020, Bodily et al., 2018, Broos et al., 2020, Brown, 2020, Gutiérrez et al., 2020, Herodotou et al., 2020, Millecamp et al., 2018]. These studies mostly reported demographics related to sex, male/female percentages or numbers. Three of the studies that reported sex, only reported a number for females, thus suggesting that sex is binary, and the rest were male. In studies that relied on participants with expert knowledge, age and/or experience were also reported. One of the studies did report a percentage of "under-represented minority groups," but it was unclear what identities were included in this group. While most of these studies addressed inter-rater reliability for the coding of the qualitative portions of the study, none of them reported information about the coders themselves to determine if they

were similar or dissimilar to each other and to the participants.

Other studies employed interviews or focus groups where a dashboard was presented to people to elicit their opinions and suggestions [Echeverria et al., 2018a, Howell et al., 2018, Klein et al., 2019, Lim et al., 2020, Roberts et al., 2017, Sun et al., 2019, Wise and Jung, 2019, Zheng et al., 2021]. Like the previous studies, these studies mainly reported participant sex. However, all studies in this group reported numbers for all sexes instead of just one number and assuming a binary distinction. More frequently than sex, participants' experience and age were reported. In the case of studies focused on teachers, teaching experience was reported. In the case of student-focused studies, year in school or age was reported. These differences may be attributed to the smaller number of participants in interviews and focus groups.

In other studies, the researchers were able to observe how participants interacted with the dashboards in the wild by analyzing log data from the dashboard [Broos et al., 2017b,a, Foster and Siddle, 2020, Kim et al., 2016]. It was not surprising that most of these studies did not include any participant demographic information since log data tend to have limited user information. One study did present socio-economic status in their dataset and reported results based on this indicator. Another study conducted at a "women's university" stated that their sample was therefore "100% female." This conflation of sex and gender leading to, incorrectly interchanging sex for gender, was present in almost all papers in the sample.

Although major funding agencies like the Institute of Education Sciences (IES), the research arm of the US Department of Education, is prioritizing randomized controlled experiments to answer education policy questions, such as

"what works, what doesn't,"[Spybrook et al., 2016, IES, 2021], only five studies in our sample used an experimental design. Two of those studies were conducted in a live course where a random sample of students was granted access to a dashboard [Aljohani et al., 2019, Hellings and Haelermans, 2020] and reported the general information about the course, but not demographic information about the students. Another two studies conducted a within-subjects experiment in a live course with students granted access to a dashboard in some but not other weeks [Amarasinghe et al., 2020, Han et al., 2021]; and one experimental lab study in which participants evaluated four different dashboard conditions [Lim et al., 2019]. The former study reported just the sex of the participants, while the latter reported both sex and age. Controlled experiments are an essential methodological tool to demonstrate the effectiveness of dashboards that have been deployed into university environments. At the same time, the studies included in this sample provided little information about the study participants, and none of them provided a demographic breakdown by experimental condition.

As research results have been used to justify and advocate for policy changes, this exclusion could exacerbate societal issues for minoritized groups who are not sufficiently represented in the research. Interviews and focus groups have been shown to amplify the voices of minoritized participants more effectively than quantitative methods, but there is a risk that smaller samples omit voices from marginalized groups [Griffin et al., 2011]. This issue is exacerbated when a colorblind approach is taken to data analysis by not accounting for or addressing participants' demographics in the study. This lack of data implies a narrative that all participants are the same and reinforces the norms associated with those most privileged in a context [Collins, 2015]. As a field, the inclusion

of demographic data can help other researchers understand which communities or contexts must be investigated to understand the boundaries of theories and frameworks, and to prevent potentially harmful policies from being deployed in contexts that the research evidence would not support.

In addition to missing participant demographic information, we did not find researchers positioning themselves within the research. By positioning, we mean reflections from the researcher about how their experiences and identities may impact their research from study design to the interpretation of results [Milner, 2007]. Nevertheless, it was encouraging to see numerous studies, typically qualitative studies, explicitly state their epistemology in the study context, and we hope this continues across methodological disciplines.

2.3.2 Surveillance Concerns

The large amount of data that LADs use to generate visualizations has sparked critical conversations about privacy and ethics of learning analytics and educational data mining. Regardless of whether research studies have an ethics or privacy goal, consideration of the ethical implications of their studies is important. This theme is grounded in the privacy and ethical concerns brought up by study participants. In our sample, there were six studies where privacy and ethical concerns were brought up by participants even though these concerns were not being studied [Brown, 2020, Heath and Leinonen, 2016, Howell et al., 2018, Roberts et al., 2017, Sun et al., 2019, Wise and Jung, 2019]. Although issues of surveillance were not central to these studies, it was the first time some of the participants became aware that their institutions were mining their data. While

the data collection and mining was happening independently from and probably well before the intended research, participants still linked the potential of surveillance and lack of privacy to the research project.

In some studies, participants were cautious of how the display of the data could impact individual privacy. Participants in Roberts and colleagues' study [Roberts et al., 2017], who were students at the university, were concerned that dashboard comparison features with other students could reduce their own privacy. Participants wanted the comparison features, but also wanted anonymity, which was possible for this particular research. In other studies, the research prompted ethical questions such as should students have the ability to completely remove themselves from the collection or display, instead of remaining anonymous [Heath and Leinonen, 2016]? This dilemma is currently being addressed at numerous institutions, weighing the risk of individual privacy with the learning benefits that can only be gleaned by full participation in the data. Some of the instructor-focused studies reported that faculty were also worried about their students' privacy [Howell et al., 2018, Sun et al., 2019, Wise and Jung, 2019]. These faculty noted that many students were unaware of the data mechanisms of the university and that information should be provided to students about data collection [Sun et al., 2019]. Other studies took this idea a step further, acknowledging the power relationship between instructors and students and suggesting that data could make this relationship more oppositional if instructors used the dashboard data as facts or surveillance against presumed future student behavior [Han et al., 2021, Wise and Jung, 2019].

Although privacy concerns were not the main focus of the dashboard studies, since in most settings they were using existing data infrastructure, these

concerns came up in student interviews and focus groups. Not only were there concerns about student privacy, but there were also concerns from faculty about surveillance of their courses using data displayed in the dashboards. Brown [Brown, 2020] found that faculty were seeing the data collection as "unwelcome surveillance" of their teaching practices. They felt it was unclear who had access to the data and what decisions were being made with them. In one extreme example, a faculty member decided to remove all LMS data from their course dashboard, which limited the amount of course insights that the dashboard was designed to provide. Other faculty members, just like the students, expressed that their data used in predictive modeling should be anonymized. A remarkable feature of many of Brown's [Brown, 2020] observations was that most of the concerns extended beyond the dashboard. For example, a faculty member can prevent LMS data from showing up in their dashboard, but the university still has access to that data for modeling and data-informed decision making. While faculty and students expressed privacy concerns, the same concerns did not arise in advisor-focused studies. This could have been because early educational data mining research focused on early alert systems that were created to help advisors reach at-risk students. Thus, many advisors were already aware of the data collection and analysis that are happening at their institutions. Dashboards are sometimes students, faculty, and staff's first encounter with the large data systems at their institution, even though the dashboards merely display the data and do not collect it.

At its core, this theme reflects a privacy concern with implications for all aspects of learning analytics, including LAD research. From the perspective of JEDI, we identify a need for LAD studies to be transparent, use accessible language, and thoughtfully consider all decisions that have to be made by re-

searchers, institutions, staff, instructors, and/or students around how data will be visualized in dashboards. Studies in this literature review exemplify the numerous decisions that must be made when conducting research about data, including but not limited to: who has access to what data; who has access to compare data; what data should be displayed; how should individuals process and use the data; and when can an individual remove themselves from the data. As researchers, so many of these choices have become automatic or predetermined by academic institutions. Even if changes cannot be made, we should still interrogate what those decisions mean for JEDI in the research. Take for example Wise and Jung's [Wise and Jung, 2019] paper, they suggest future research should develop a new dashboard view that anonymizes student information to the instructor. In testing these strategies, it is critical to consider how this change might impact JEDI both individually and at an institutional level. Some might argue that hiding the student information can foster equal treatment of all students in the class, while others might argue that student learning is a function of individual student experiences, including their social identities, which should therefore be visible to instructors.

2.3.3 Implicit Pedagogies

Another theme that emerged is resistance by faculty, and sometimes administrators, to accept dashboard systems because the design does not align with their individual pedagogies. This theme is echoed in many papers in the learning analytics community as a missed opportunity to design educational technologies with pedagogy in mind [Castañeda and Selwyn, 2018]. Multiple studies found that even with helpful dashboard insights, experienced participants still relied

on their own pedagogy to address issues or discrepancies when their interpretation of the data did not align with their own pedagogy [Gutiérrez et al., 2020, Wise and Jung, 2019, Zheng et al., 2021]. In our sample, we identified studies that focused on integration issues of faculty and advisor pedagogy, and studies that explored how dashboards could be designed with a focus on pedagogy.

Faculty concerns about pedagogy were grounded in the fear that incorporating LADs into teaching might result in extra work for them [Atif et al., 2020, Brown, 2020, Howell et al., 2018]. This fear is not unfounded, as one of the current issues of learning analytics is the lack of uniformity in data. One strategy to unify the data would be to enforce data standards. For example, if an institution was looking to design a dashboard from LMS data, the designers would need some level of assurance that they could pull consistent data from multiple courses. If one course uses modules to organize course content (e.g., assignments, quizzes) but another course uses pages to organize course content, it becomes challenging to design one dashboard to display the same information for the instructors of both courses. Instead, a course design policy would need to be implemented to choose one of these options as a standard and some instructors would need to adjust their course design and/or pedagogy in order to use the dashboard [Herodotou et al., 2020].

Additional workload was mentioned by Wise and Jung [Wise and Jung, 2019] as a reason why they did not conduct more interviews with their faculty participants: they thought more interviews throughout the semester would be a burden on the faculty in addition to modifying their courses to use a dashboard. In Howell and colleagues [Howell et al., 2018], the faculty acknowledged that these decisions might need to be made by the university and com-

promise was possible, but also felt that faculty should have a seat at the table where these design decisions are being made. In some studies, faculty not only were dealing with the added workload, but also failed to see how the insights from a dashboard could be used to inform their teaching practices [Brown, 2020, Herodotou et al., 2020]. Wise and Jung [Wise and Jung, 2019] suggested that faculty might not be able to use dashboards for teaching practices, because they could find the insights incongruent with their observations outside of the dashboard, partly due to the time it takes for the data in the dashboard to update. In these cases, the faculty may lose trust in the dashboard. But not all changes to teaching practices were considered bad, in fact some studies pointed out the opportunity for dashboards to initiate a reflection process for faculty about their pedagogy [Herodotou et al., 2020, Wise and Jung, 2019]. Using dashboards in this way could allow faculty to identify ineffective assignments or help them to better adapt their teaching to particular students. These studies show the importance of both incorporating teaching pedagogy into the design of teacher dashboards and using the dashboards as a reflective tool for pedagogy.

While most mentions of pedagogy were directly related to faculty, some papers addressed the pedagogical issues of using dashboards for advising [Gutiérrez et al., 2020, Millecamp et al., 2018]. Gutiérrez et al. [2020] found that different types of advisors had differing levels of adoption of an advising dashboard. In their study, they compared advising "done by professionals: i.e., trained academic advisers" [Gutiérrez et al., 2020, p. 11] to advising done by faculty. They found that faculty advisors were more likely to trust the dashboard and underlying model as compared to their professional advising counterparts. This difference highlights an issue of pedagogy, because to faculty the LAD was a tool to help them with a secondary responsibility, whereas professional advi-

sors felt their expertise/pedagogy was not fully leveraged by using the LAD. In another study concerning advisor's behaviors with dashboards, Millecamp et al. [2018] looked at how their dashboard could support advisors meeting with students. They found that advisors typically interacted with the dashboard at the beginning of the meeting to understand a student's situation, but as the meeting progressed, the advisors relied less on the dashboard and more on their pedagogy for helping students. This level of interaction may be sufficient, but it raises the question of whether it is possible to design an advisor-centered dashboard that is useful for the entire advising meeting.

Some of the papers in the sample designed a dashboard to incorporate pedagogy [Atif et al., 2020, Echeverria et al., 2018b]. Echeverria et al. [2018b] set out to understand how a dashboard could be designed with pedagogy as an input in the design process. The result was a dashboard that allowed instructors to customize visualization rules to match their own pedagogy. Unfortunately, these customizations take time to program and additional training would need to be provided to instructors. So it has yet to be determined if this is a feasible solution. While less customized than the previous example, Atif and colleagues [Atif et al., 2020] found that instructors were willing to put in extra hours to initially configure a system in the hopes that they would be able to deliver a better learning experience to students. The authors cautioned future research to understand the behavior of how and when instructors tweak configurations in order to make future designs more useful.

The emergent issue in this theme is tied to the formal power of faculty, derived from their status in the institution [Tatum, 2000]. JEDI-conscious researchers should therefore ask themselves: How might this LAD reinforce

and/or support the dominant pedagogy? How does the choice of research questions and design reinforce and/or support the dominant pedagogy? Lastly, what losses can result from forcing individuals into the dominant pedagogy or leaving individuals out of the process? These are hard questions to answer, but grappling with them can yield benefits for LAD research. Looking deeper at Wise and Jung's [Wise and Jung, 2019] study, one of their findings highlights that instructors were unwilling to adopt a LAD if its insights contradicted their own knowledge or experience. This dissonance could just be one of many messages signaling to a minoritized instructor that they are wrong, while a non-minoritized instructor may dismiss the LAD insight without questioning themselves. This raises a critical question of how one can design and research a system that proves useful to both instructors without disregarding their experience or knowledge.

2.3.4 Software Development Resources

The fourth theme that arose from the sample of papers was the scarcity of commercialized or open-source software used to create dashboards. The majority of papers used homegrown dashboards built either by the researchers or in cooperation with their institution's IT departments [Aguilar et al., 2021, Aljohani et al., 2019, Amarasinghe et al., 2020, Atif et al., 2020, Bodily et al., 2018, Broos et al., 2020, 2017b,a, Echeverria et al., 2018b, Gutiérrez et al., 2020, Han et al., 2021, Hellings and Haelermans, 2020, Herodotou et al., 2020, Kim et al., 2016, Lim et al., 2019, Millecamp et al., 2018, Sun et al., 2019, Wise and Jung, 2019]. Foster and Siddle [2020] contracted their dashboard development to a company and were able to request modifications as a consultant to the company's dash-

board software. Brown [Brown, 2020] did not indicate the source of their software, but given the study spanned more than one university, it was likely a commercial software.

There are two sides to the question of whether researchers should use homegrown or commercial software. Homegrown software gives the research team more flexibility to design a usable dashboard for their context along with avoiding adding to the increased commercialization of higher education [Castañeda and Selwyn, 2018]. However, the homegrown customizations which are benefits in one study context may not transfer well to another institution or study without extensive technical work. This was evidenced by Guerra and colleagues [Guerra et al., 2020], who used the same base implementation of an advisor dashboard across three different institutions. Each institution underwent their own software development to customize the dashboard to their needs. The creation of dashboards is resource intensive in terms of financial and human capital, potentially creating a barrier for researchers trying to conduct LAD research. While commercialized software tends to be expensive, the results and findings from studies using commercialized software may have more potential to be transferable to multiple contexts.

Another option for creating dashboard software that can be scaled is to use open-source software. Only two papers in the sample utilized open-source software to construct their dashboards [Klein et al., 2019, Lim et al., 2020]. Another two open-source software papers were identified in the article search process, but they did not include empirical studies [Leitner and Ebner, 2017, Cobos et al., 2016]. Implementing an open-source dashboard can be resource intensive, as evidenced by the technology stack and setup of Leitner and Ebner [2017] open-

source dashboard software paper. Another study created open-source software to be used with edX and Open edX courses [Cobos et al., 2016]. Where there is no correct correct answer for the type of dashboard software to use for LADs, the financial and technical resources required to conduct this type of work deserve awareness and discussion in the field.

There was also a geographical trend in efforts to create LADs, with a bias towards countries in the global north. The availability of software development resources not only influences how data will be displayed, but also what data that is used in a dashboard. Although study sites spanned across all continents except Antarctica, there still was limited globalization in terms of the number of countries represented. Figure 2.2 illustrates this point about the global distribution of LAD research. Relying on mostly individualistic countries to set forth what data are important to visualize could have serious consequences for deploying these dashboards in other cultures [LaBrie et al., 2018]. We are not the first to note this bias against the global south; other education researchers have highlighted the same issue in regards to educational research [Kross and Guo, 2018, Tuti et al., 2020]. With most research sites being located in western cultures, eastern perspectives may be unwittingly excluded in dashboard narratives. This could result in problems if dashboards are deployed at institutions without an evaluation of their local effects. Further research that investigates these issues can further our understanding of the globalization of LADs.

Collaboration can serve as an approach to address JEDI by combining software development resources. Combining resources, through open-source code or formal partnerships, can enable scientists without the necessary resources to conduct LAD research. This type of collaboration is already happening in the

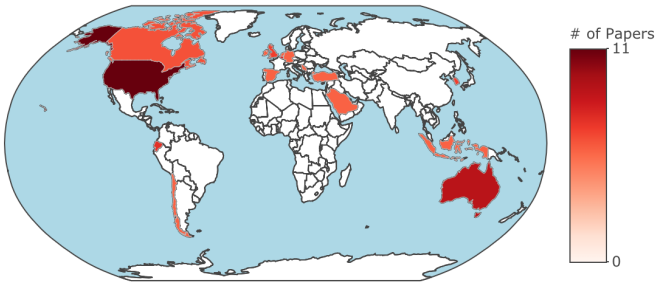


Figure 2.2: LAD research papers by the country where the study was conducted. The majority of LADs are developed and researched in North America, Europe, and Australia.

LAD literature: Guerra et al. [2020] and Gutiérrez et al. [2020] deployed dashboards in both European and Latin American countries and added partnerships with research teams from multiple countries. Researchers or institutions who have the resources to undertake LAD development are encouraged to provide open access to the software from the start. While open-source code is a first step, we also acknowledge the significant resources needed to maintain these dashboards. As evident in the sample of papers for this review, many research teams had the support and backing of their institutions. In the near future, this institutional dependency will need to be addressed as a barrier to research progress.

2.4 Discussion

We conducted a critical review of LAD research focusing on how the literature has engaged with issues of justice, equity, diversity, and inclusion in higher education. We have highlighted four major themes in how research in the broader learning analytics and educational data science community has engaged with these topics thus far. Now we discuss our findings and propose future research directions.

2.4.1 Current LAD Research for Justice, Equity, Diversity, and Inclusion

Researchers have made substantial efforts to advance our understanding of how to develop LADs in higher education. This presents an opportunity to build upon this body of knowledge to strategically use LADs to improve JEDI in higher education. When we posed RQ1 at the beginning of this investigation, we expected to find and report on LAD research that focused on improving JEDI. Specifically, we expected to present one, if not multiple, themes dedicated to unpacking JEDI in LAD research. However, none of our themes addressed JEDI in LADs, because except for two studies, JEDI concepts were not addressed in any of the papers. Each study had the opportunity to address JEDI, given that all of these studies exist in contexts with issues of power and injustices that have led to groups being underrepresented or marginalized [Collins, 2015]. Foster and Siddle [2020] initially considered the use of demographic data, but then they removed all demographic information, except for socioeconomic status, after discussions with their university community raised concerns that these indicators could stereotype students. While critically examining demographics can foster additional concerns, we risk allowing existing inequities to proliferate unfettered if researchers do not pursue opportunities for deeper investigation. In contrast, an exemplar paper that considered JEDI principles, even though the study was not primarily about JEDI, is Li et al. [2021]. They studied a dashboard that granted instructors the ability to compare student behaviors across predefined subgroups. The purpose of the LAD for this study was not to create a Diversity Dashboard or a dashboard that caters to people with JEDI-specific questions. Instead, this study incorporated JEDI principles by providing a LAD

with the flexibility to use it to visualize and uncover disparities in course contexts. This work exemplifies the opportunity to address JEDI questions in research that is not framed as a paper on JEDI.

2.4.2 Maintaining Systemic Inequities in LAD Research

In recent years, a common institutional strategy to address social inequity has been to create Diversity, Equity, and Inclusion initiatives and offices [Patton et al., 2019]. Yet these offices are often siloed and unwittingly contribute to a narrative that diversity-related work is only done by individuals working in these offices or for initiatives created by these offices. This problem is further complicated by the gap between those who research education and those who practice education in the day-to-day. This gap between those that conduct research and those that enact the research into practice has contributed to maintaining and exacerbating historical inequities in higher education.

However, LAD research is unique in that it interweaves multiple contexts, uniquely placing the research team in a position to affect the research design and their use of tools in educational contexts. Our thematic analysis addressed RQ2 by highlighting researcher practices that can contribute to maintaining systemic inequities in LAD research. In particular, the sampling and description of participant populations in studies, as well as unspecified researcher positionality; student and faculty-based concerns about surveillance that often remain unaddressed; the implied behavior changes required for user adoption of LADs; and the abundance of resources needed to conduct LAD research. We hope researchers will critically examine their research to understand how their research

practices are related to the themes we identified and may contribute to creating or maintaining inequities.

2.4.3 Future Work

Lastly, we address answers to RQ3 by highlighting opportunities within LAD research to improve JEDI in higher education. We call attention to the following directions for future research that can help to improve LADs.

Shared Software Resources and Cross-Border Collaborations

As we enter a new phase in higher education, where students use technology to study virtually across borders [Joyner et al., 2020], LAD researchers have an opportunity to also cross borders to connect and collaborate. This connection could include partnerships that gain access to new populations or sharing resources to develop new LADs. When researchers collaborate to develop LADs, additional data can be incorporated, and more opportunities for improved practices and pedagogies are possible. Researchers looking to engage in this type of research should look at the LALA project and the lessons learned from this large-scale dashboard collaboration [Hilliger et al., 2020]. Two of the studies based on this project were included in this review. Moreover, Hilliger et al. [2020] present more case studies of this collaboration along with lessons learned from their cross-border collaboration.

Improved LAD Design and Usability

To advance the design and usability of LADs, we encourage more perspectives from users and learning theories to be taken into account. This also means that projects should also consider making LAD development more accessible to researchers and institutions, and adopting more open-source software. Research conducted by Echeverria et al. [2018b] and Atif et al. [2020] are exemplars for designing LADs with pedagogy as a design requirement. Echeverria et al. [2018b] had instructors create rules related to teaching pedagogy, such as participation in the class discussion board. An example dashboard rule related to this type of participation highlights students who have posted less than a minimum number of postings. To incorporate JEDI into this rule, the LAD could convey to the instructor available demographic information that the highlighted students have in common. This type of design that allows the end-user to create rules based on their pedagogies or needs will also allow for these rules to have JEDI extensions that can highlight previously unknown inequities.

Studying LADs for JEDI

By tackling challenging issues of JEDI, researchers can model how to embed meaningful use of demographic data into dashboards. For instance, Aguilar [2020] studied a summer bridge program with a high proportion of "underrepresented minorities", but they missed an opportunity to examine or discuss how students' demographic characteristics could have been embedded into their dashboard. At a time when institutions are grappling with how to identify and reduce systematic inequities on their campuses, our research community has an opportunity to take up the call to study how to design dashboards that can ad-

vance this goal. Taking advantage of the current momentum to advance social justice, we encourage more LAD research to focus on this critical topic or at least critically engage with the implications of LADs for JEDI. Researchers interested in creating dashboards that center the experiences of historically marginalized students may find inspiration in the Equity Scoreboard project [Harris and Bensimon, 2007, Bensimon, 2004]. The project "combines a theoretical framework with practical strategies to initiate institutional change that will lead to equitable outcomes for students of color" [for Urban Education, 2021], and at the end of the process, a dashboard is created to show context-specific metrics for long-term evaluation of initiatives. While this project concentrates on macro-level institutional data, the practices from this project can be adapted to more granular, course-level data, which are more representative of LAD data.

2.4.4 LAD Research that Does Not Focus on JEDI

Every research study sets out to address a set of research questions and this does not require a focus on JEDI. One purpose of this review is to continue the conversations started by many organizations via "Call for Actions" statements. SOLAR's Statement of Support and Call for Action in 2020 included such a call to their research community: *"We encourage members of our Society to mobilise our expertise and connections with communities to actively contribute to the hard work of promoting social justice and dismantling injustices in education. [...] It is our duty to educate ourselves and to focus more actively on how to create an equitable and just environment for all academics, and people our work impacts, free from racial discrimination."* [SoLAR, 2020]

Our interpretation of SOLAR's call to action is to examine how all research, regardless of its research questions, can actively help dismantle injustices in education. Previous research has documented the issues that can occur when JEDI is merely a symbolic notion without active prioritization [Patel, 2015, Patton et al., 2019]. Our goal in highlighting specific studies as having missed opportunities to engage with JEDI is not to suggest that the research questions in these papers should change, but to indicate places where small intentional changes could actively help dismantle injustices in education.

2.4.5 Limitations

This study is a critical review and therefore subject to the limitations of a critical review. It cannot be compared or used like a systematic literature review because we used a sample of papers addressing a specific concern. This limitation does not invalidate a critical review [Paré et al., 2015]. Moreover, a critical review is dependent on the researchers and their experiences when creating themes. As a US-born doctoral student with multiple US underrepresented and minoritized identities and a European-born professor now living in the US, our results were constructed through these lenses. Once again, this does not invalidate the review, but it allows our readers to further contextualize the results for their purposes.

2.5 Conclusion

We sought to understand the potentially significant benefits of LADs are being leveraged to improve JEDI in higher education. Through a critical literature review, we identified four areas in LAD research where changes could be made to improve outcomes for justice, equity, diversity, and inclusion. Our findings support one conclusion in particular: there is a need to incorporate JEDI research into LAD studies. There are many ways, identified here through our themes, for researchers to incorporate JEDI into their studies, and without having to change the focus of their scientific inquiry. They range from the simple addition of details about participants so that policies based on research studies are applied in appropriate contexts, to the harder effort of recruiting more diverse participants early in the research to understand pedagogical issues and needs of the faculty, staff, and students using the dashboards. Another more complex recommendation, but one that can be strategically underwritten by funding agencies, is to create cross-cultural and cross-boundary collaborations that allow LADs and LAD theories to be tested in multiple contexts. Most of these directions for future work can be addressed immediately and the list is certainly not exhaustive. We are keen to see our field use JEDI principles to promote justice, equity, diversity, and inclusion in LADs and, more generally, in learning analytics.

Table 2.2: Four core themes identified in this critical literature review. Each theme is summarized by providing a description along with challenges and opportunities for justice, equity, diversity, and inclusion.

Theme	Description	Challenges	Opportunities
Participant Identities and Researcher Positionality	The disclosure of participant and research identities in studies.	Protecting participants' privacy while also including enough demographic information so that context can be better determined to create policies based on research.	LAD researchers should collect and report demographic information from participants. LAD researchers should also reflect on their identities and experiences and how they influence their research studies.
Surveillance Concerns	Conflation of dashboard research with larger learning analytics privacy and surveillance issues.	Decisions about data access and visualizations are made by LAD researchers, and these decisions have consequences for how users make meaning of the dashboards.	Researchers should be transparent about all decisions made in the research, including ones they consider to be implied. Explaining these decision will give researchers an opportunity to interrogate the choices they make that could have negative impacts.
Implicit Pedagogies	The need for incorporating pedagogy into LAD research.	LADs have been created with the goal of supporting learning and instruction in a scalable way, but they are designed with certain values and user pedagogies in mind. This can neglect pedagogies that fall outside of the dominant narrative.	LAD researchers can design dashboards to be accessible to different pedagogies by making the dashboards more customizable, and use outcome measures that reflect the varying goals of instructors or advisors.
Software Development Resources	The development of LADs for research is a resource-intensive process where a large share of the development happens just a few countries.	The financial resources and software development expertise required to develop and deploy LADs, as well as the need for close relationships with institutional IT offices, make LAD research inaccessible to many researchers.	Making dashboard software open-source can significantly reduce the upfront costs of creating a LAD for research and foster research collaboration across institutions and borders, which can expand the reach of LAD research to more global contexts.

CHAPTER 3

SCALABLE MEASURES OF COURSE EFFECTIVENESS, EQUITY, AND INCLUSIVENESS

While the preceding sections have highlighted the limitations of current LADs in addressing JEDI issues, it is evident that these tools hold significant potential to promote equity and inclusion in higher education. To fully harness the power of LADs, we must focus on developing innovative metrics that can provide a more comprehensive assessment of course-level equity and inclusivity. In the following section, we delve into the development of two new metrics: course equality and course inclusiveness. These metrics are designed to offer a scalable and holistic approach to identifying and addressing inequities in educational settings. By combining these metrics with the insights gained from our critical review of LAD research, we can pave the way for a future where LADs are powerful tools for advancing JEDI in higher education.

3.1 Introduction

Regular assessments and quality assurance are important practices for maintaining the credibility of educational institutions. For institutions with hundreds or even thousands of educational offerings each year, this requires a scalable approach to identify issues and allocate resources to address them. Educational triage, as defined by Prinsloo and Slade [2014], provides a framework to help answer the question of "how do we make moral decisions when resources are (increasingly) limited?" Resources like professional development are often scarce in education, and general large-scale training may not effectively address

critical issues. In institutions or school districts with many course offerings and classes, there is a need to scalably identify courses that fall short of meeting students' needs so that tailored support can be provided to the educators who need it the most, not just those who self-select into professional development programs.

Course grades have long been used to create simple and scalable metrics for course effectiveness, for example, by computing pass rates, or their inverse, fail rates, also known as DFW rates (i.e., the percentage of students earning D or F grades, or withdrawing from the course late) [Urtel, 2008]. However, there has not been an equivalent course-level metric to identify issues with equity and inclusion, even though educational institutions strive to provide equitable and inclusive learning environments to their students. Researchers and practitioners have been working on promoting equity and inclusion for years, focusing on increasing educational access, fostering curricular diversity, and closing achievement disparities [Tolossa et al., 2023]. Scalable methods for auditing educational environments and monitoring educational equity and inclusion are important to allocate scarce resources and assess the effectiveness of interventions. While existing tools like the equality, diversity, and inclusion (EDI) learning analytics dashboard [Bayer et al., 2024] and the Course Diversity Dashboard [Sloan-Lynch and Morse, 2024], are steps in the right direction, they have notable limitations, including their reliance on simple measures of performance gaps [Hubbard, 2024]. Developing scalable, reliable, and valid measures of EDI experiences and perceptions is essential for longitudinal efforts to address critical issues of equity and inclusion in education.

In this paper, we first review research on course evaluation methods and

metrics for capturing EDI in courses. We then introduce two new EDI metrics, course equality and course inclusiveness, which are designed to provide a holistic and scalable assessment of EDI. We demonstrate the usefulness of the new metrics through an experimental study, in which educators make data-driven decisions about professional development allocation. This research contributes new metrics for EDI efforts and inspires practical applications of these metrics.

3.2 Background

Recent research has explored various approaches to quantifying and promoting EDI in educational settings. Several studies have focused on developing metrics and frameworks to assess EDI in courses and academic departments [Ludwig, 2021, Nair et al., 2024, Khanuja et al., 2023]. These include the Inclusive Excellence Ratio [Ludwig, 2021] and the use of the Gini coefficient to measure equity [Khanuja et al., 2023]. Some studies have explored alternative approaches to evaluating equity beyond demographic identifiers [Pai, 2023] and proposed methods to improve fairness in course recommendation systems [Polyzou et al., 2021]. Overall, these papers highlight the growing importance of EDI analytics in education and the need for robust metrics and strategies to assess and promote EDI.

3.2.1 The DFW Rate

The DFW rate, which measures the percentage of students receiving a D, F, or W in a course, has been used to measure course effectiveness and an outcome

variable in higher education research [Urtel, 2008]. While studies have used the DFW rate to evaluate course redesigns and interventions [Colvard et al., 2018, Van Dusen and Nissen, 2020], it may miss serious issues related to inclusion and equity for underrepresented groups of students [Urtel, 2008, Colvard et al., 2018, Rahal and Zainuba, 2016]. For example, a course could have a low DFW rate, yet all first-generation college students failed the course. The DFW rate alone cannot reveal this disparity, but an equity-focused metric might.

Research using DFW rates to address EDI in higher education has shown mixed results. For example, Van Dusen and Nissen [2020] used DFW rates to evaluate their Learning Assistant intervention, a program where undergraduates who previously passed the course are used as peer educators. While the intervention showed an overall reduction in DFW rates, a more detailed analysis revealed that this reduction was skewed towards overrepresented groups, with a smaller decrease for underrepresented students. This study highlights the potential limitations of relying solely on DFW rates as a measure of equity and inclusion, as these rates may mask disparities among different student groups. Simply focusing on gateway courses with high DFW rates may not be the most effective use of EDI resources [Swan et al., 2018, Bloemer et al., 2017]. Instead, a more nuanced approach is recommended considering student characteristics and their grade level. Machine learning models have also been used to help identify at-risk students, but the low predictive value of demographic variables, along with the many additional predictors in those models, may still disguise inequities [Yang et al., 2020]. Within learning analytics, dashboard research has been used as a tool to visualize EDI issues. While past reviews of the literature have indicated that learning analytics dashboards have largely neglected EDI issues and potentially reinforced existing inequities [Williamson and Kizilcec,

2022], more recent research has centered dashboards as a space to highlight and focus on educational inequities and data [Bayer et al., 2024, Sloan-Lynch and Morse, 2024].

3.2.2 Student Evaluations of Teaching

Student Evaluations of Teaching (SETs) are surveys used to assess course effectiveness. They are widely used but have important limitations. Research has shown that SETs are not strongly correlated with student grades [Uttl et al., 2017] and are influenced by factors unrelated to teaching effectiveness, such as instructor accent, class size, and class meeting time. Studies show that SETs are biased against female instructors and faculty of color [Boring et al., 2016, Austin, 2020]. These biases can significantly impact hiring, promotion, and tenure decisions and contribute to the marginalization of underrepresented academics [Austin, 2020]. While there is ample research exploring students' biases in course evaluations towards faculty based on gender and race, not much work with SETs has focused on how students' experiences in classrooms vary along gender and racial dimensions [Sengupta et al., 2019, Fisher et al., 2019, Wallace et al., 2019]. Yet examining socio-demographic heterogeneity in the student experience is critical because it can highlight issues of equity and inclusion, even if certain groups of students make up a small proportion of the entire course. Overall, while SETs are widely used, scalable, and potentially useful for assessing EDI-related issues, the documented concerns about bias in this self-report measure need to be addressed first. We therefore focus on metrics that are grounded in self-report measures.

3.2.3 Measures of Course Equity

This section explores existing metrics that have been developed to measure equity in higher education. Specifically, we look at three types of equity metrics: the Equity Scorecard, Disproportionate Impact, and STEM Gendered Performance Differences.

The Equity Scorecard [Harris and Bensimon, 2007] created the groundwork for emphasizing the importance of measuring inequities in higher education to promote institutional change. The goal of the Equity Scorecard project was to create a tool that continually assesses and addresses equity gaps within institutions. The Equity Scorecard created an environment where university administrators across a university can come together as a team and ask equity-minded questions about their institutions. Once the questions had been agreed upon, the team then created metrics to evaluate and monitor these equity questions on their campus [Harris and Bensimon, 2007]. The metrics were campus-specific and revolved around the successful completion percentages for certain gateway courses [Harris and Bensimon, 2007]. While the Equity Scorecard excelled at improving equity at individual institutions, it did not provide a large-scale solution to identifying and quantifying inequities in higher education.

Matz and colleagues introduced new course metrics to understand the nature of gender-based achievement gaps in the context of introductory courses at five large research institutions [Matz et al., 2017]. Specifically, they propose the average grade anomaly (AGA) as a measure of disproportionate impact. It quantifies how much students perform better (or worse) in a given course relative to how they perform on average in all their other courses. Thus, a high AGA indicates that students do better in that course as compared to their other

courses. Building on the AGA metric, the study proposes the gendered performance difference (GPD) metric, which quantifies the gender effect in the AGA for a particular course while controlling for students' standardized test scores and their average grades in other courses. Using these metrics, Matz and colleagues found patterns of significant gender achievement gaps in STEM courses, especially when looking at the performance differences between labs and lectures. Two notable limitations of this approach are (1) controlling for ACT/SAT and prior grades can mask inequities that are perpetuated in these ability measures, and (2) the metrics are designed to examine only one binary demographic group at a time. Still, we took inspiration from their approach in the development of the metrics we propose here.

3.3 Development of Course Metrics

3.3.1 Course Effectiveness

While Course Effectiveness is not an EDI metric, it measures the overall successful operation of a course under the assumption that the goal is to have most students master the course content and score well enough on assessments to at least pass the course. Thus, Course Effectiveness is operationalized in terms of the percentage of students who receive a passing grade in a course (see Equation 3.1). It is essentially the inverse of the DFW Rate. It has a lower bound of 0 (no students passed the course) and an upper bound of 1 (everyone passed), and it is course enrollment size invariant.

$$\textbf{Effectiveness}_c = \frac{\textbf{\# Students with Passing Grades}_c}{\textbf{\# Enrolled Students}_c} \tag{3.1}$$

3.3.2 Course Equality

Course Equality is an aggregate measure of achievement gaps between multiple social groups in a given course. It is operationalized in terms of the percentage of variance explained (R^2) in student grades by multiple social group indicators. To quantify this for each course, we compute the R^2 of a linear regression model that predicts each student's final course grade with their sociodemographic characteristics as predictors (see Equation 3.2). The model can include any number of sociodemographic predictors available. For demonstration, we show gender, Pell-grant eligibility status, first-generation college student status (FG), and indicators for each of J ethno-racial groups. In our empirical evaluation study, we also included first-order interactions of the predictors in the model to account for the compounding impact of intersectionality in student identities (higher-order interactions could be included if feasible) [Griffin and Museus, 2011, Bowleg, 2008]. The final course grade was converted from a letter grade (A+, A, A-, etc.) to a numeric grade (4.33, 4, 3.67, etc.), such that grades ranged from 0 to 4.33 (withdrawals were excluded in this computation because they are accounted for in the Course Inclusiveness metric).

To ensure that underrepresented groups are not overlooked in the regression analysis, we employed a weighting method that assigned higher weights to observations from smaller groups (akin to oversampling observations). Specifically, we assign an unnormalized weight of $(10 * \sqrt{n^2 + 100} - 1)^{-1}$ to observations

with group size n. For example, in a course with 30 students and only two who are FG, each FG student would get a normalized weight of 0.0861 in the regression model (for a total of 0.1722), while each continuing-generation college student would receive a normalized weight of 0.0296 (for a total of .8288)—the two students represent 6.67% of the class but receive a combined regression weight of 17.22%. This reweighting approach helps mitigate the potential bias that can arise in imbalanced datasets. We encourage analysts to compare different weighting regimes based on the distribution of their local dataset.

$$\mathbf{grade}_i = \beta_0 + \beta_1 \mathbf{gender}_i + \beta_2 \mathbf{pell}_i + \beta_3 \mathbf{FG}_i + \sum_{j=4}^{J} \beta_j \mathbf{race}_{ij} + \epsilon_i \tag{3.2}$$

The R^2 value is computed by using Equation 3.3, where $\bar{grade}_c$ is the mean grade for course c, $grade_{ci}$ is student i's grade in course c, and $\hat{\epsilon}_{ic}$ is the model error for student i in course c. Due to the properties of the R^2 value, the Course Equality metric has an upper bound of 1 (i.e., sociodemographic characteristics do not explain any variation in grades) and a lower bound of 0 (i.e., sociodemographic characteristics explain all the variation in grades). The Course Equality metric is also course-size invariant.

$$\mathbf{Equality}_c = 1 - R_c^2 = 1 - \frac{\sum_i^n \hat{\epsilon}_{ic}^2}{\sum_i^n (grade_{ic} - \bar{grade}_c)} \tag{3.3}$$

Note that we intentionally use the term Equality instead of Equity, as we do not make an explicit judgment about the desired value for each student group. Instead, the Course Equality metric casts a larger net, ultimately capturing a more complete picture to identify courses with any inequities.

3.3.3 Course Inclusiveness

Course Inclusiveness is an aggregate measure of the diversity in sociodemographic composition at the beginning and later on in a course. Course Inclusiveness is operationalized using data on who was initially enrolled in the course and who later withdrew from the course.

First, to quantify diversity in initial enrollment, we compute the Normalized Generalized Variance (NGV) [Budescu and Budescu, 2012] of the sociodemographic makeup of students enrolled in the course. Intuitively, the NGV is the probability that two students are from the same sociodemographic groups. The NGV is comprised of the sums of the squared proportion (P) for each category (C). A category is a division within a grouping; for example, a grouping could be ethnicity, and the categories could be American Indian, Asian, Black, Hispanic, and White. For example, we can define a category as the combination of all sociodemographic indicators to account for the intersections of a student's identities [Griffin and Museus, 2011, Bowleg, 2008]. Then, a first-generation college student (1) who is not Pell-grant eligible (0) and who identifies as a Black (B) man (M) would have a combined category of "10BM", whereas a continuing generation college student who is also Pell-grant ineligible, and a White woman would have a combined category of "00WF".

To compute the NGV, we sum all of the proportions of categories for a given course and multiply by a normalizing factor to make course with varying numbers of categories (C) comparable (see Equation 3.4). The NGV has a lower bound of 0 (i.e., courses with no variance where one sociodemographic category makes up the entire class) and an upper bound of 1 (i.e., courses where sociodemographic groups are evenly distributed).

$$\mathbf{NGV}_c = \frac{C}{C-1}\sum_{i=1}^{C} P_i^2 \tag{3.4}$$

Second, we compute the percentage of variance explained (R^2) in student withdrawals from a course based on students' sociodemographic characteristics. For each course, we computed McFadden's R^2 of a logistic regression model that predicts whether a student withdrew from the course using the students' sociodemographic characteristics as predictors (see example model in Equation 3.5; in our empirical study, we also included first-order interactions between all predictors).

$$\mathbf{withdraw}_i = \beta_0 + \beta_1 \mathbf{gender}_i + \beta_2 \mathbf{pell}_i + \beta_3 \mathbf{FG}_i + \sum_{j=4}^{J} \beta_j \mathbf{race}_{ij} + \epsilon_i \tag{3.5}$$

McFadden's R^2 can be computed from the value of the maximum likelihood function for the full model shown in Equation 3.5, L_{Full}, and the value of the likelihood function for the intercept-only model, $L_{Intercept}$. Each value is transformed using the natural logarithm ($ln(.)$) before computing their ratio, as shown in Equation 3.6. The upper bound of this R^2 metric is 1 (i.e., sociodemographic characteristics do not explain any variation in course withdrawals) and a lower bound of 0 (i.e., sociodemographic characteristics explain all of the variation in course withdrawals).

$$\mathbf{R}_c^2 = 1 - \frac{\ln L_{Full}}{\ln L_{Intercept}} \tag{3.6}$$

We compute the Course Inclusiveness metric by combining the NGV and R^2 measures using the harmonic mean to obtain an average of two rates (see

Equation 3.7).

$$\textbf{Inclusiveness}_c = \frac{2\textbf{NGV}_c(1 - \textbf{R}_c^2)}{\textbf{NGV}_c + (1 - \textbf{R}_c^2)} \tag{3.7}$$

3.3.4 Simulation

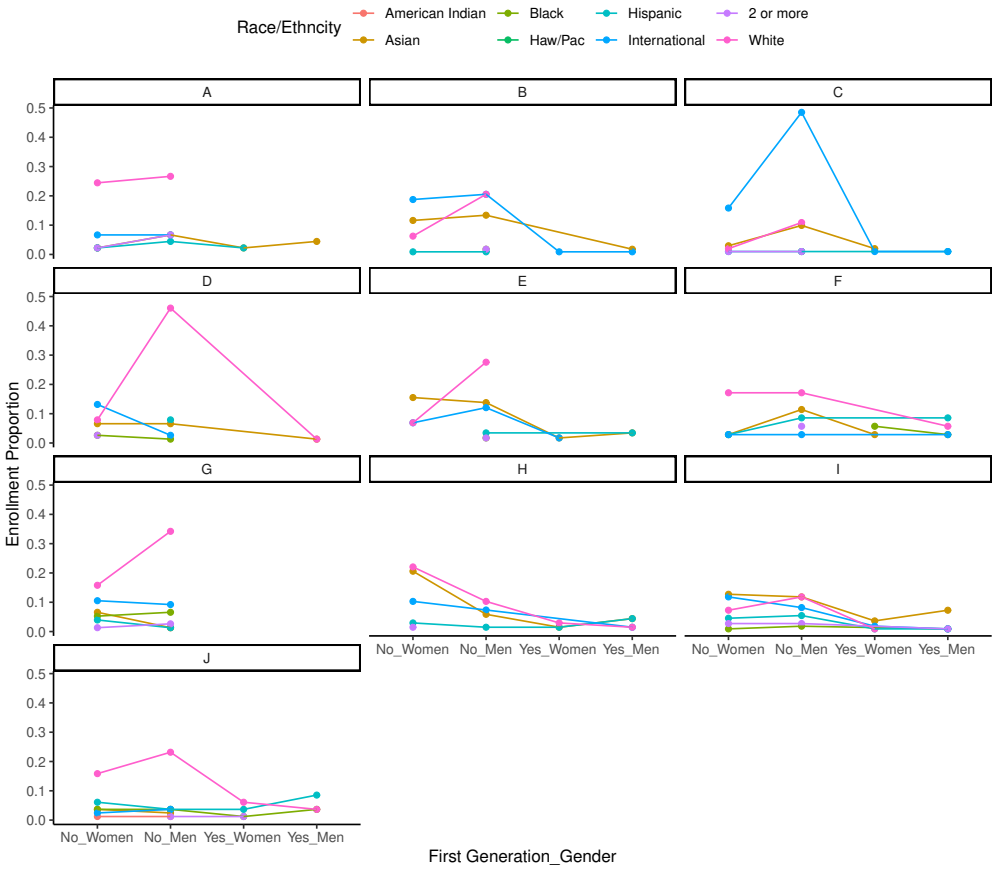


Figure 3.1: The enrollment proportions of the 10 simulated courses, broken down by face/ethnicity, gender and first generation status.

We conducted simulations to evaluate the behavior of the Course Equality and Inclusiveness metrics under various conditions. Specifically, we explored how these metrics change based on course composition, performance

gaps, and course size. We simulated 10 courses with diverse compositions and grades. Figure 3.1 illustrates the course compositions. Courses H and I exhibited balanced enrollment across race/ethnicity, gender, and first-generation status, while courses C and D demonstrated unbalanced proportions. Table 3.1 presents the resulting Course Inclusiveness scores. As expected, courses C and D had low inclusiveness scores, reflecting their unbalanced compositions. Course I, despite appearing balanced, had a lower-than-expected score due to withdrawal patterns.

Course	Inclusiveness	Equality	Effectiveness
A	0.95	0.48	1.00
B	0.96	0.60	1.00
C	0.28	0.54	0.99
D	0.54	0.77	1.00
E	0.96	0.66	1.00
F	0.96	0.38	0.88
G	0.94	0.14	0.98
H	0.96	0.74	1.00
I	0.76	0.68	0.89
J	0.59	0.66	0.96

Table 3.1: The course Inclusiveness, Equality, and Effectiveness results for the 10 simulated courses. For each metric, the lowest values have been highlighted in pink.

To assess the impact of performance gaps on course equality, we analyzed student grades aggregated by race/ethnicity, gender, and first-generation status. As depicted in Figure 3.2, courses with significant disparities between different groups indicate inequality and lower Equality scores. For instance, Course A exhibited a clear pattern of lower grades among Non-First Generation women, particularly Hispanic or multiracial students. Conversely, Course D demonstrated relatively equal grades across demographics, resulting in a higher Equality score. Table 3.1 further supports these findings, showing that Course A indeed has a lower Equality score compared to Course D. Additionally, the ta-

ble highlights how courses can have high pass rates but still exhibit issues with inclusiveness and equality, as demonstrated by Course A.

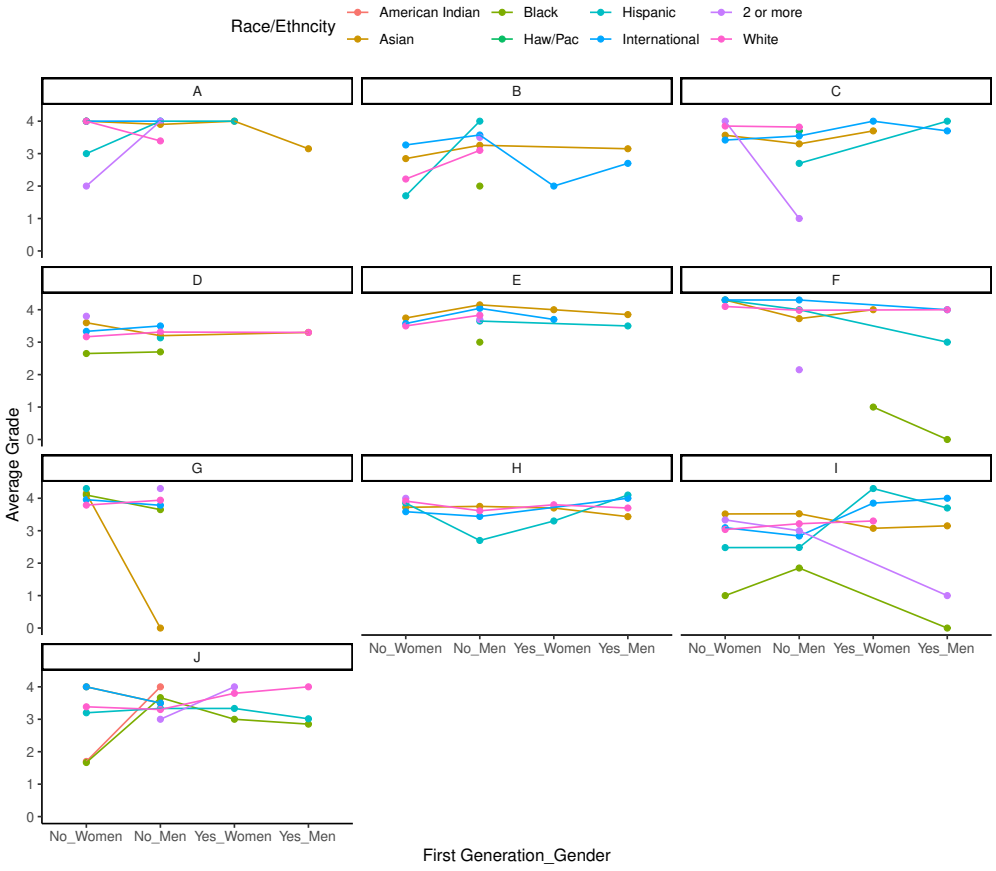


Figure 3.2: The student grades of the 10 simulated courses, broken down by face/ethnicity, gender and first generation status.

To investigate the impact of course size on the metrics, we simulated doubling and quadrupling the size of the 10 courses while maintaining their original composition and grade distributions. By carefully controlling the variables, we aimed to isolate the effect of course size. Our analysis revealed that most courses maintained relatively consistent metric scores, regardless of size (Figure 3.3). However, some courses, such as Course F, experienced notable changes. Course F initially had a relatively high Inclusiveness score. However, when doubled and quadrupled, it experienced a significant decrease in the score.

This shift can be attributed to the influence of withdrawal patterns. In the original simulated dataset, a single withdrawal had minimal impact on the Inclusiveness metric. However, as the course size increased, the relative impact of a single withdrawal became more pronounced. In particular, the withdrawal of a student from a minority group significantly impacted the Inclusiveness score. This finding underscores the importance of considering withdrawal patterns when evaluating course inclusiveness, especially in larger courses.

3.3.5 Data Example

We applied all three metrics to a real-world dataset comprising enrollment and grade data from undergraduate courses at a large private university in the Northeast United States. Figures 3.4 and 3.5 illustrate the distribution of Equality and Inclusiveness scores, respectively, compared to the conventional effectiveness metric. While many courses exhibited high effectiveness scores (passing rates above 85%), the Equality and Inclusiveness scores varied significantly, indicating potential disparities in student outcomes.

To delve deeper into the nuances of these metrics, we selected two courses with notable differences between effectiveness, equality, and inclusiveness scores. Table 3.2 presents a detailed analysis of these courses. In the case of course A, despite a high effectiveness score (98.8% pass rate), the low Equality score revealed significant disparities in grade distributions across different demographic groups. A closer examination showed that a disproportionate number of students from historically underrepresented racial groups failed the course. This highlights the limitations of relying solely on overall pass rates and

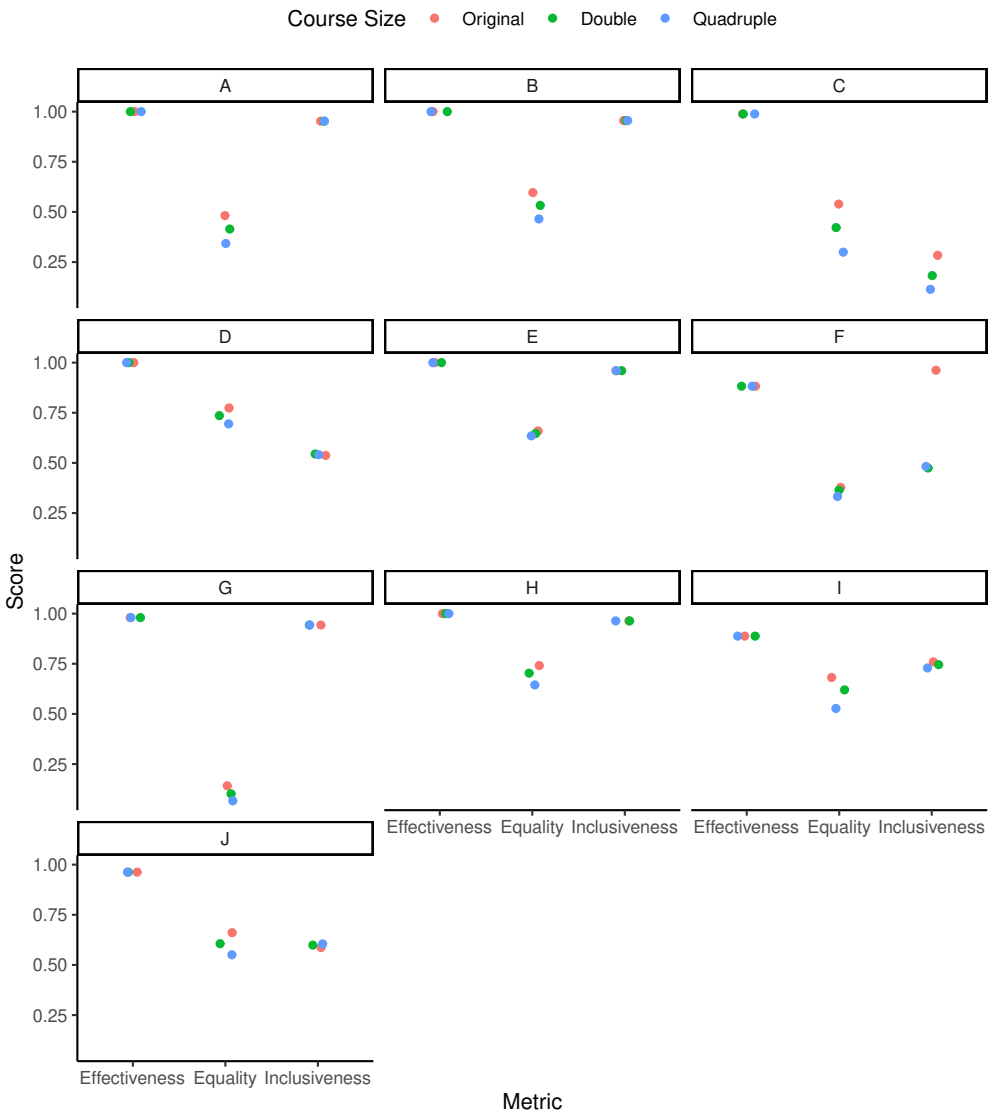


Figure 3.3: For each simulated course, the resulting metric scores of doubling and quadrupling the simulated courses.

the importance of considering intersectional factors.

For course B, this course exhibited a high effectiveness score but a low Inclusiveness score, indicating a lack of diversity in the student body. Figure 3.7 further illustrates this imbalance across gender and race/ethnicity. This finding underscores the need to consider both student outcomes and representation when evaluating course quality. These case studies demonstrate the value of

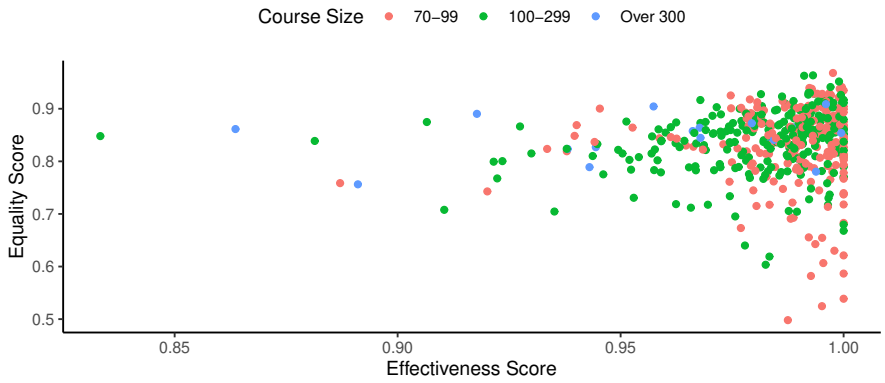


Figure 3.4: 565 courses showing a weak relationship, $r = -0.006 (p = 0.762)$ between the effectiveness and equality scores.

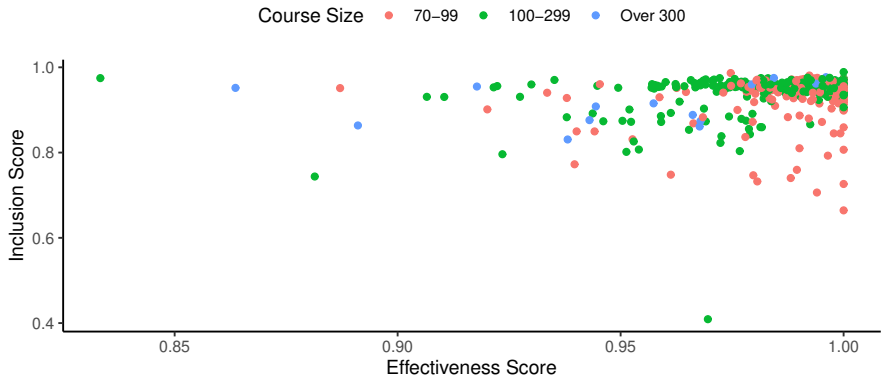


Figure 3.5: 565 courses showing a weak relationship, $r = 0.066 (p = 0.001)$ between the effectiveness and inclusiveness scores.

Course	Effectiveness	Equality	Inclusiveness
A	0.988	0.498	0.883
B	1	0.683	0.664

Table 3.2: Table showing the Effectiveness, Equality, and Inclusiveness scores for 2 sample courses.

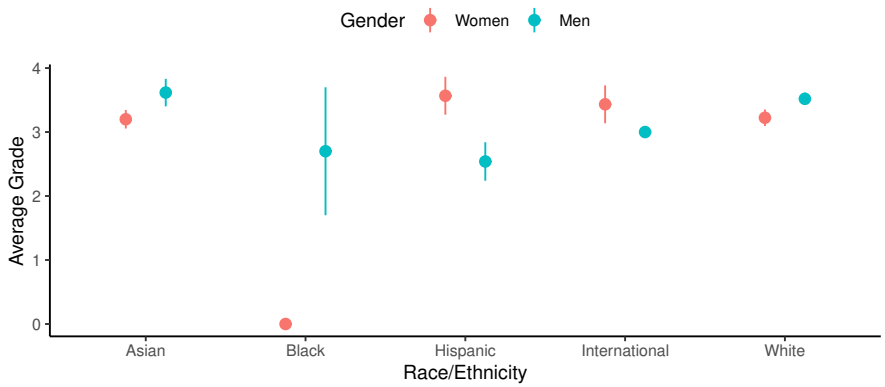


Figure 3.6: The grade distributions for Course A group by Race/Ethnicity and Gender.

the new metrics in uncovering hidden disparities and providing a more comprehensive assessment of course quality. These metrics offer a more nuanced understanding of equality and inclusiveness in higher education by considering factors such as course composition, performance gaps, and withdrawal patterns.

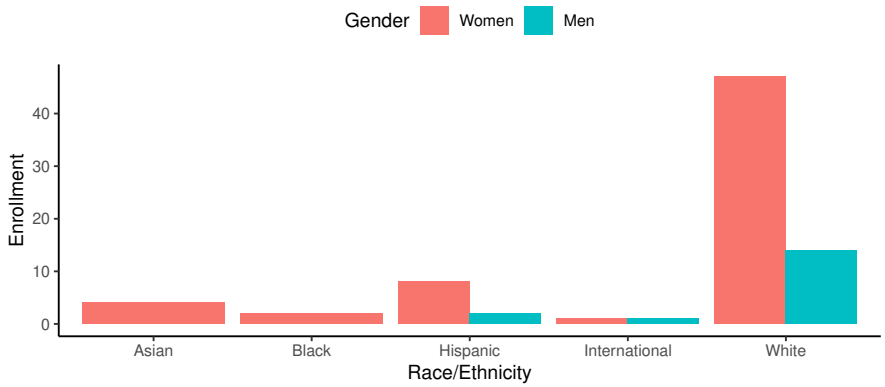


Figure 3.7: The class composition for Course B group by Race/Ethnicity and Gender.

3.4 Empirical User Evaluation

To empirically evaluate the usefulness of the Course Equality and Inclusiveness metrics, we conducted an experimental study with educators to understand the usefulness of these metrics. Specifically, our goal is to understand (1) how educators would use the metrics to make decisions and (2) how much educators express confidence and trust in using these metrics. We use a scenario-based design in which participants were shown a table of courses with metrics for course. We simulated course data to appear realistic (varying the number of students, socio-demographic composition, and grades) and computed metrics as defined above. Participants were asked to recommend specific courses for two different programs using only the data provided to them to make these decisions. Our focus is on how the Equality and Inclusiveness metrics influence participants' ability to make empirically-grounded decisions to recommend courses for an Inclusive Design program (described as helping educators improve equality and inclusion in their courses). As a point of comparison, we also ask participants to recommend courses for an Instructional Design program, which is described as unrelated to equity or inclusion.

We preregistered[1] the following hypotheses and research questions which address the usage of the new metrics.

- H1: In the treatment condition, educators are more likely to select courses with lower Equality and Inclusiveness scores for the Inclusive Design program than the Instructional Design program.
- H2: The treatment increases the likelihood of selecting courses with lower

[1]Our preregistered report included identifiable information, so we created a link with a copy of the non-identifiable information.

Equality and Inclusiveness scores for the Inclusive Design program.

- RQ1: How does the treatment change the characteristics of courses selected for the (a) Instructional Design program and (b) Inclusive Design Program?

In addition, we preregistered the following hypotheses and research questions which address educators' trust and confidence using the new metrics.

- H3: In the control condition, educators have lower (a) confidence and (b) trust in selecting courses for the Inclusive Design program than the Instructional Design program.
- H4: The treatment increases educators' (a) confidence and (b) trust in selecting courses for the Inclusive Design program.
- H5: The treatment reduces the difference in educators' (a) confidence and (b) trust in selecting courses for the Inclusive vs. Instructional Design program (by increasing educators' (a) confidence and (b) trust for the Inclusive Design Program relative to the Instructional Design program).
- RQ2: How does the treatment change educators' (a) confidence and (b) trust in selecting courses for the Instructional Design program?

3.4.1 Methods

The study was preregistered on OSF. All code, data, and supplementary materials for this study are openly available on OSF.

Participants

We recruited a sample of 250 educators residing in the United States through Prolific, an online platform for research participation. We a Participants received $2.75 for completing an online survey. Following our preregistered exclusions, data from participants who took longer than 25 minutes to complete the task were excluded (n = 8). The excluded participants' completion times ranged from 26 to 59 minutes. This resulted in a final sample size of N = 242 educators, 72% of them were K-12 teachers and 28% instructors in higher education settings. Participants had been teaching for 11+ years (47%), followed by 2-5 years (25%), followed by 6-11 years (24%), followed by 4% who taught for less than a year. The majority identified with feminine pronouns (67%), followed by masculine pronouns (29%), followed by non-binary pronouns(2%). 2% of participants preferred not to answer or selected "Not Listed.". The majority of participants identified as White (86%), followed by Black or African American (7%), followed by Asian (5%), American Indian or Alaska Native (2%). 7% of participants indicated "Not Listed."

Procedure

Participants first identified their primary teaching domain (K-12 or Higher Education). They then received a scenario relevant to their chosen level, explaining the context of resource allocation based on class data. For example, K-12 educators read, "We are studying a new way to allocate resources to middle and high school classes. Teachers are sometimes asked to review classes in another school district. This review process may lead to additional resources being supplied to teachers who teach classes that are deemed less successful than others."

Following the scenario, participants were introduced to two support programs:

- **Instructional Design:** This program focuses on developing the knowledge and skills to create effective learning experiences in general. This includes reviewing learning objectives, instructional strategies, assessment methods, and course development tools.
- **Inclusive Design:** This program focuses on designing learning experiences that cater to a diverse range of learners and reducing grade gaps. This includes understanding accessibility needs, incorporating a representative curriculum, and fostering a welcoming learning environment.

Participants were then randomly assigned into one of two conditions: Traditional Metrics (n = 125) or Enhanced Metrics (n = 117). The only difference between the two conditions was the course information provided to participants to make PD recommendations. All participants were informed that they would review data and recommend classes for PD programs. They were then presented with a table of courses with metrics to make recommendations (they were presented with the following definitions):

- **# of Students:** The number of students in the class.
- **Average Grade:** The average grade out of 100.
- **Effectiveness Score:** The percentage of students who received a passing grade ('C' or better).

Participants assigned to the Enhanced Metrics condition received two additional metrics presented with the following definitions:[2]

[2]We renamed the Course Inclusiveness metric to Composition Score in the study design to

- **Equality Score:** What percentage of the variation in student grades is NOT explained by student attributes, such as their gender, lunch program eligibility, and IEP (Individualized Educational Plan). A low Equality Score indicates that grades vary systematically by student attributes (e.g., gender achievement gaps). A high Equality Score indicates that grades are independent of these student attributes. For example, if eligible for a class lunch program (lower income) students receive grades that are 20 points lower on average than lunch program ineligible (higher income) students, then this class would have a lower Equality Score compared to a class where students from both groups did equally well.
- **Composition Score:** How diverse are the students who took the class (in terms of gender, lunch program eligibility, and IEP (Individualized Educational Plan), etc.). A low Composition Score indicates a low diversity of students in the class. A high Composition score indicates high diversity. For example, a class that only has lunch program ineligible students will have a lower composition score than a class that has an equal balance of lunch program eligible and ineligible students.

Immediately after participants saw each definition for the metrics above, they were asked a comprehension question that assessed their understanding of the metric. For example, participants were asked which of the following statements best described the Effectiveness Score: (a) The effectiveness score measures the likeability by students. (b) The effectiveness score measures how many students passed. (c) The effectiveness score measures how many students missed classes. If participants chose b, they moved on in the survey; otherwise,

reduce demand effects for recommendations for a program that is called "Inclusive Design". The presented definition still reflects the same metric.

participants were given another opportunity to respond with the correct answer. We discontinued the survey for participants who were unable to provide the correct answer after two attempts.

Following comprehension checks, participants were presented with a table containing 26 simulated courses and their associated metric values. The simulated dataset was generated using the following steps. First, we initialized 26 courses (labeled a to z) and assigned 5,000 students to a randomly selected course (course size: mean = 192.308, sd = 90.805).[3] Then, we randomly assigned each student independently a binary first-generation label (Bernoulli p=0.2), a binary low-SES label (Bernoulli p=0.3), and a binary failing grade label (Bernoulli p=0.3). Depending on the failing grade label, we sampled a numeric grade for the student using a uniform distribution (between 25 and 50 if failing grade is true, between 85 and 100 if failing grade is false). Finally, we computed course metrics based these data and made adjustments to five courses to lower grades for FG and low-SES students.

Participants in the Enhanced Metrics condition saw all five metrics, while those in the Traditional Metrics condition did not see the equality and composition scores. Figure 3.8 shows a portion of the table presented to participants to aid them in making recommendations. Definitions for each column in the table were provided above the table in case participants required clarification. Participants reviewed the table and recommended classes for the Instructional Design and Inclusive Design programs. The order of program recommendations was counterbalanced to reduce order effects. Participants then answered a set of survey questions for each program and then answered two open-ended

[3]We adjusted course size down by a factor of 10 for the table presented to K-12 educators in our study.

Class	# Students	Average Grade	Effectiveness Score	Equality Score	Composition Score
a	9	67.55	65.91	41.81	77.65
b	9	71.69	60.92	93.73	83.57
c	25	69.27	58.06	98.36	80.32
d	26	62.8	60.23	62.2	75.84
e	11	73.24	65.14	98.47	85.49
f	20	70.14	59.8	98.3	83.9
g	12	70.51	61.02	98.19	78.46
h	30	71.83	62.46	99.56	81.67
i	10	74.02	70	70.01	36
j	11	68.72	58.18	91.85	86.66
k	16	75.31	68.29	98.54	80.1

Figure 3.8: A portion of the table that participants were presented with that showed the 26 courses along with the metrics. Participants could click the table headers for each column to sort the rows in ascending or descending order.

questions about their decision-making process, and finally a few demographic questions.

Dependent Measures

Course Recommendations. Participants reviewed a list of all 26 courses and used checkboxes to select the five they would recommend for each program (Instructional Design and Inclusive Design). We calculated the average value for each metric (# of Students, Average Grade, Effectiveness Score, Equality Score, and Composition Score) across the five chosen courses for each program. This yielded an average score for each metric of each program for each participant.

Confidence. Confidence in their recommendation was measured with a four-item scale. Participants rated their level of agreement with four statements: ”I had enough information to make an informed decision”; ”I have confidence in the recommendation I gave”; I doubt that the courses I selected are good candidates for this program”; and ”There simply was too little information about the courses for this decision.” Each question was rated on a 5-point Likert scale (Strongly Disagree=1, Disagree, Neutral, Agree, Strongly Agree=5). The average score across the four items provided a single confidence score for each

participant for each program (Cronbach's alpha = 0.84).

Trust. Trust in the recommendation process was assessed using the question: "How much do you trust the analytic process (i.e., relying on course-specific measures) to recommend courses for the [Instructional Design or Inclusive Design] program?" The participants responded on a 5-point Likert scale (Not at all=1, A little, A moderate amount, A lot, A great deal=5).

Exploratory Analyses

To gain deeper insights into participants' decision-making processes, we included two open-ended questions: Recommendation Process: "How did you make your recommendations?" and Desired Information: "What further information would you have liked to see to make recommendations for each program?" Content analysis was employed to categorize responses from the Recommendation Process open-ended question. Responses were examined for the presence of specific metrics. For instance, a response stating, 'For the inclusive program I chose classes with low average grades. For the instructional program I chose classes with low average grades and low effectiveness scores,' would be assigned the following codes: Average Grade - Inclusion, Average Grade - Instructional, and Effectiveness - Instructional.

Thematic analysis was conducted to identify and categorize patterns within the open-ended responses to the Desired Information question. Responses were examined for recurring themes, and similar data needs were grouped accordingly. For example, responses expressing a desire for state assessment data or more information about course assessments were categorized under the theme

'Assessment Data.' Responses indicating a need for longitudinal data, such as multiple years of course performance, were grouped under the theme 'Longitudinal.' Responses indicating that participants had all the necessary data were assigned to the category 'Nothing Else.'

3.4.2 Results

Behavioral Effects

We examine the characteristics of recommended courses to evaluate if educators understood the new course metrics as intended (H1) and how their availability affects educators' course recommendations (H2). Figure 3.9 illustrates the average equality and composition score of courses recommended for each program and in each experimental condition.

First, we examine the pattern of course recommendations among participants in the enhanced metrics condition. We fitted a linear mixed-effects regression model for each outcome measure (equality score and composition score) to estimate the difference in course characteristics for courses recommended to the Inclusive Design program relative to the Instructional Design program.[4] As hypothesized (H1), we found that educators in the Enhanced Metrics condition chose courses for the Inclusive Design program with lower equality scores ($b = -5.420, t_{116} = -4.024, p < 0.001$) and lower composition scores ($b = -1.363, t_{116} = -2.011, p = 0.046$) than they chose for the Instructional Design program. This finding confirms that educators use the additional infor-

[4]We fitted a mixed-effects model because we have two observations for each participant derived from the two sets of course recommendations for the Inclusive and Instructional Design programs.

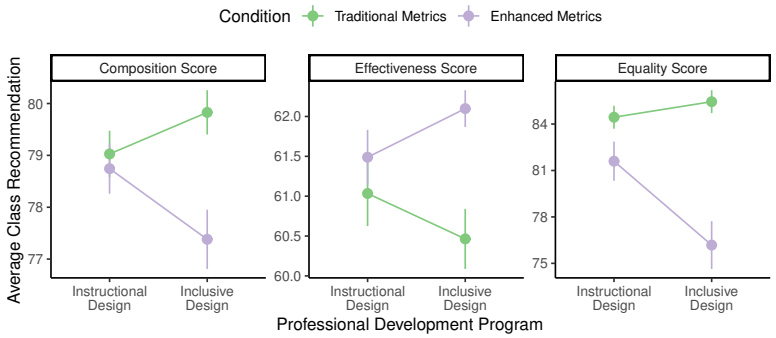


Figure 3.9: Average Effectiveness score, Equality score, and Composition score for each program by experimental condition.

mation to make decisions consistent with our design intentions.

Second, we examine the causal effect of providing the Equality and Composition Scores by comparing the selected courses between the experimental conditions. We focus the analysis on recommendations for the Inclusive Design program for which the new metrics should provide relevant information. We used a linear regression model with robust standard errors to estimate the effect of Enhanced Metric availability on the average equality score and composition score of selected courses. As hypothesized (H2), we found that presenting the new course metrics caused educators to recommend more courses with lower equality scores ($b = -9.266, t_{240} = -5.417, p < 0.001$) and lower composition scores ($b = -2.448, t_{240} = -3.433, p < 0.001$).

While our focus is on how the new course metrics affected course recommendations in terms of their equality and composition scores, we also examine effects on the composition of courses along other dimensions that we did not target (RQ1). Specifically, for the set of recommended courses for each program, we observe the average number of students, average grade, and average effectiveness score. Using the same linear regression approach, we ex-

amined how the availability of enhanced metrics influenced the characteristics of courses that educators recommended for each program along these dimensions. For the Instructional Design program, we found that the recommended courses had similar effectiveness scores ($b = 0.454, t_{240} = 0.851, p = 0.395$), but similar numbers of students ($b = -1.924, t_{240} = -0.195, p = 0.846$) and average grades ($b = -0.2638, t_{240} = -0.621, p = 0.536$). For the Inclusive Design program, we found that the recommended courses had higher effectiveness scores ($b = 1.634, t_{240} = 3.699, p < 0.001$), but similar numbers of students ($b = 0.116, t_{240} = 0.012, p = 0.991$) and average grades ($b = -0.433, t_{240} = -1.161, p = 0.247$). Overall, we find that the unintended effect on the effectiveness score is smaller than the desired effects on the equality and composition scores, as seen in Figure 3.9[5].

Attitudinal Effects

We examine educators' trust and confidence in their recommendations to evaluate if educators without access to enhanced metrics feel lower levels of confidence and trust in their decision-making for the Inclusive Design program than the Instruction Design program (H3). We then examine how the availability of enhanced metrics affected educators' confidence and trust in their recommendations for the Inclusive Design program (H4). Figure 3.10 illustrates the average level of trust and confidence in the course recommendations for each program and in each experimental condition.

First, we examine patterns of trust and confidence in course recommendations among educators in the traditional metrics condition. As before, we fitted

[5]The different y-axis shows similar gaps, but the magnitude of the Effectiveness Score is smaller in comparison to Equality and Composition Score.

a linear mixed-effects regression model for each outcome measure (confidence and trust) to estimate the difference in course recommendation attitudes to the Inclusive Design program relative to the Instructional Design program. As hypothesized (H3), we found that educators in the Traditional Metrics condition had lower confidence ($b = -0.159, t_{123} = -2.765, p = 0.007$) and lower trust ($b = -0.200, t_{124} = -3.215, p = 0.002$) in their course recommendations for the Inclusive Design program than they did for the Instructional Design program. This finding confirms that traditional course metrics provide insufficient support for making decisions concerning equity and inclusion.

Second, we examine the causal effect of providing enhanced course metrics by comparing confidence and trust between the experimental conditions. We focus the analysis on recommendations for the Inclusive Design program, for which the new metrics should provide relevant information and are expected to confidence and trust (H4). Again, we used a linear regression model with robust standard errors to estimate the effect of enhanced metric availability on educators' confidence and trust in the recommendations. As hypothesized (H4), we found that presenting the new course metrics caused educators to have more confidence ($b = 0.576, t_{239} = 5.394, p < 0.001$) and more trust ($b = 0.436, t_{240} = 3.761, p < 0.001$) in their decisions.

Third, based on the assumption that educators in the traditional metrics condition have more confidence and trust in their recommendations for the Instructional Design program than the Inclusive Design program, we hypothesized (H5) that the enhanced metrics would close these gaps in trust and confidence between programs. Consistent with this hypothesis, we found that the attitudinal gap between programs was closed and even reversed in terms of confidence

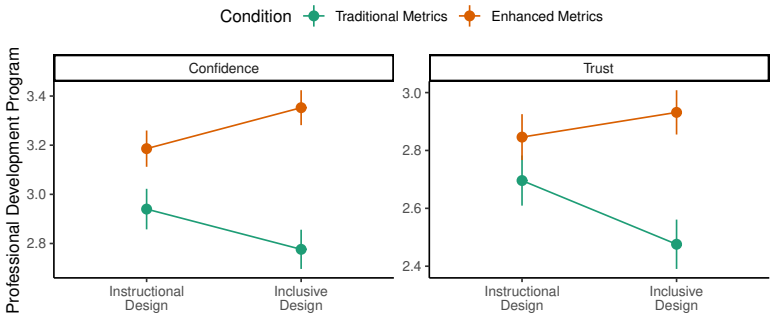


Figure 3.10: Average ratings of educators' confidence and trust in their decisions for each program by experimental condition.

($b = -0.324, t_{239} = -3.964, p < 0.001$) and trust ($b = -0.286, t_{240} = -3.512, p < 0.001$). Figure 3.10 highlights this result, with the slopes in the Traditional Metrics condition being negative, but the slopes turning positive in the Enhanced Metrics condition.

While we focused on how the new course metrics affected educators' confidence and trust in course recommendations in the Inclusive Design program, we also examined the effects for the Instructional Design program (RQ2). Using the same linear regression approach, we examined how the availability of enhanced metrics affected educators' confidence and trust in recommending courses for the Instructional Design program. We found that the enhanced metrics increased educators' confidence ($b = 0.246, t_{240} = 2.221, p = 0.027$) but not significantly their trust ($b = 0.150, t_{240} = 1.279, p = 0.202$). Overall, we find that the side effect on confidence for the Instructional Design program is notably smaller than the intended effect on confidence in the Inclusive Design program, as shown in Figure 3.10.

Exploratory Analysis of Open-Ended Responses

We systematically analyzed responses to two open-ended questions to better understand (1) educators' process of making recommendations based on the course metrics provided and (2) what additional information they would desire to make recommendations. These analyses are exploratory and can offer a more nuanced understanding of the decision-making process than the qualitative data alone.

Educators in the traditional metrics condition mentioned class size significantly more often than those in the enhanced metrics condition, who mentioned it the least. Effectiveness was a key consideration for those in the traditional metrics condition, with over half of the educators mentioning it as part of their decision-making process. In contrast, in the enhanced metrics condition, Effectiveness was mentioned at a similar rate as Equality, Composition, and Average Grade. As expected, the equality and composition scores were not mentioned in the traditional metrics condition due to unfamiliarity. Figure 3.11 visually represents the findings from the content analysis.

Overall, educators in the traditional metrics condition used significantly more words explaining their decision-making process (M=32.782, SD=24.366) compared to educators in the enhanced metrics condition (M=27.267, SD=17.563; $t_{238} = -2.021, p = 0.044$). This difference may be due to less descriptive text or more streamlined processes. The more streamlined processes in the enhanced metrics may be attributed to the availability of more applicable metrics. Instead of relying on multiple calculations, as many in the traditional metrics did, educators in the enhanced metrics used single metrics to make their decisions. For example, an educator in the traditional metrics condition de-

Figure 3.11: Percentage of educators who mentioned each metric (Average Grade, Class Size, Composition, Effectiveness, and Equality) in their decision-making process by experimental condition.

scribed their process as follows:

> For inclusive design, first I prioritized enrollment numbers. Anything over 150 was prioritized, and then I looked at the lowest average grades and effectiveness scores. These large courses are more likely to have varieties of learning styles and engagement levels at play, and engaging people in different ways will probably raise the average and effectiveness score...

This explanation demonstrates that the educator used multiple metrics and heuristics to make their recommendation. In contrast, an educator in the enhanced metrics condition wrote: "I started with the [classes] with the lowest scores on the related measures (e.g., equality score for the inclusive design question), then from the lowest 5 I looked for any reasons not to include them..."

This example demonstrates the reliance on the equality and composition metrics by the treatment group educator.

Second, we conducted a thematic analysis of open-ended responses about desired data which yielded 17 themes. We consolidated themes with insuffi-

cient data into twelve more comprehensive themes. All themes with definitions are available on OSF. Educators in both conditions, particularly those in the traditional metrics condition, expressed a need for more curriculum-related data, such as subject, teaching level, difficulty, and modality. For example, one educator noted, "I'd like to know what these courses were even about, to get a sense of how difficult material might be impacting metrics."

Information on accommodations, disabilities, and IEP/504 Plans was another frequently citepd data point. Although not explicitly mentioned in the experimental scenario, our new metrics could be scaled to include attributes related to accommodations. Both groups also sought longitudinal data to understand whether low scores were isolated incidents or a recurring pattern. As one educator stated, "I would be interested to see if the scores would change significantly between years and semesters."

Some themes were more prevalent in one condition than the other: demographic data was mentioned more often in the traditional metrics condition (98%), presumably because they did not have access to equality and composition metrics. These educators sought information such as "Statistics on gender and ethnic representation for students in various courses" or "DFW rates by demographic." However, even those with access to enhanced metrics desired more granular demographic data. For example, one educator responded, "I would have liked to have seen some breakdowns of exact differences. I wouldn't have minded seeing low income, first generation, minority status of these courses..." Educators in both conditions also wanted more detailed class composition data, along with information about the instructor's experience and general school data. Overall, educators in the enhanced metrics condition (69%) were more

likely than those in the traditional metrics condition (31%) to indicate that they had all the necessary data to make decisions.

3.5 Discussion

In this paper, we developed two new scalable course-level metrics designed to support EDI-related resource allocation and institutional research. We conducted a user-centered evaluation of these metrics by testing them with a sample of educators using a resource allocation scenario study. Our results demonstrate that educators exposed to the new metrics found them useful enough to inform their decision making. They also felt more confident and expressed higher trust in their decisions compared to educators who had access to only traditional course metrics. This finding was further supported by our qualitative data, which showed that educators provided more details about their process and the additional data they wished they had. Educators who only saw traditional metrics desired demographic data to help them make decisions and several participants mentioned they wished they had some diversity score or rating. Conversely, educators who saw enhanced metrics were more likely to indicate that they did not need additional data and that the metrics provided were adequate for the decision-making task.

Educators in both the quantitative and qualitative data relied solely on the effectiveness score when making decisions, highlighting its perceived value. In contrast, those in the enhanced metrics condition utilized effectiveness, composition, and equality scores. Despite our efforts to create scalable metrics, participants noted several limitations of the metrics. In the enhanced metrics con-

dition, educators still wanted to see more granular data regarding demographics. The practical solution might involve a combination of course demographics with global metrics that aggregate across them to triage courses with potential problems from a large set of courses. The metrics we propose are flexible to accommodate any set of sociodemographic, cultural, and other background factors. Still, there appears to be a need for more granular data to evaluate selected courses, understand issues more deeply, and develop concrete recommendations. In the sections below, we continue unpacking these findings further and conclude with future avenues for research.

3.5.1 The Right Data for the Task

EDI decisions require EDI metrics. Educators not equipped with EDI-related analytics could not make informed, confident, and trustworthy decisions related to EDI. This is not an education-specific issue; holistic EDI metrics have been called for when evaluating small businesses [Bruni et al., 2023]. Tang et al. [2023] has also called for more decision science research on EDI's economic and social impacts. In education, Nair et al. [2024] has suggested the need for contextualizable institution-wide metrics to advance EDI initiatives. These studies collectively emphasize the need for data-driven approaches, inclusive decision-making, and ongoing efforts to promote EDI across various sectors. Our study serves to propose a set of metrics and demonstrate the need and usability of these metrics in an educational context.

We also evaluated the role of using the right data to help improve trust and confidence in decision making. Prior work suggests that educators' confidence

and trust play crucial roles in their use of data for decision making. Teachers often need help with data interpretation and graph literacy, which can be influenced by experience and confidence [Oslund et al., 2021]. This underlines the need to make sure any EDI-related metrics are relatable and understandable to educators and validate their experience. A few educators saw the enhanced metrics and felt they wanted to use those metrics but needed help understanding the definitions, or it was too much information at once. Applications using these metrics must pay attention to how these metrics are introduced, ensuring the descriptions also help instill confidence. One area for future improvement would be to more explicitly compare the equality and inclusive metrics. While we constructed them to detect differing signals of EDI, we did not include that information in our experiment. This could help alleviate concerns about potential manipulation. For example, an instructor might limit student enrollment to improve their equality score, but this could negatively impact course inclusiveness. By explicitly comparing these metrics, we can better understand their interrelationships and potential for unintended consequences. We also have to be careful not to erase or devalue educators' experiences and options by the use of data-driven metrics. Performative accountability and data-driven logic have eroded trust in teachers' professional judgment [Daliri-Ngametua et al., 2022]. In our sample, a few educators wanted to know the teachers' perception of those courses. All of this provides evidence that these metrics are just one part of the solution. Overall, the effective use of data in education requires addressing issues of trust, confidence, and collaborative development to support educators in making informed decisions.

3.5.2 Course Equality and Inclusiveness Metrics in Practice

As promising as our results are in suggesting that our metrics are generally useful, we also need to acknowledge that they alone cannot fix issues associated with EDI. As mentioned in the metrics development section, the Course Equality metric does not directly measure equity. Instead, we chose a scalable measure that should still identify courses with inequities, but additional analysis is needed to investigate courses that score low on the Equality and Inclusiveness metrics to fully understand what is happening in those courses. Educational triage has been emphasized as a method for prioritizing limited resources [Prinsloo and Slade, 2014] and classification of courses using our proposed course metrics can help in the triage process. This was echoed by several participants in the enhanced metrics condition who wished they knew more about the class demographic breakdowns, especially when deciding between four or five courses that looked similar on the metrics, additional breakdown data such as achievement gaps or class composition might be needed to make a final decision. Though this is not the final step toward a solution, metrics like these play a key role in narrowing down the number of courses that require human evaluation. In line with previous research suggesting that using broader metrics can effectively narrow resource-constrained decision-making into smaller groups for human inspection [Vul et al., 2014], we found that educators trust the use of these metrics in this context. In higher education, department chairs or deans could use these metrics to inform them of course performance across their units. Principals could use these metrics to allocate professional development or assign mentorship teams. However, relying solely on scalable course metrics can lead to unintended consequences, and ultimately, a multi-faceted approach is needed [Thomas and Uminsky, 2022]. As stated in the

introduction, we see these two new metrics as part of a solution. They should not be relied upon solely but used in conjunction with other methods. The potential impact of these new metrics on EDI initiatives could be significant if used effectively by institutional stakeholders in secondary and higher education.

3.5.3 Future Research

Future research in this area is ongoing and needs to focus on more comprehensive solutions to support efforts that promote EDI. Our work proposes an initial triage system for EDI resource allocation, but we hope that future work will examine ways to provide participants with selected breakdown data and determine its usefulness in improving confidence and trust in decision making. Additional research should also highlight the attributes of accommodations and disabilities. While we mentioned the metric could consider IEP Plans, many educators in our sample asked for guarantees that this type of data was part of the metric. We also chose a convenient sample of Prolific educators, which is not representative of the broader educational community. Future research should explore additional avenues to survey a more representative sample. Another interesting research avenue is understanding how this metric could help in longitudinal accountability efforts. While our metrics can be tracked over time, there is no mechanism in the metric that tracks it over time. That may give more weight to more recent years or include the ability to reduce the weight on abnormal years, such as the pandemic years. This ongoing research and development process can be part of ongoing efforts to improve EDI in education.

CHAPTER 4

ALGORITHM APPRECIATION IN EDUCATION: EDUCATORS PREFER COMPLEX OVER SIMPLE ALGORITHMS

Although our research has made significant progress in developing scalable metrics to address equity and inclusion in education, it is imperative to examine the broader context of technological advancements and their implications for educational practices. As AI-powered tools become increasingly integrated into educational environments, comprehending the factors influencing their adoption and effective utilization is essential.

A principal challenge in this context is addressing algorithm aversion, which describes the tendency of educators to be reluctant to trust and adopt AI-based systems, even when these systems present potential advantages. In the subsequent section, we will investigate the factors contributing to algorithm aversion, including insufficient transparency and perceived complexity. We will analyze how providing detailed explanations for complex algorithms can alleviate these concerns and promote trust among educators. By gaining insight into the psychological underpinnings of algorithm aversion and formulating strategies to mitigate it, we can facilitate the successful integration of AI-powered tools within the educational sphere.

4.1 Introduction

Trust is an important factor influencing the pace of AI adoption in education. It is essential for accepting and effectively integrating AI-based educational tech-

nologies [Kizilcec, 2024, Viberg et al., 2023]. Educators are reluctant to accept AI recommendations that contradict their prior knowledge about students, and they expect AI to be infallible even in subjective situations [Nazaretsky et al., 2021]. Factors that influence trust generally include perceived benefits, ease of use, transparency, and anxiety [Ayanwale et al., 2024, Choi et al., 2023]. In higher education, Aladi [2024] highlights the lack of transparency, reliability issues, and ethical concerns as key factors slowing down the integration of AI technologies. These factors need to be addressed to understand how best to handle concerns of adoption and trust and successfully implement AI-powered tools in education.

Algorithm aversion has been proposed as an explanatory framework to understand how trust, transparency, and confidence issues might undermine the adoption of AI tools. Algorithm aversion posits that educators would be reluctant to use tools with algorithms, even though these tools' capabilities could benefit their instructional practices [Dietvorst et al., 2015, Kaufmann, 2021]. Algorithm aversion suggests that educators would prefer simple algorithms (or get human advice) over seemingly complex algorithms. The framework also suggests remedies to reduce algorithm aversion, such as increased transparency [Turel and Kalhan, 2023], giving users control over the algorithm's outcomes [Cheng and Chouldechova, 2023], and providing feedback mechanisms [Xu et al., 2023], to raise the likelihood of technological adoption. However, how to provide effective explanations for AI in education, explanations that raise algorithmic transparency in such a way that promotes technology adoption, remains an open question.

This research, through a series of experimental studies, investigates the ef-

fects of receiving detailed explanations for a complex algorithm to lower algorithm aversion and encourage adoption. Our overarching research question was: How does the presence of a detailed explanation for a complex algorithm affect educators' attitudes and intent to use an AI-powered tool? Study 1 employed a two (explanation) by two (visualization) between-subjects design with repeated measures to compare educators' attitudes and preferences for a complex algorithm versus a simple one when exposed to both algorithms. Study 2 compares three conditions (simple algorithm vs. complex algorithm with explanation vs. without explanation) in a between-subjects design to measure the effects on educators' attitudes and preferences after completing the task of explaining the algorithm to another person to simulate the need to explain and justify decisions informed by the algorithm. This work contributes to the literature on human-centered AI in education by demonstrating the extent of algorithm aversion among educators, which appears to have evolved over time, at least in what has come to be a common area of application for AI-based education technology, intelligent tutoring systems that promote mastery learning.

4.2 Background

4.2.1 Algorithm Aversion

Algorithm aversion refers to people's reluctance to use algorithmic decision aids despite their superior performance [Dietvorst et al., 2015]. This phenomenon has been observed in various domains, including autonomous vehicles [Kang, 2022] and forecasting [Dietvorst et al., 2015]. In education, it has been studied

with both in-service and pre-service teachers, who often prefer human advice over expert models [Kaufmann, 2021].

Several factors influence algorithm aversion, including perceived task subjectivity [Castelo et al., 2019], familiarity with algorithms [Mahmud and Islam, 2023], and trust in the system [Ireland, 2020]. However, Kaufmann [2021] found that commonly studied personality traits like openness and neuroticism were not related to teachers' perceptions or behaviors regarding expert models using algorithms. Introducing AI tools for instructors requires considering the effects of algorithm aversion to increase adoption and use and identifying instructor-related factors that might influence initial levels of algorithm aversion. Researchers have proposed using various levels of algorithm transparency to help reduce algorithm aversion [Cheng and Chouldechova, 2023, Turel and Kalhan, 2023]. Xu et al. [2023] tested algorithm aversion reduction strategies with higher education administrators using an AI tool to assist with making credit articulation decisions. They found mixed-results for reducing algorithm aversion, as allowing users to provide feedback about the algorithm increased aversion rather than reducing aversion. This research, like other educational research has emphasized the importance of human-centered design for crafting effective strategies in education [Buckingham Shum et al., 2019, Guo et al., 2024]. More research is needed to understand what approaches can influence algorithm aversion, such as providing higher levels of algorithm transparency.

4.2.2 Measuring Algorithm Aversion

Various measures have been used to assess algorithm aversion. One common approach is to compare human and algorithmic decisions [Dietvorst et al., 2015, Kaufmann, 2021, Castelo et al., 2019]. Another measure is to assess participants' perceived competence in the recommendations proposed by the algorithm [Castelo et al., 2019]. Finally, researchers have examined participants' intention to use tools with algorithms [Dietvorst et al., 2015, Kang, 2022, Cheng and Chouldechova, 2023]. The measurements' diversity provides flexibility for researchers to adapt algorithm aversion to their context.

Algorithm Comparisons

Many studies investigate the effects of algorithm aversion by providing two versions of decision-making agents, typically a human or an algorithm [Dietvorst et al., 2015, Kaufmann, 2021, Castelo et al., 2019]. Participants are typically exposed to decisions made by both agents and asked to answer which decision-making agent they would prefer to use. While effective, this approach may not always be realistic or efficient in educational contexts, such as teachers' use of intelligent tutoring systems (ITS). Instead of comparing a human to an agent, in this context, we need to compare educators' heuristics for learning and the ITS algorithm in the ITS environment [Holstein et al., 2017]. One way to simulate this comparison would be to study the difference between a simple heuristic algorithm and a more complex algorithm used for knowledge tracing.

Knowledge tracing (KT) algorithms have been widely used in education to model and track students' knowledge states during the learning process [Dai

et al., 2021, Zia et al., 2021]. These algorithms are employed in ITS to personalize instruction and improve student outcomes [Hicke, 2023, Zhang and Maclellan, 2021]. KT has been applied to sequence educational content, leading to improved student performance and engagement compared to expert-designed approaches [David et al., 2016]. Two methods of KT, Bayesian Knowledge Tracing (BKT) and N-Consecutive Correct Responses (N-CCR), have been widely used [Kelly et al., 2015]. N-CCR is based on a simple heuristic: once a student answers N questions correctly in a row, the student has demonstrated proficiency. For example, 3-CCR would determine proficiency after a student answered three questions right in a row. BKT has also been widely used in ITS for student modeling and performance prediction. BKT utilizes Bayesian statistics to maintain internal states of student proficiency. BKT is a two-state Hidden Markov Model where the unobserved hidden state being modeled is student learning, and for a given knowledge component, a student has a state of either learned or not learned [Kelly et al., 2015]. BKT can use individualized parameters and personal priors to update the hidden knowledge state constantly. As students answer questions, the internal state can be updated based on accuracy, time, hints, and other question-related factors. Comparatively, BKT is a more complex algorithm than N-CCR, and fewer people are expected to know the inner workings of BKT initially. This research informs our first research questions.

RQ1: What are the differences in educator perceptions of tools using a simple (N-CCR) versus a complex (BKT) algorithm?

Attitudes

Prior research indicates that attitudinal constructs such as perceived confidence, accuracy, and trust are strongly associated with lower algorithm aversion and increased adoption of new technologies [Dietvorst et al., 2015, Cheng and Chouldechova, 2023, Castelo et al., 2019, Nazaretsky et al., 2021, Ireland, 2020]. Likewise, educators' trust and confidence in education technology tools significantly influence their adoption and usage. Key factors affecting trust include perceived benefits, transparency, self-efficacy, and understanding of AI [Nazaretsky et al., 2021, Viberg et al., 2023]. Trust can be enhanced through professional development programs that explain AI decision-making processes [Nazaretsky et al., 2022] or other essential factors, including minimizing additional workload, increasing teacher ownership, and addressing ethical concerns [Cukurova et al., 2023]. Further research on educators' perspectives and needs can guide approaches to optimize AI use in education [Kizilcec, 2024, Qin et al., 2020]. This will require measuring educator attitudes related to trust and confidence, given their association with both algorithm aversion and technology adoption.

4.2.3 Explanations as an Intervention

A number of studies have demonstrated that providing explanations can reduce algorithm aversion and stimulate intentions to adopt technology [Turel and Kalhan, 2023, Wang et al., 2024]. Yet research on algorithmic transparency in educational settings reveals a complex pattern of relationships between explanations, trust, and perceived fairness. Providing explanations for algorithms like

BKT can increase trust, perceived accuracy, and user confidence [Williamson and Kizilcec, 2021]. Explainable AI (XAI) techniques have been increasingly applied in education to enhance trust and confidence in AI tools among educators. These techniques have the potential to significantly improve educators' trust and technology acceptance without increasing cognitive load too much [Wang et al., 2024]. However, challenges remain, such as tailoring explanations to different stakeholders and the relative nature of explicability across populations and domains (Farrow, 2023). In general, there is a need for more research to understand what types of explanations can reduce algorithm aversion towards education technology for different stakeholder groups. This research informs our second research question.

RQ2: How does explaining a complex algorithm (like BKT) influence educators' attitudes towards tools using complex or simple algorithms (like BKT and N-CCR)?

We conducted a series of studies to better understand how explanations can influence algorithm aversion and technology adoption. We explore the effects of explanations on attitudes related to algorithm aversion and technology adoption, using a comparison between a simple algorithm (N-CCR) and a more complex one (BKT) with varying levels of explanation.

4.3 Study 1

In preparation for Study 1, we conducted a pilot study with teachers who actively use the ASSISTments platform (recruited via an email newsletter). Our study design and stimuli are inspired by the Skill Builder feature in the AS-

SISTments platform, and we were eager to understand how teachers using ASSISTments would react to algorithms. We used feedback from the pilot data to improve study materials to make them more realistic. Notably, even in the small pilot study (n=39 educators), we saw evidence consistent with the results reported in Study 1 and 2. Grounded in our research questions and the review of the literature, our goal was to test the following hypotheses in Study 1:

- **H1.** Verbal and visual explanations of BKT lead participants to prefer it over N-CCR.
- **H2.** Verbal and visual explanations of BKT will positively increase participants attitudes about the BKT algorithm.

4.3.1 Methods

Participants

A total of 170 participants were recruited from Prolific for Study 1. Participants received $1.70 for completing the 10-minute survey, advertised as seeking input on an Adaptive Teaching App. A power analysis conducted with G*Power determined a target sample size of 170. The analysis aimed for 95% power to detect a medium effect size of 0.25 at a significance level (alpha) of 0.05. The analysis considered six repeated measures and four participant groups. Of the 170 participants, 2 were excluded due to lack of teaching experience (1) or prior experience with Bayesian Knowledge Tracing (BKT) algorithms (1). Analyses were conducted on the remaining 168 participants. Participants had been teaching for 11+ years (44%), followed by 6-11 years (31%), followed by 2-5 years

(23%), followed by 2% who taught for less than a year. The majority identified with feminine pronouns (59%), followed by masculine pronouns (36%), followed by non-binary pronouns(2%). 3% of participants preferred not to answer or selected "Not Listed.". The majority of participants identified as White (81%), followed by Asian (9%), followed by Black or African American (6%), followed by Native Hawaiian or Pacific Islander (1%), and American Indian or Alaska Native (1%). 2% of participants indicated "Not Listed."

Procedure

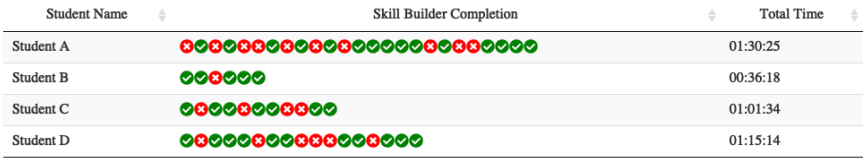


Figure 4.1: Sample report that participants were shown for Study 1. Participants in the Detailed Visualization condition saw percentages below every 4 questions indicating the student's proficiency at that point.

Participants were given a narrative introducing a learning tool called Skill Builder, that helps teachers determine when a student has learned a particular skill in any subject area. They were informed that Skill Builder uses an algorithm to determine if a student has learned a topic and adjusts the questions accordingly. Participants were then asked to review two algorithms and provide their opinions on both.

Next, they were presented with a description of the 3RR algorithm and a sample report showing multiple students and their progress to proficiency. The table looked similar to Figure 4.1. Participants answered questions about their attitudes towards the 3RR algorithm.

Participants were randomly assigned to one of four conditions based on a 2x2 factorial design. The next page introduced the BKT algorithm, providing a description and sample learning progress visualization (the content of the explanation and visualization dependent on their condition), followed by the same set of attitudinal questions. Finally, participants were asked to compare the two algorithms and provide a rationale for their preference.

Experimental Manipulation

In the no-explanation condition, participants received this one-sentence description of the BKT algorithm: "This algorithm determines that a student has mastered a skill once they reach a high probability of mastery based on their responses up to that point." In the BKT explanation condition, participants additionally received the following information about the BKT algorithm:

The algorithm uses all of the following information to estimate the probability that a student has mastered a skill. If the probability is above 95%, the algorithm determines that the student has mastered the skill.

- *an initial probability that the student has mastered the skill based on their first answer (this will be higher if they answer correctly)*
- *a guess probability for multiple-choice questions (e.g., the chance of guessing correctly is 50% for a True/False question)*
- *a slip probability for answering incorrectly even though the student already mastered the skill (i.e., they accidentally get it wrong)*
- *the question difficulty based on how many other students got it wrong before*

- *other process information, such as how many hints the student asked for or how much time it took them to answer the question*

In the BKT simple visualization condition, participants received a simple table depicting a sample student's learning progress mirroring the one shown in the 3RR algorithm. In the BKT detailed visualization condition, the same table was enhanced to show the estimated probability of proficiency on the report.

Measures

We measured participants' attitudes towards each algorithm using six questions with 5-point unipolar response scales ('Not at all," 'Somewhat," 'Moderately," 'Very," 'Extremely'). We adapted this set of measures from prior work that used them to examine people's attitudes towards algorithms in a similar educational context [Williamson and Kizilcec, 2021].

Confidence: "How confident are you that the Skill Builder feature with this algorithm will help you teach your students?"

Understanding: "How well do you understand how this algorithm determines if a student has mastered a skill?"

Sophistication: "How sophisticated do you think this algorithm is for determining if a student has mastered a skill?"

Accuracy: "How accurate do you think this algorithm is at determining if a student has mastered a skill?"

Trust: "How much do you trust this algorithm to determine if a student has

mastered a skill?"

Speed: "How much effort will it take you to gauge your students' mastery using Skill Builder problem sets with this algorithm?"

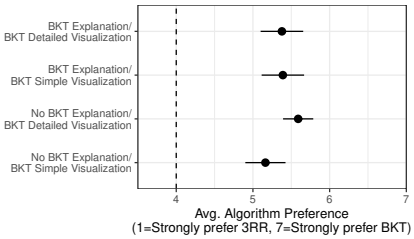
Use: "How likely are you to start using Skill Builder problems with this algorithm in your teaching practice?"

At the end of the survey, participants rated their general preference over the two algorithms in response to the following question (questions adaped from [Williamson and Kizilcec, 2021]): "You just learned about the 3 Right in a Row (3RR) algorithm and the Bayesian Knowledge Tracing (BKT) algorithm. Which algorithm would you prefer to use for Skill Builder problem sets in your teaching practice?" Response options were on a 7-point bipolar scale: 'Strongly prefer 3RR', 'Moderately prefer 3RR', 'Slightly prefer 3RR', 'Neither prefer 3RR nor BKT', 'Slightly prefer BKT,' 'Moderately prefer BKT,' 'Strongly prefer BKT.' Participants were invited to provide a rationale for their preference using an open-ended question: "Please tell us why you prefer the algorithm that you choose above."

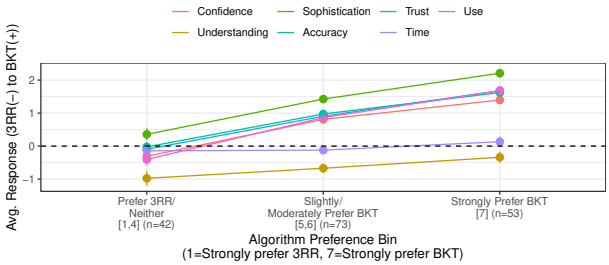
4.3.2 Results

Effects of Explanations on Preference and Attitudes

We examined how algorithm preference varies across experimental conditions by fitting a linear regression model. We find no evidence in support of H1: algorithm preference did not change as a result of adding explanations or visualizations ($F_{3,164} = 0.4454, p = 0.7209$). Figure 4.2a shows that educators,



(a) Average algorithm preference by condition for all 168 participants. Participants consistently preferred the BKT algorithm, regardless of the provided explanations or visualizations.



(b) The average responses on each attitudinal measure demonstrate a significant positive correlation with preference for the BKT algorithm. This trend is evident across three preference levels: Prefer 3RR to Neither, Slightly and Moderately Prefer BKT, and Strongly Prefer BKT.

independent of their experimental condition, tended to prefer BKT on average. While we did not find a significant effect on preference, we observed that the detailed visualization (but not textual explanation) raised confidence in BKT ($b = 0.466, t_{164} = 2.583, p = 0.011$), trust in BKT ($b = 0.4216, t_{164} = 2.124, p = 0.035$), and its perceived accuracy ($b = 0.410, t_{164} = 2.086, p = 0.039$). Thus, we find that detailed visualizations can improve some attitudes related to algorithms. Overall, the finding that educators prefer the complex over the simple algorithm is surprising, considering prior work on algorithm aversion.

Relationship Between Attitudes and Preferences

To analyze the relationship between participants' attitudes and algorithm preference, we calculated the difference between their responses to the 3RR and BKT tools for each attitudinal measure. A negative difference indicated a preference for 3RR, while a positive difference indicated a preference for BKT. For example, a participant who rated the 3RR tool 4 for confidence and the BKT tool 2 would have a difference score of -2, indicating a higher confidence in 3RR. Figure 4.2b shows the average response on each measure at three levels of preference. Re-

sults indicate that all seven measures are positively correlated with preference (all Pearson's r, between 0.17 and 0.71, with all $p < 0.001$). Accuracy, confidence, use, and trust had the strongest correlations with algorithm preference ($r > 0.63$). These four constructs are particularly influential in determining educator's preference for the BKT algorithms. Together, the seven measures explain 60.1% of the variance in preferences ($F_{7,160} = 34.43, p < 0.001$).

While participants in this study generally preferred BKT, our attempts to improve this preference through explanations and visualizations were limited. However, we did successfully influence Confidence, Accuracy, and Trust, key predictors of algorithm preference. Given the unexpected findings, we wondered if our manipulations could have been more effective if participants had engaged in an activity that would challenge their perceptions of BKT. In the scenario we set up in Study 1, participants might choose the more sophisticated algorithm without too much hesitation and not show a strong reaction to the explanations because they do not have to be accountable to the algorithmic recommendations or explain them to other stakeholders. To explore this, we changed the design in Study 2 to add an activity designed to reinforce the need to understand and trust the algorithm.

4.4 Study 2

In light of the surprising findings in Study 1, we updated the scenario-based design in Study 2 to add a realistic activity that would force educators to engage with the algorithm in ways that demand trust and explainability. Specifically, we created a situation in which the algorithm raises a concern about a student's

progress and we asked educators to explain the educational application with its algorithmic prediction to the student's parent/guardian. Our findings from the open-ended questions in Study 1 suggested that teachers might face trust and confidence issues when explaining algorithms to parents or students. Several participants in Study 1 expressed concerns about the complexity of explaining the BKT algorithm to parents and other teachers. Some participants preferred the 3RR algorithm, citing its perceived simplicity and ability to provide positive affirmations to students. These responses highlight the importance of understanding an algorithm when explaining it to others, as noted by Chaushi et al. [2023]. These concerns motivated us to design Study 2, which aimed to create a more realistic scenario where educators would need to explain the algorithm to a parent or guardian. This more realistic scenario allowed us to explore further the potential impact of explanations on attitudes and intentions.

In addition to the new study design, we also modified our dependent measures. Rather than using single-item measures, we adapted established scales to assess trust [Schoeffer et al., 2022] and competence [Ooge et al., 2022]. We also introduced a new measure, informational fairness [Schoeffer et al., 2022], to better understand users' comprehension of the algorithm. Still grounded in the same overarching research questions, we formulated the following two hypotheses for Study 2:

- **H1:** Educators rate the application with the simple algorithm (3RR) higher on (a) trust, (b) competence, (c) informational fairness, and (d) usage intention than the same application with the complex algorithm (BKT) if no further explanation is provided.
- **H2:** Adding an explanation for the complex algorithm increases educa-

tors' ratings of (a) trust, (b) competence, (c) informational fairness, and (d) usage intention for the application with the complex algorithm.

The study was preregistered on OSF at `https://osf.io/7c5zt/?view_only=a2a3c5629d3f4601aa9d30919e56aee9`.[1] All code, data, and supplementary materials for this study are available on OSF.

4.4.1 Methods

Participants

A total of 300 participants were recruited from Prolific for the study. Participants received $2.00 for completing the 10-minute survey, advertised as seeking input on an Adaptive Teaching App. Anyone who participated in Study 1 was excluded from participating in Study 2. All 300 participants were included in the analysis. Participants had been teaching for 11+ years (46%), followed by 2-5 years (26%), followed by 6-11 years (23%), followed by 5% who taught for less than a year. The majority identified with feminine pronouns (70%), followed by masculine pronouns (26%), followed by non-binary pronouns(1%). 3% of participants preferred not to answer or selected "Not Listed.". The majority of participants identified as White (86%), followed by Black or African American (7%), followed by Asian (6%), American Indian or Alaska Native (2%), followed by Native Hawaiian or Pacific Islander (1%). 3% of participants indicated "Not Listed."

[1] Our preregistered report included identifiable information, so we created a link with a copy of the non-identifiable information.

Experimental Manipulation

Unlike the previous study, participants in Study 2 were randomly assigned to one of three conditions that determined the algorithm and the level of detail about the algorithm they received. Participants in the 3RR condition were told that the Skill Builder application determines a high level of proficiency once the student answers three questions correctly in a row. Participants in the BKT condition were told that the Skill Builder application uses Bayesian Knowledge Tracing to determine a high level of proficiency. Participants in the BKT with Explanation condition were also told that the Skill Builder application uses Bayesian Knowledge Tracing. However, they were provided with a more detailed explanation of BKT (with similar text from Study 1) and a more detailed visualization indicating the percentage of proficiency in their report.

Procedure

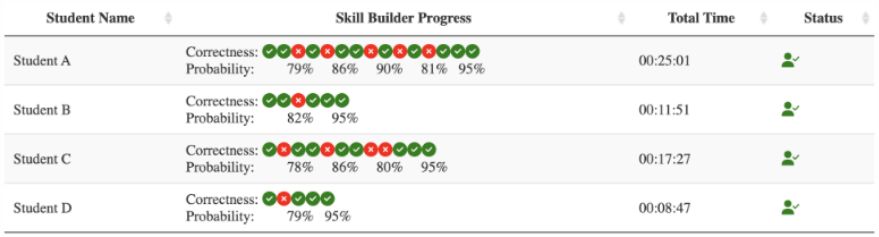


Figure 4.3: A detailed Skill Builder sample report was shown to participants in the BKT with Explanation condition. With the exception of the row indicating Probability, the same report was shown for the other two conditions.

As in Study 1, participants were provided with a contextual narrative that introduced a general learning tool called Skill Builder that helps teachers determine when a student has learned a particular skill in any subject area. Unlike the previous study, where participants were told they would see two versions

(algorithms) of the application, participants in Study 2 were only asked to give their thoughts on one version.

At this point, participants were randomly assigned to one of three conditions to determine the algorithm they were told that Skill Builder uses. There were 101 participants in the 3RR condition, 104 in the BKT condition, and 95 in the BKT with Explanation condition. Next, participants were given more information about Skill Builder and an introduction to the algorithm (dependent on randomization) that Skill Builder uses to determine proficiency. At the bottom of the page, they are shown a Skill Builder sample report that visualizes student performance. Figure 4.3 shows the report that participants were shown for the BKT with Explanation condition. The visualization shows that the student performance in all three conditions was the same. For example, Student A took 15 questions in all three conditions to reach proficiency. Since we prioritized consistency amongst the participants, the BKT proficiency percentages were not created by running the BKT algorithm as we had done in the previous studies.

After the detailed descriptions, participants were introduced to Chris, a student in their class who, according to Skill Builder, has yet to reach proficiency. They were also shown a sample report of Chris's performance, showing that Chris has not reached proficiency. While Chris's performance metrics (Question Answers, Total Time, and Status) were the same across all three conditions, participants in the BKT with Explanation condition did see the proficiency percentages in their report. After reviewing the report, we asked the participants to prepare a note for parent-teacher conferences that addressed Chris's performance on the Skill Builder application. The verbiage for this task was as follows:

For the parent-teacher conference, you decide to write a note explaining Chris's progress with the goal of developing a collaborative plan to help Chris with practicing the skill. Write a note explaining the situation to Chris's parent(s)/guardian(s). In your note, you should:

- *Briefly introduce the Skill Builder tool.*
- *Explain how the tool assessed that Chris requires more practice with the skill.*
  - If choosing a specific skill (i.e., in math, writing, history) helps you write the note, feel free to pick one.

After finishing their note, the participants answered several questions about their perceptions of the Skill Builder application (see Measures) and demographics.

Measures

We assessed participants' attitudes towards the Skill Builder application using the following pre-registered measures:

Intention. Intention to use the Skill Builder application was assessed using the question: How likely or unlikely would you be to use Skill Builder for your subject area in your classroom practice? The participants rated the question on a 5-point Likert scale (Very Unlikely, Somewhat Unlikely, Neither Likely or Unlikely, Somewhat Likely, Very Likely).

Competence. Competence in the Skill Builder application was measured with a four-item scale. The four questions asked were: It is effective at recommending decisions on student proficiency; It lacks the expertise to estimate my

student's proficiency; It fails to understand the student's level of proficiency; and It performs its role of recommending decisions on student proficiency well. Each question was rated on a 5-point Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree). The average score across the four items (Cronbach's alpha = .81) provided a single competence score for each participant.

Informational Fairness. Informational fairness in the Skill Builder application was measured with a four-item scale. The four questions asked were: It explains the decision procedures thoroughly; Its explanations regarding procedures are reasonable; I cannot understand the process by which the decision was made; and I don't have enough information to judge whether the decision procedures are fair or unfair. Each question was rated on a 5-point Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree). The average score across the four items (Cronbach's alpha = .82) provided a single informational fairness score for each participant.

Trust. Trust in the Skill Builder application was measured with a six-item scale. The six questions asked were: I trust that it makes high-quality decisions about student proficiency; I believe its decision-making procedure is unbiased; I know it is trustworthy based on my understanding of the decision-making procedure; I think I cannot trust it; It cannot be trusted to carry out student proficiency decisions faithfully; and In my opinion, it is not trustworthy. Each question was rated on a 5-point Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree). The average score across the four items (Cronbach's alpha = .89) provided a single trust score for each participant.

Note to Parents. A thematic analysis was conducted on the participants' responses. Responses were examined for recurring themes and tone, and similar

themes were grouped to develop our distinctive themes.

4.4.2 Results

Table 4.1: Results from the robust linear regressions to explain the difference in dependent measures for participants in the 3RR and BKT condition.

	Dependent variable:			
	Intention	Informational Fairness	Competence	Trust
	(1)	(2)	(3)	(4)
BKT	0.190	−0.412***	0.370***	0.236**
	(0.131)	(0.134)	(0.103)	(0.105)
Constant	3.743***	3.621***	3.327***	3.488***
	(0.098)	(0.093)	(0.080)	(0.084)
Observations	205	205	205	205
R^2	0.010	0.045	0.060	0.024
Residual Std. Error (df = 203)	0.936	0.957	0.737	0.752
F Statistic (df = 1; 203)	2.114	9.501***	12.936***	5.048**
Note:				*p<0.1; **p<0.05; ***p<0.01

Impact of Explanation of Intention and Attitudes

We first examine if the participants demonstrated initial algorithm aversion. We fitted a linear regression model to estimate the attitudinal and intention differences between participants in the 3RR and BKT conditions (Table 4.1). Contrary to our hypothesis (H1), participants in the 3RR condition rated the tool lower in Competence and Trust, and there were no differences between the conditions for Intention. Consistent with our hypothesis, Informational Fairness was higher for the 3RR condition. While participants indicated the BKT tool was

less understandable, the participants still had high ratings for Competence and Trust.

Table 4.2: Results from the robust linear regressions to explain the difference in dependent measures for participants in the BKT and BKT with Explanation condition.

	Dependent variable:			
	Intention	Informational Fairness	Competence	Trust
	(1)	(2)	(3)	(4)
BKT with Explanation	−0.112	0.191	−0.058	0.191
	(0.135)	(0.134)	(0.097)	(0.134)
Constant	3.933***	3.209***	3.697***	3.209***
	(0.087)	(0.096)	(0.066)	(0.096)
Observations	199	199	199	199
R^2	0.003	0.010	0.002	0.010
Residual Std. Error (df = 197)	0.947	0.945	0.684	0.945
F Statistic (df = 1; 197)	0.691	2.025	0.353	2.025
Note:				*p<0.1; **p<0.05; ***p<0.01

We then examine if providing explanations improved participants' perceptions of BKT. We fitted another linear regression model to estimate the attitudinal and intention differences between participants in the BKT with and without explanation conditions (Figure 4.4). Contrary to our hypothesis (H2), adding explanations to the presentation of the BKT algorithm did not improve any of the attitudinal or preference outcomes. Table 4.2 summarizes these results for H2. Overall, there were no differences between the two BKT conditions. While not significant, the Informational Fairness ratings indicate evidence that the BKT with Explanation condition may have helped with understanding the algorithm. The difference in Information Fairness and all other outcomes across conditions is shown in Figure 4.4.

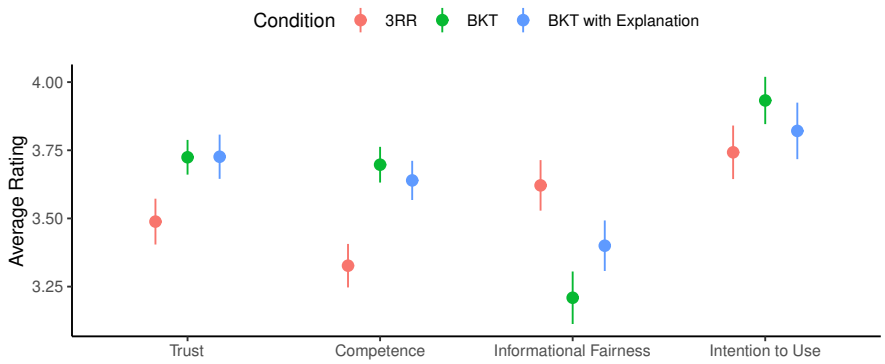


Figure 4.4: Attitudinal and Intention to Use measures by experimental condition. Trust and Competence are significantly lower for 3RR compared to both BKT conditions.

Exploratory Measures

Relationship Between Intention and Competence, Informational Fairness, and Confidence. We examined the relationship between intention and the other attitudinal measures. Regression analysis revealed that Competence, Trust, and Informational Fairness collectively explained 36% of the variance in Intention to use the tool. This indicates that these factors are significant predictors of educators' intentions to adopt and use the tool with Competence ($b = 0.581, p < 0.001$) and Trust ($b = 0.234, p = 0.026$) being the most influential in this analysis. This indicates that higher levels of competence and trust are associated with a higher likelihood of using the application. However, this relationship is only apparent for participants with higher levels of intention. Informational Fairness (understanding how the application works) showed no statistical relationship with intention ($b = 0.004, p = 0.943$).

Note. To better understand participants' attitudes and experiences, we analyzed the notes they wrote to parents. We extracted two types of themes from this analysis: deductive themes related to the level of detail about Skill Builder

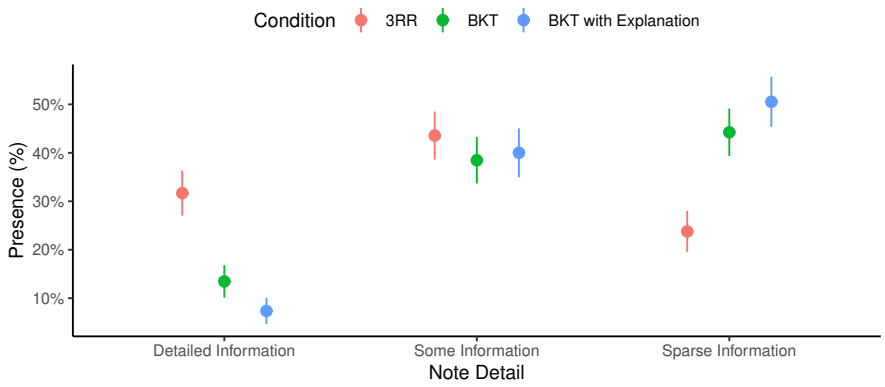


Figure 4.5: Note detail by experimental condition. Participants in the 3RR condition provided more detailed information about the algorithms in their notes to parents/guardians compared to those in the BKT conditions, regardless of whether an explanation was provided.

and inductive themes that emerged from the content of the notes.

We categorized the notes into three levels of detail: Sparse Information, Some Information, and Detailed Information (Table 4.3). 40% of responses were tagged as Sparse Information, 41% as Some Information, and 19% as Detailed Information. Figure 4.5 visualizes the distribution of note details by condition. As expected, participants in the 3RR condition provided the most detail about the Skill Builder application. This finding and the informational fairness ratings suggest that 3RR was the easiest to understand and explain.

While most analyses of the note details did not reveal new information, one notable finding emerged from the interaction between those who wrote little information and the dependent measures. Specifically, participants who wrote little information and were assigned to the BKT condition had a significantly higher intention to use the algorithm than those in other conditions.

Outside of the three themes about algorithm details in the note, we identified 12 other inductive themes from the notes. The complete list of themes

Theme	Definition	Example
Sparse Information	No or very little specific details about Skill Builder.	”Hello, we have a new tool to help Chris identify where he needs help. It uses AI to determine where his weakest points are. Please improve the indicated areas.”
Some Information	There is mention of a tool that helps identify low proficiency areas along with either information about how it determines proficiency or acknowledgement of what happens after proficiency is achieved.	”Skill Builder is an online tool that helps students review content and master skills by answering questions and trying to build streaks of correct answers.”
Detailed Information	Skill builder is explained with clear definitions of how it determines proficiency and what happens after proficiency is achieved.	”This year, I have began to use a tool call Skill Builder to help my students gain proficiency in their skills. So often in education, we tend to move on to new topics without ensuring that students have mastered the current ones. Skill Builder allows me to see that students are ready to move to new material after they can successfully answer three consecutive questions.”

Table 4.3: Descriptions of the themes about algorithm details in the note.

can be found on OSF. While helpful for providing a deeper interrogation into the participant's thoughts, most themes were irrelevant to our initial hypothesis, except for one theme, which we called ”Data from Reports.” The Data from the Reports theme was coded anytime a participant referenced the visualized report in their note. This code came in the form of mentioning how many answers Chris got correctly, adding in the proficiency percentages, or indicating the amount of time Chris took on the set of questions. It gives some evidence that the participants found the report understandable and clear. For example, one participant wrote, ”[Chris] needs to obtain a higher percentage score. Skill Builder recorded scores of 52%, 64% and 46%,” and another participant wrote,

"It looks as if it's getting harder for [Chris] since his score has dropped so he will need more practice." There were about 30% of the participants that included some aspect of the reports in their note. An analysis of this code by condition showed that it was even across the conditions, without around 30% of participants in each condition including information from the report. While there were no differences by condition, there was an interesting pattern with the Informational Fairness dependent measure, where those taking information from the note appeared to have higher Information Fairness ratings. Including data from the report might be an additional measure of understanding and clarity of the data.

4.5 Discussion

In this paper, we conducted two studies to understand the effects of providing algorithm transparency in additional text and detailed visualizations of educators' attitudes and intentions to use AI-powered tools. In Study 1, we provided educators with two algorithms, Simple (3RR) and Complex (BKT), and they overwhelmingly preferred and had generally positive attitudes toward the tool with the complex algorithm regardless of whether they were provided transparency into the algorithm. In Study 2, we only provided educators with one algorithm but added an explanation task where educators needed to explain the tool to a parent/guardian. Similar to Study 1, educators had similar positive attitudes toward the complex algorithm regardless of the level of transparency they were exposed to. In addition, Study 2 also indicated that understanding how the algorithm works (Informational Fairness) is not pertinent to predicting future intention to use a tool. Given these findings, we discuss whether algo-

rithm aversion has dissipated.

4.5.1 Algorithm Appreciation

Algorithm aversion was theorized when initial research regarding humans' perceptions of using algorithms to assist with decision-making showed humans being resistant to tools using algorithms [Dietvorst et al., 2015]. However, recent research shows that this trend has dissipated or reversed in what researchers call algorithm appreciation [Logg et al., 2019, You et al., 2022]. Algorithm appreciation is the tendency to rely on algorithmic advice over human advice, which might help explain the results of the studies in this paper. While we initially hypothesized that educators would demonstrate algorithm aversion and prefer simple heuristics to assist them in decision-making, educators instead responded that a more complex algorithm would be better to demonstrate learning. Turel and Kalhan [2023] framed algorithm aversion as an implicit bias or prejudice against AI and reasoned that, like other types of prejudices, exposure to AI would move people from aversion to appreciation. Exposure to AI is frequently coming from an increased usage of generative AI applications. Educators have employed GenAI tools like ChatGPT and Bing Image Creator for lesson planning, brainstorming, and professional development [Ruiz et al., 2024, Chen et al., 2023]. It may be that the recent explosion of generative AI tools geared toward educators has changed educators' perceptions of algorithms from aversion to appreciation. Even in these two studies, we saw a change in attitudes over time. We ran Study 1 in December 2022 and Study 2 in August 2024. In that time, we saw an increase in average differences in attitudes and intentions towards BKT. For example, in Study 1, the average intent to use

the tool with BKT was 2.899, but in Study 2, it was 3.875 for educators in one of the two BKT conditions. Trust displayed the same pattern, 3.130 for Study 1 and 3.725 for Study 2. It is hard not to think that the increased usage of generative AI has led to more favorable attitudes and usage of algorithms in general.

4.5.2 AI Literacy

The growing prevalence of AI literacy initiatives specifically designed for educators may have played a significant role in shifting their attitudes toward algorithms. Research suggests that increased AI literacy correlates with educators' willingness to learn and use AI-powered tools [Du et al., 2024]. A focus on integrating AI literacy into teacher education programs, including technical skills, ethical considerations, and practical applications, has been emphasized [Ayanwale et al., 2024]. This could have contributed to a positive evolution from algorithm aversion to appreciation. Educators with higher AI literacy are more likely to have positive attitudes and intentions regarding AI-based education technology. Our findings align with previous research highlighting the importance of psychological and social factors in technology adoption [Kizilcec, 2024]. While algorithm aversion might have had a limited impact, understanding the factors that facilitated this shift is crucial for designing and implementing AI-powered tools effectively in education. Tailoring these tools to individual needs and perceptions is essential for successful adoption.

4.5.3 Future Research

The results from this paper follow similar existing calls to design more studies focused on understanding educators' needs and perspectives of AI-powered tools. Our results indicate a few possible future research areas. We chose BKT as our complex algorithm, even though it is not a black-box algorithm. We favored the ability to explain the algorithm to novice users over a truly black-box algorithm. Our results show positive trends in understanding and informational fairness when additional explanations are provided, indicating that it was, on average, less understandable without any explanation. However, the gap is small, and running this type of study with a more complex algorithm may provide lower values for the complex algorithm without explanation, allowing for a greater space for the manipulation to work.

Another popular study design in this literature is about understanding AI attitudes and intentions where the AI gives wrong answers. We made our tool to provide realistic recommendations that educators would agree with and want to adopt. Another study could look at the effects of AI on student performance, namely when it gives a mix of right and obvious wrong recommendations.

The last area we encourage researchers to explore is experimenting with the context of how they need to explain the algorithm. Study 2 asks participants to explain the tool to a parent/guardian. The task of explaining the tool to a parent may have unintentionally influenced participants to report more positive views, as people are generally more likely to present a positive image of a classroom tool when communicating with parents. Future research could consider creating a narrative where the participant explains the algorithm and tool to a principal or school leader, along with their feedback about if and how to use the tool.

This design might give more space for participants who disagree with the tool to indicate more negative sentiment.

CHAPTER 5

USING INSTRUCTOR DASHBOARDS TO IMPROVE EQUITY AND INCLUSION IN COLLEGE COURSES

Continuing the prior discourse concerning the technical aspects of artificial intelligence tools and their impact on educators' perspectives, it is imperative to investigate the broader societal implications of technology within the realm of higher education. As artificial intelligence continues to advance, it is essential to consider how these tools may assist in addressing systemic inequalities and promoting social justice. In the subsequent section, we shall revisit the importance of data-driven strategies in enhancing equity and inclusion within higher education. We will examine instructors' perceptions of and engagement with sociodemographic data to inform decisions that can enhance student outcomes. By acknowledging the challenges and opportunities associated with data-driven methodologies, we can formulate strategies to effectively support instructors in cultivating more equitable and inclusive learning environments.

5.1 Introduction

Improving equity, diversity, and inclusion (EDI) in higher education institutions has been a daunting challenge due to the long-standing unaddressed obstacles that can exacerbate systemic barriers to success for minoritized students [Harper and Hurtado, 2007]. However, a critical area of promising research is the investigation of decisions that occur continuously about pedagogy, curriculum, admissions, and promotion and tenure [Fairweather, 2002]. Recently, these decisions have been guided by the increased use of administrative

data from university systems, climate and engagement data collected by surveys, and multimodal data collected by sensors in classrooms [Blikstein and Worsley, 2016, Hora et al., 2017]. The increased use of data by university staff offers researchers new opportunities to gain insights into decision-making processes in higher education and identify interventions to improve EDI. More specifically, the increased use of instruction-related data creates opportunities to provide instructors with valuable information regarding student behaviors in their courses and the accompanying outcomes. Considering the increased use of instructional data and the desire to create more equitable and inclusive environments [Harackiewicz and Priniski, 2018, Hora et al., 2017], what would happen if instructors had consistent and accessible equity and inclusion information?

Across many academic fields, researchers have been concerned with improving equity and inclusion in college courses. This uptake in interest has been evidenced by the numerous studies investigating whether various behavioral interventions have improved equity and inclusion measures [Harackiewicz and Priniski, 2018, Yeager and Walton, 2011]. Some studies have focused on active-learning interventions [Haak et al., 2011], while others explored affirmations and utility-value interventions to improve academic outcomes for minoritized students [Miyake et al., 2010]. The increase in investigations about improving equity and inclusion measures via behavioral interventions demonstrates that researchers across disciplines are concerned with issues of equity and inclusion in college courses. In particular, college classrooms are a fruitful context to deploy research interventions, given that these instructors have many touch points with students that play a critical role in facilitating a more inclusive classroom and a more inclusive college experience [Mayhew and Fernández,

2007]. Commonly, studies use various measures (achievement gap, belonging) to decide if an intervention was successful. However, very few studies have considered providing instructors with the data used to calculate these measures [Williamson and Kizilcec, 2022]. The lack of inquiry into this strategy informs the central theme of this research.

Some claim that providing data associated with equity and inclusion goals, typically socio-demographic data, is controversial. In the learning analytics community, this debate is focused on whether data should be color-blind (hiding racial data) or not (including racial data in analyses) [Shum, 2020, Li et al., 2023]. However, empirical studies exploring the usage of protected attributes have primarily explored predictive models [Yu et al., 2021, Li et al., 2023] rather than highlighting historical trends or systemic inequities for protected attributes. The limited research in this area, further underscores the importance of this study and its use of socio-demographics for auditing purposes to detect potential unfairness in a course [Baker et al., 2023].

Acknowledging the controversy of including socio-demographic data helps to justify the need for specific studies that look specifically at the mechanisms in play for instructor data-driven decision-making to improve equity and inclusion in their courses [Williamson and Kizilcec, 2022, Baker et al., 2023]. Along with investigating the types and presentation of data to instructors, additional queries are needed to understand how instructors' experiences created from their own identities and experiences might influence how they use racial and gender data [Lowery et al., 2007]. To that end, we designed a study to ask and observe how instructors interpret and use data containing socio-demographics. Our findings contribute to the literature by adapting technology acceptance

models to account for the discomfort and avoidance behaviors that can happen when instructors are presented with socio-demographic data. By accounting for these behaviors, institutions can better deploy data dashboards to help improve equity and inclusion in the university context.

5.2 Background

5.2.1 Instructor Data-Driven Decision-Making

While research on providing data to instructors to improve equity and inclusion in their classrooms is limited, many educational studies have investigated the process of providing learning analytics data to instructors [Brown, 2020, Hora et al., 2017, Herodotou et al., 2023]. Hora et al. [2017] examined cultural practices of data use and the impact of context and cognition upon decision-making by 59 faculty (36% response rate) across three universities to understand better how faculty use teaching-related data. Through a descriptive case study design, they explored how faculty used data and other types of information for planning their courses. These findings support other research [Echeverria et al., 2018b], indicating that faculty often rely on prior experience and knowledge about teaching rather than empirical evidence. While prior experience and knowledge can be informative for instructor decision-making, relying solely on intuitive expert decision-making and rapid decision-making can lead to incorrect decision-making due to biases or heuristics. For example, instructors may recall previous student behavior, such as frequent requests for extensions on a group assignment by students of color. However, additional empirical evi-

dence using course discussion data may suggest that students of color had difficulty finding groups; thus, changing the process for assigning groups rather than changing the due date might be a better solution. Findings from these studies support the argument that overcoming incorrect decision-making due to biases requires accessible and guided data to ensure the effective adoption of equity and inclusion objectives in teaching practices.

5.2.2 Threats to Equity and Inclusion-Related Data

In the U.S., social dynamics occur within a historical context that created a structured society with hierarchies assigned to race, gender, and other identities. Within these hierarchies, some identities are afforded more social power than others (referred to as privilege). For some, being confronted with one's privilege or reminded of one's place within a U.S. social hierarchy disrupts the comfort zone created by the hierarchy of racial power dynamics [Mills, 2019]. In turn, this internal discomfort can urge those with dominant identity positions to ameliorate the situation or avoid it to return to a familiar sense of comfort [Ahmed, 2017, Lowery et al., 2007].

Researchers have described patterns where encountering data that makes participants consider their own privileged positions causes feelings of discomfort, leading to avoidance behaviors and withdrawal [Chadwick, 2021]. Similar behavior has negatively impacted transgender and gender-expansive school children in the U.S. when encountering instructors who experience discomfort with the gender identity of a student [Luecke, 2018]. In Luecke's study, most instructors were accustomed to encountering children who used the binary gen-

der identities, boy and girl, and these aligned with the child's assigned sex at birth, male or female, respectively. Children whose gender identity did not align with these expectations, in this case transgender (an individual whose gender is different from their sex assigned at birth) and gender-expansive (individuals whose gender expands outside the binary of boy/girl) children, caused their instructor to feel uncomfortable. To mitigate the discomfort, instructors would avoid the student. The instructor's avoidance of the source of discomfort (the gender-expansive student) leads to adverse academic and social outcomes.

A growing body of research explores discomfort as a learning tool [Zembylas, 2020]. Using this approach, discomfort is a feeling of uneasiness linked to a disturbance in one's own 'comfort zones,' and cultivates an opportunity for deeper learning [Zembylas, 2015]. However, it is crucial to understand that an alternative to productively using discomfort as learning is instead to use privilege and power to avoid systemic issues [Ahmed, 2017]. In a study of school personnel navigating policy changes related to gender identity, participants who responded to or voiced discomfort about trans and gender-expansive identities then took action to restrict or "roll back" decisions supporting progress [Payne and Smith, 2022]. These actions allowed the personnel to relieve their discomfort by asserting a form of control. This body of research shows the importance of designing interventions that reduce discomfort through reflection rather than avoidance.

5.2.3 Theoretical Framework: Unified Theory of Acceptance and Use of Technology (UTAUT)

In the context of education, data is typically presented to instructors in the form of Learning Analytics Dashboards (LADs) [Wise and Jung, 2019, Ahn et al., 2019]. Given that LADs are a technology intervention to help provide feedback to instructors, previous researchers have used technology adoption and acceptance models to guide efforts to promote the usage of instructor-focused LADs [Herodotou et al., 2023, 2019]. To fully understand the role that discomfort and its accompanying avoidant behaviors might play in instructors' acceptance of a LAD, this study will heavily rely on a technology acceptance model.

Historically, the Technology Acceptance Model (TAM) has been used to explain the factors influencing a user's decision to adopt a technology [Davis, 1989]. The two significant variables affecting technology adoption are Perceived Ease of Use (PEOU) and Perceived Usefulness (PU). These two variables indirectly influence a user's behavior by directly influencing a user's attitude about the technology. Educational research has used TAM to explore how teachers/faculty/instructors adopt new technology. In their TAM meta-analysis, Scherer et al. [Scherer et al., 2019] found that including instructors in pre-implementation decisions and continued training from implementation through post-implementation led to greater adoption of new technology. Schoonenboom [Schoonenboom, 2014] took a more in-depth look at technology adoption of Learning Management Systems (LMS) and found that it is not just about the tool for instructors, but designers must also consider the instructor's task. These studies confirmed non-education TAM research, indicating that PU and PEOU may not be sufficient to model technology acceptance on their own.

The UTAUT was proposed to unify the many extensions of the TAM along with similar models for predicting technology adoption [Venkatesh et al., 2003]. Marangunić and Granić [2015] literature review identified 32 articles that either extended TAM or developed new models to describe technology adoption. Many of these extensions were created to address criticisms of the base TAM model, such as the need to account for more than just the attitudinal intention influenced by Perceived Ease of Use (PEOU) and Perceived Usefulness (PU) on technology use. For example, Mathieson et al. [2001] found that TAM models could be strengthened by accounting for a user's Perceived Behavioral Control (PBC). The inclusion of PBC brought together TAM attitudinal measures (PU and PEOU) [Davis, 1989] with constructs from the Theory of Planned Behavior (TPB) [Ajzen, 1991] to account for contexts where the user may not have the resources or knowledge to use a technological system appropriately. In education, Chen et al. [2013] adopted the TAM model with PBC to explore students' intentions to use a mobile Learning Management System and found that perceived resources had a significant effect on PEOU and Behavioral Intention (BI). The UTAUT incorporates PBC into the model through the construct of Facilitating Conditions (FC).

The UTAUT also expands TAM by including a construct for Social Influence (SI) and expanding the definition of TAM constructs, PU and PEOU, to Performance Expectancy (PE) and Effort Expectancy (EE), respectively. SI captures the degree to which users might be influenced by the attitudes towards the system from superiors or peers. Given that previous learning analytic dashboard research has found little support for SI in large-scale adoption [Herodotou et al., 2023], we did not expect to find evidence of SI influencing technology usage in our small and targeted sample. However, we do anticipate replicating

Herodotou's [Herodotou et al., 2023] qualitative finding that lack of resources or knowledge related to equity and inclusion (FC) will have a negative effect on an instructor's use of the dashboard (BI) and the degree to which they think the dashboard helps them improve their courses (PE).

5.3 Study 1

To begin to understand how the presence of socio-demographics might affect instructors' use of of dashboard, we first conducted a study with a set of qualitative interviews. This study was conducted using a two-wave research design. The first wave was a semi-structured interview with university instructors to understand their current data practices, and their perspective on incorporating equity and inclusion into their courses. We then designed and built a custom dashboard based on participants' responses about the data that would be helpful to them. The second wave was conducted approximately three months later and used a think-aloud approach to observe instructors' use of the custom dashboard displaying course data specific to the course the participant taught. This study was guided by the following research questions:

- **RQ1:** What conditions predict when an instructor may demonstrate discomfort (verbal or physical) when discussing socio-demographic data?
- **RQ2:** Do those with a higher presence of discomfort lead to engaging more in avoidant behaviors?
- **RQ3:** What additional influencers are important to understanding the successful adoption of an instructor dashboard displaying socio-demographic data?

5.3.1 Methods

Educational Context and Participants

Six instructors teaching large lecture courses were recruited from a selective research university in the United States. While all the participants resided in the same academic department, there was a variety of course content wherein some instructors taught programming-heavy data science courses, while others taught human-computer interaction design-based courses—half of the participants identified as men and the other half of the participants identified as women. The range of teaching experience ranged from one to twelve years, with an average of five years. While the instructors do teach many courses, we asked them to cater their responses to their large-lecture (¿100) undergraduate courses. Instructors have previously had access to course evaluation results (mid-semester and end of term) and enrolled students' major and academic level (e.g., sophomore, senior, graduate). All instructors indicated they had not been provided any socio-demographic information such as race or gender from the university. However, some instructors used custom course surveys at the beginning of their courses to gather some of this socio-demographic information. For the subset of instructors that surveyed students, they did not combine this data with performance data to see potential performance differences.

Wave 1: Semi-Structured Interview

Participants were interviewed to document their perspectives and responses to questions concerning general data use and, more specifically, using data to improve equity and inclusion in their courses. Each interview was recorded using

an online video conferencing platform (i.e., Zoom) and had an approximate duration of 45 – 60 minutes. While the interview questions covered many topics related to instructor use of course data, the following three questions were the most pertinent for this study (complete interview protocol available on OSF):

- **EDI Definition:** How would you define equity, diversity, and inclusion [in your course]?
- **Performance Differences:** Are there areas in your courses where you perceive performance and behavioral differences between different groups of students?
- **Aggregations:** Are there certain types of aggregations that should happen with non-sensitive [major, GPA] data that should not be allowed with sensitive data [race, gender]?

Creating the Dashboard

Following the results of wave 1, a dashboard was designed and built to support instructors in improving equity and inclusion in their courses. The dashboard was constructed using administrative data (i.e., Final Grades, Major, Co-Enrollments) and Learning Management System (LMS) data (i.e., Assignment Grades and Deadlines) for each instructor's large-lecture course. The dashboard was divided into three sections: Current Semester, LMS, and Historical. The Current Semester tab showed data about students enrolled for the next semester, such as majors, GPA, and demographic breakdowns using race/ethnicity [1] and gender. Figure 5.1 shows a couple of examples from

[1] While the researchers understand the difference between race and ethnicity, the university combines these two identities into a single field.

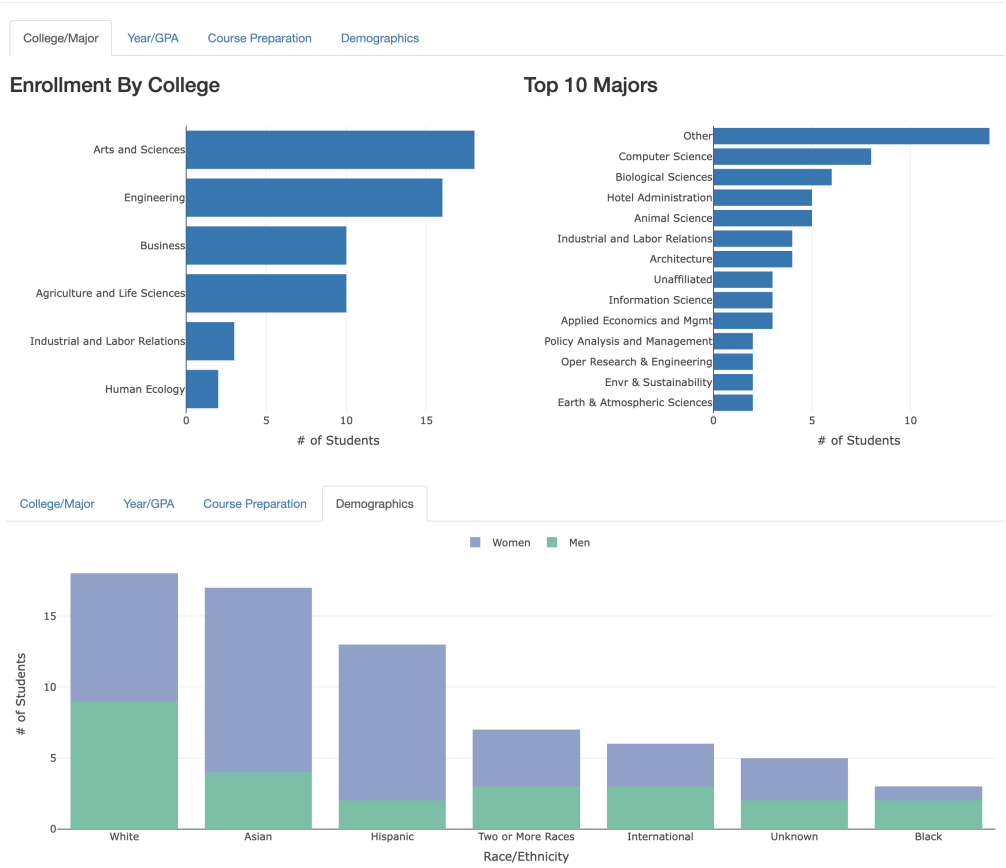


Figure 5.1: Screenshots from the *Current Semester* tab. The visualization on the top shows the current enrollment by college and the most frequent majors. The visualization on the bottom displays the gender and race/ethnicity breakdown. Data in this figure was simulated to not disclose actual student performance data.

this tab. The LMS tab showed average grades on assignments throughout the semester, and plots displaying the Average Grade Anomaly (AGA) and Performance Difference (PD) [Matz et al., 2017] by assignment for gender and race/ethnicity. The Historical tab displayed enrollment and final grade trends over the course offering time period. Visualizations displayed in the Historical tab included a filter to choose the semesters to display and a toggle to choose if the data should be aggregated by race/ethnicity, gender, or no aggregation. Figure 5.2 shows a couple of examples from this tab. Additional screenshots for the dashboard can be found on OSF. The dashboard was created using *RShiny*

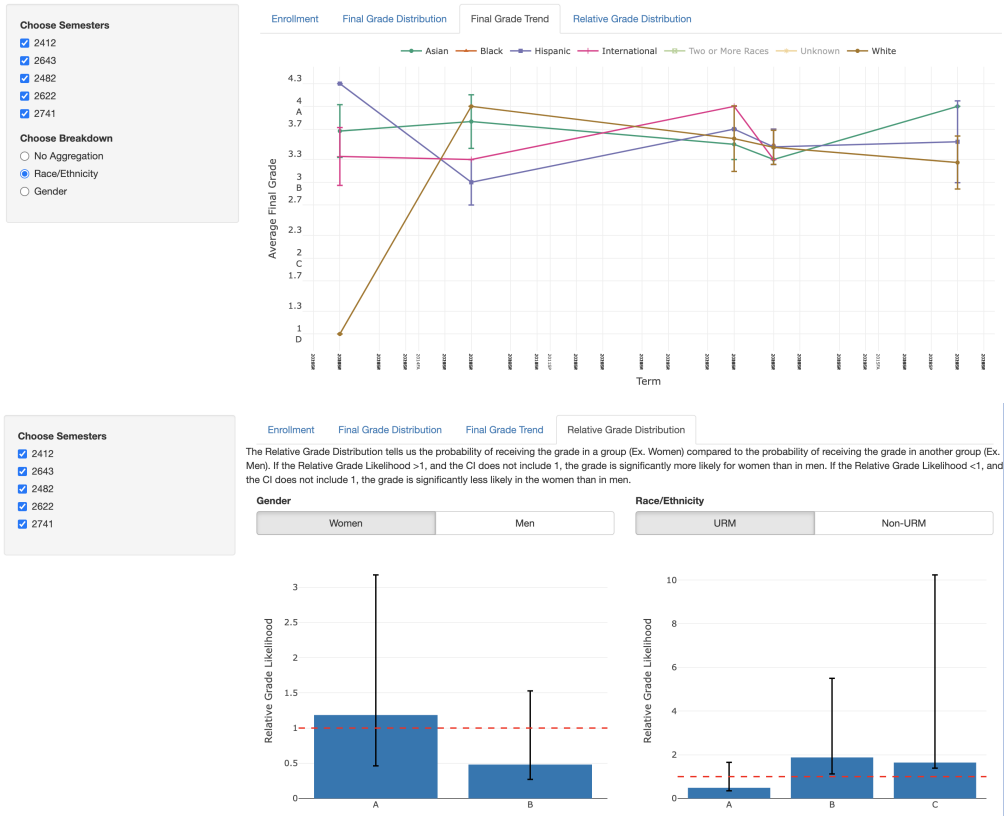


Figure 5.2: Screenshots from the *Historical* tab. The visualization on the top shows the final grade trend aggregated by race. The visualization on the bottom displays the relative risk for specific letter grades for gender and URM status. Data in this figure was simulated to not disclose actual student performance data.

and *Plotly* packages for the visualizations and *shinymanager* for authentication and authorization to ensure instructors could only view data for their courses.

Wave 2: Dashboard Think-Aloud

In the second wave, we aimed to understand which visualizations were helpful to instructors, how easily they could navigate the dashboard, and how they would act upon the data presented if the visualization were useful. Each interview was recorded using an online video conferencing platform (Zoom) and had an approximate duration of 45–75 minutes. The same participants who

participated in Wave 1 also participated in Wave 2. The procedure for the think-aloud portion began by sharing the dashboard URL and the login credentials with the participant. After the participants could log in to the dashboard, they were asked to share their screen with the researcher. As the participants encountered each visualization element, they were asked to answer the following questions: (1) What information is being conveyed by the visualization? (2) Is the visualization element useful? (3) If it was useful, how would you use the information? If it was not useful, why wasn't it useful? At the end of the session, the participants were asked if there was additional data not displayed that would help inform their course decision-making practices.

5.3.2 Analytic Approach

We conducted a hybrid thematic analysis that included a data-driven inductive approach and a deductive approach in which the UTAUT model served as the *a priori* template [Fereday and Muir-Cochrane, 2006]. The transcripts from both waves and memos created after each interview were initially reviewed and coded based on their relationship to key constructs in the UTAUT. The researchers then reviewed the transcripts to code phenomena unexplained by UTAUT. The whole process took several readings of the transcripts until the relationship between individual codes uncovered distinct themes. Our final list of themes was reduced based on the inclusion factor that multiple participants needed to have mentioned the theme. Once primary themes were identified, the transcripts were re-read to ensure that all instances had been coded consistently. In addition to the transcripts, memos were also written after each interview in both waves. Memos were used to record behavioral observations that were not

Table 5.1: Themes identified from coding the transcripts from Wave 1 and Wave 2.

Theme	Description
Performance Expectancy (PE)	Performance Expectancy - when participants indicated that the visualization element was useful.
Effort Expectancy (EE)	Effort Expectancy - when participants were able to make use of dashboard features to further drill down into the results.
Ease of Use - Issues	When participants had problems with the interface which kept them from making the data actionable.
University Data	When participants had problems with how the university data was collected therefore caused them to be uncertain on how to use the data.
Additional Uses	Participants suggested additional data or visualizations that were not related to the goal of the study.
Barriers	Participants indicated a resistance to using racial or gendered data to inform them about their course.
Facilitating Conditions (FC)	Facilitating Conditions - when participants indicated they had the control to make a course change based on the data.
Reflection/Help-Seeking	Participants indicated that they needed additional time/support to process the data in order to come up with actionable interventions. Some indicated a need for external consulting from an EDI expert, while others needed to informally reflect on their teaching practices or discuss with their teaching team.

caught in the audio transcripts. While we coded these behaviors in the data as descriptors, they were not thematic codes.

5.3.3 Results

We analyzed the data produced from both interview waves and identified key themes shown in Table 5.1. Grounded in the UTAUT, we coded themes related to PE, EE, and FC. Moreover, we found several additional themes that need to be considered to adapt UTAUT to explain technology adoption for improving equity and inclusion in college courses. We highlight several key themes related to our research objectives.

Dashboard Usefulness and Ease of Use

PE was the most prevalent theme in the interview transcripts, which indicates that participants found the dashboard generally useful. Some found it useful because it confirmed trends they thought were present in their courses. This sentiment is demonstrated by Participant 4 (P4), who said, "This is a trend I see in my class. This is like not surprising. But it's nice to see it. It's a good visual." Other participants found the dashboard useful because it helped them answer a question they had previously been considering: for instance:

> I think whenever we have the discussions about sort of controlling the growth of the major... which gets into like harsher enrollment caps, we really were worried that any kind of access restriction will translate into equity issues. We're not kind of aware of how that plays out. So it seems at least in enrollment that doesn't play out in this way. (P2)

EE was coded with two themes (Ease of Use Issues and EE). Ease of Use Issues were times when the interface design disrupted the participant's ability to interpret and use the data. Many of these coded instances can be attributed to a desire to see the course gender breakdown without race and the lack of readability of a stacked bar chart with many colored groups. On the other hand, EE indicated times when a participant used one of the interactive components to drill further down into the data; for example:

> What I'm gonna do is just start cutting away and then a strategy I often use for these sorts of visualizations is bringing in things. Rather

than cutting out things. Because it kind of helps me focus more on individual pieces of information. (P1)

All participants could navigate the dashboard pages and access all of the information, and all but one participant utilized all of the filters and aggregation options. A typical use of the filters in the *Historical* tab was to remove semesters in which a participant did not teach the course. These filters allowed them to narrow down the data to terms where they were in control of the curriculum.

Agreeableness to University Data

We found additional themes that influence the successful adoption of technology for improving equity and inclusion. While participants found most of the visualizations helpful, issues with labels caused all of the participants to question the data and reduce their ability to act upon the data. In particular, all participants noted that they found it challenging to view the data as valid because of how the university recorded data about students' socio-demographic characteristics (i.e., gender, race, and ethnicity labels). We coded this issue as University Data to reflect issues named by participants about how the university stores socio-demographic data. Many participants did not like that data was aggregated with only two genders and shared a sentiment like "Women and men or like that's 2. There's more than 2, right?" (P3). Gender was not the only issue; many also took offense to the labels for race/ethnicity. A few participants were critical of the language being used, along with the issue of only allowing students to carry one label. For example:

So international is just like the big bucket of any international like

> does international Asian get duplicated if you're international from Asia or. Like how does this work? (P5)

Other participants were frustrated by the vastness of the labeled groups, comprising many heterogeneous groups. For example:

> I mean, international would mix together everyone who's not a US citizen, I guess. Maybe that's the basis they did it on. So anyone who's here on a visa...but the way it's constructed right now doesn't help that much because it doesn't distinguish. (P3)

Data Discomfort

Some participants were hesitant when imagining themselves using the data from the dashboard to inform changes to their course, even though they found the dashboard useful and easy to use. We will first explain how the discomfort manifested before they saw the dashboard (i.e., in the Wave 1 interview) and then how it manifested when during dashboard exposure (i.e., in the Wave 2 interview). In the first wave, we asked participants three questions (EDI Definition, Performance Differences, and Aggregations; see details in Methods) that we expected to produce some discomfort if instructors were uncomfortable talking about socio-demographic data. We did not find many signs of discomfort when the participants talked about their EDI Definition. Most participants did not include identity-specific information in their definitions and spoke broadly about creating welcoming environments that ensure all students have a high chance of success. Similar to the responses to the EDI Definition, when speaking about perceived performance differences, participants avoided racial and

ethnic identities. Instead, many referred to students' academic major or college as potential sources of performance gaps. Three of the participants also indicated that there might be gender differences in their course, with women being more active participants than men. We did notice signs of discomfort in most of the participants who acknowledged that they have no idea if there are performance differences in their courses. For example:

> So I have an intuitional sense of like, yes, this is happening, and coming from my own background, I probably am going to be ignorant of it. But beyond that, I don't have a great sense of it, and that's the thing that scares me is that there are problems that I feel are likely to be happening. But I don't necessarily know what the problems are, or how bad they are... I'm pretty scattered on this, too, because I haven't thought about it enough. (P1)

This discomfort in the interview became even more prevalent when participants responded to the question about aggregating student data along socio-demographic lines. All participants displayed some degree of visual and/or verbal discomfort when discussing seeing data dis-aggregated by socio-demographic data such as race/ethnicity or gender. In particular, one participant demonstrated multiple signs of discomfort, such as touching their head, changes in breathing pattern and voice, and changing their body orientation. For example, in the following statement we included all utterances and repetitions to help illustrate the discomfort.

> I feel really conflicted about the yeah, about this prospect. Say, of getting, you know the same data by self-reported race ethnicity, or

> by self-reported gender like that. The plus side of it is, I feel it would help me know if I am systematically misserving a group of students. I I I imagine everyone you talk to every instructor would say, I work hard not to do that. Surely every self-reflective instructor would also say, it's possible that I do do it in some way that's invisible to me, and while I can always continue trying to work harder, it sure would be an effective prompt to me if I saw data revealing that like, hey, look your your systematically failing the women in your class right? That you better do something about that fast, and if you don't have the data, you don't have the same prompt. I know though two that like I don't know how to say it, except that it feels like squeaky. It shouldn't have to be the case that you see, this data revealed disparity to know that you should be maximizing the inclusivity of of the assignments of the structure of the class of the examples that you give you whatever it is that you could do to make people welcome and supported and enable their success in the class you should be doing that, anyway. Whether the data tells you that you're failing in some way or not. (P6)

In the second wave, we noticed that the participants who demonstrated significant discomfort before being exposed to sociodemographic data also demonstrated discomfort after exposure. However, instead of displaying physical reactions, the discomfort manifested in the amount of control they perceived themselves to have for making course changes based on that data. The themes of *Additional Uses* and *Barriers* code instances when participants started to defer control. The first theme, Additional Uses, highlights when a participant would indicate that additional data was needed, but that data would not help them

accomplish broader EDI goals. For example, one of the most common data requests was related to building an early alert system to identify individual students instead of systemic issues. This theme was commonly expressed by saying, ”Yeah, so I but I want to find this student early so they can drop the class or do something else to help them stay engaged” (P4). The other theme, Barriers, is a collection of reasons that participants raised that prevented them from being able to make any changes. Sometimes, this was directly stated by participants, such as ”There are specific things that I can change about my class like prereqs [prerequisite courses]. But there is not that much else that I can do with grades, although this is nice to see.” (P5). Sometimes instructors justified the concern by referencing larger societal concerns about ”create[ing] stereotypes or strengthen[ing] stereotypes.” (P4). At other points, instructors, even when error bars were present, questioned the significance of a trend. For example:

> Is it possible to have counts next to each of these, like I recognize that there are counts...but it could be useful to see like oh wow like The 2 or more [races] students are super overperforming, but actually if there’s just like 2 students like maybe it’s not as huge of an issue. I mean, obviously an issue, so useful to see counts. (P5)

We found that the avoidant behavior of focusing on external factors to explain differences in the performances of socio-demographic groups was directly related to FC in the UTAUT model. The discomfort observed was actually reflecting their lack of ability or resources to act upon the data. This finding was highlighted by a pattern wherein some participants moved from initial discomfort and avoidant behaviors to brainstorming potential actions. A key turning point in this process for two of the participant came when they acknowl-

edged that some of their course practices might have created these differences, and they could either reinforce things that worked or rework assignments that caused the greatest differences in performances.

Reflection/Help-Seeking

A common practice we observed for participants who worked through their discomfort with socio-demographic data was to employ internal reflection or begin thinking about whom they may ask for additional guidance. We interpreted reflection and help-seeking responses as tools that participants use to better understand the resources and abilities available to address issues they might have encountered in the data. We identified one theme, *Reflection/Help-Seeking*, that could thwart some of the distractions that can affect perceived control. Additional Reflection was identified when we observed multiple participants going through internal reflection processes during the study that caused them to change their thoughts on what they could control. For example:

> And it's something I really need to think about because I like I reflected before. These [Hispanic] students are not as salient to me. In my thinking about URMs in the class. Like to be blunt, that's the reality. I don't think I have this salient in my head. And that's a sign that maybe I'm neglecting other aspects of their course. So that's something to think about this fall? This is a good concrete outcome for this visualization for me. (P1)

We also found that participants who displayed the least amount of discomfort engaged in either internal reflection or indicated the need to consult others for

help. For example:

> It would generate an idea for a proposal or a course of action and then I would treat that as a starting point for a conversation. I would look at it and say like, oh, interesting. Let me think about that...I think I better go talk to somebody, you know, other people who I know are good teachers that might have interesting perspectives on these things. (P3)

Still, some participants wanted an expert opinion or guidance on interpreting or using this data to improve equity and inclusion. In particular, P4 had identified a problem but needed help correcting the issue and tried to ask the interviewer for guidance.

In addition to discussing ideas with other instructors, most participants wanted a comparison function in the dashboard. The participants highlighted two helpful reasons for comparison: Context and Help Identification. In terms of context, the participants wanted to know if they were seeing trends specific to their course or if they were departmental/university trends. For example:

> Like you do see that like the drop in the men students who enroll is kind of sharper than women students...It might be interesting to know how much that is replicated across the major... I'd wonder if this means that in [Department] we are attracting more women into the major than men versus something in this course. (P2)

Other participants were looking for comparison information because "I would like to know who's doing better in my department, so I can go talk to them."

(P1)

In the context of UTAUT, we noticed that FC was increasing by the reflection and guidance that participants mentioned. We noticed more comments about changing behaviors to improve their courses when FC increased. For example, P4 initially believed they could not implement changes based on performance data aggregated by socio-demographics. However, after reflection, they believed that "Highlighting stereotypes might help us think about what we're doing wrong, and we need to improve, for certain groups, especially say, first-generation students." More often, this theme was characterized by participants taking on the responsibility when an undesirable outcome has been identified, for example:

> Maybe there is something I can create at the level of my class to create structures that support students from certain groups. I mean, if I support students, it's like it's like designing for accessibility. If you design something that helps, say people with a certain disability, it's going to help everyone right? (P4)

Factors Moderating Discomfort

We have focused our attention on findings pertaining to participants who displayed significant amounts of discomfort. While everyone demonstrated some discomfort, there were a few participants who did not display large amounts of discomfort and who also did not engage in avoidant behavior. From the beginning of the study, these participants always accepted that any data showing differences is "only indicative of how I'm doing rather than how that group of stu-

dents is doing" (P2). The participants who exhibited this attitude also shared an experience of either not being born in the U.S. or spending significant amounts of time outside of the U.S. as compared to participants who were more likely to display discomfort. Additionally, we noticed a pattern with participants who were more likely to engage in reflection activities after they were exposed to the dashboard. These participants taught courses that are more closely related to human factors than to computer or data science. This is evident in P4's statement: "I try to tell students that [this topic] is for everyone. Everyone's invited. It's a place where regardless of your gender and race, you can do really well. You have to have good people skills. And that doesn't come with a specific race or gender."

5.3.4 Discussion

In this study, we investigated the reactions of university instructors when discussing socio-demographic data about students. We first examined which instructors demonstrated the most discomfort when asked to share philosophies regarding equity and inclusion and the aggregation of course data along socio-demographic lines. Since we could not directly tie discomfort to intended behaviors or actions to improve course climate in the first wave, we followed that interview with a think-aloud protocol, which allowed instructors to explore a dashboard of their data with options to aggregate by race or gender. Except for comments about gender categories, the instructors in the study did not appear to be uncomfortable when talking about data concerning gender. Even before they were prompted to think about socio-demographic data, many of the instructors mentioned gender when we asked them initially about perceived

performance differences. This pattern was a stark contrast to most not mentioning race/ethnicity until we brought it up in the interviews. Therefore, our discussion will focus on discomfort due to race/ethnicity.

Discomfort Discussing Student Race/Ethnicity Data

This study's findings offer crucial implications about how instructors understand and act upon socio-demographic data. As exemplified in the results section, several instructors displayed discomfort when discussing potential performance differences and the potential of socio-demographic aggregation to highlight potential differences. This discomfort was replicated when we presented course data to the instructors. A primary display of discomfort was to engage in increased distractions, either through offering up different uses for a dashboard or mentioning barriers, such as statistical significance or external factors out of their control. We found that these distractions diverted attention from using the data as an opportunity to change the courses to improve the climate. This use of distraction can be interpreted as strategic colorblindness, which the psychology literature defines as an individual decision to avoid talking about race in an effort to be perceived as less racist [Apfelbaum et al., 2008]. While this behavior can have relieve the immediate discomfort, it is counterproductive in working to address systemic issues. Importantly, we found that distraction was more a temporal state than a personal trait as some instructors worked through this discomfort and started thinking about actions to improve their courses. This finding suggests that deploying a dashboard with some form of guidance may help instructors avoid counterproductive distractions.

Since we found varying degrees of discomfort across the instructors in our

sample, we further investigated what factors might affect discomfort. In the first wave, we found an association between discomfort and spending time outside the United States. Instructors who did not grow up in the U.S. displayed fewer signs of discomfort. This finding is confirmed by existing research indicating that international students do not have strong concepts of the U.S.'s social construction of race and do not see race as an issue to be made uncomfortable [Mitchell et al., 2017]. This pattern can allow these instructors to reduce discomfort by distancing themselves from the historical nuances or perceived blame from systemic issues. This finding is further supported by Lowery et al. [Lowery et al., 2007], who found an association between perceived racial threat and the degree to which individuals identify with racial categories imposed on them. Instructors not raised with identifying into U.S. racial categories will be less inclined to feel threatened or discomfort when systemic issues are highlighted. We also noticed this distancing in stereotype usage as a barrier, with instructors who spent a significant amount of time outside of the U.S. being less likely to worry about the potential of reinforcing stereotypes if they were shown racial data.

In the second wave of interviews, we identified similarities between instructors who could work through their discomfort. We found that the course subject matter may be associated with being able to talk through issues and think about potential actions. Instructors who taught more policy or HCI-based courses showed more signs of reflection and were able to process areas in their course design that might explain significant performance differences. This result was reinforced by statements made in the first wave of interviews, where policy and HCI instructors indicated that inclusion was a foundational tenet of their curriculum and that they already had experience discussing equity and inclusion

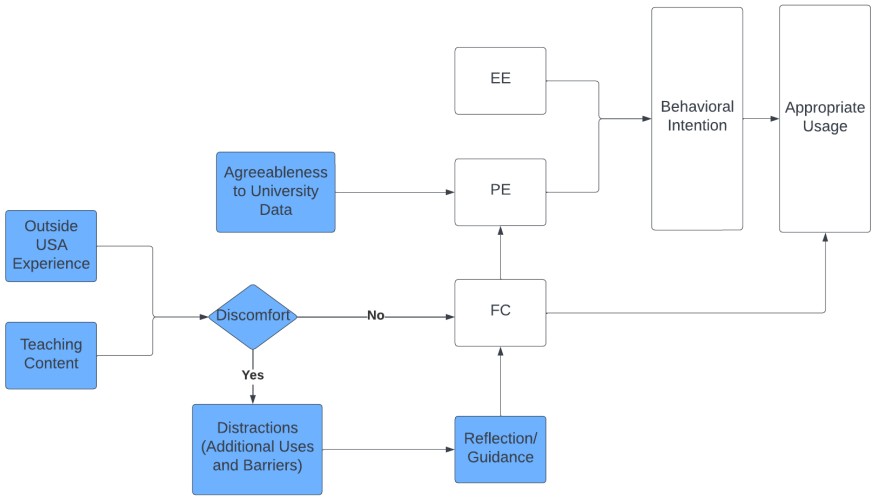


Figure 5.3: UTAUT model applied to EDI dashboards with new components (colored background) based on this study's findings.

issues in their courses. This finding indicates that more support may need to be provided to instructors who teach courses where topics around equity and inclusion are not regularly taught in the course.

Accounting for Data Discomfort in UTAUT

Building on UTAUT, we propose a mechanism for the acceptance and adoption of dashboards for improving equity and inclusion in a context where data discomfort plays an important role. Given our findings that discomfort can distract instructors from using data to improve their courses and that discomfort can be mitigated, we propose a set of additions to the UTAUT (see Figure 5.3). We account for potential issues of displaying socio-demographic data to instructors by adding several factors. First, we indicate the importance of agreeableness towards the data source (here university data) for PE. This new factor accounts for the possibility loss in perceptions of usefulness if instructors do not agree with the socio-demographic categories that are used, or they find them too expansive

to be actionable.

Next, we posit that data discomfort does not affect PE or EE but FC. Similar to Herodotou et al. [Herodotou et al., 2023], we found that qualitatively, FC has a strong influence on instructors successfully acting upon the dashboard data, so we included two pathways for instructors to feel they have developed the resources and abilities (FC) to act upon the data. One pathway is for instructors who initially approach the dashboard data with little discomfort. Due to significant experiences outside the U.S. and/or the content of their course, these instructors approached the dashboard with more resources (FC) to interpret and use the data to improve their courses. In cases when instructors show high discomfort, additional interventions to provide expert guidance or help with reflection are needed to help increase FC to the point where they have the abilities and resources to make course changes that help to improve their course.

The extended model of UTAUT can inform university efforts to deploy data presentations with socio-demographic data. In particular, it suggests the need for targeted interventions to accompany these dashboards. The results from this study indicate that a successful intervention would need to involve tools for reflection and personalized advice for their teaching contexts. Universities could also target instructional areas, such as STEM departments, that could benefit the most from an intervention program. Alternatively, universities looking for partners to promote a new dashboard may prefer to initially reach out to human-centered subjects where the level of initial discomfort could be lower than average. While the findings of this study focused on socio-demographics that are uncomfortable in the context of the U.S., stakeholders in other geographic areas can interpret our findings as relevant for displaying potentially

uncomfortable data in their context (e.g., data related to immigration status in Europe, native populations in Australia, or the cast system in India).

This study finds varying degrees of discomfort when instructors are asked to discuss or review socio-demographic student data. This discomfort can lead to avoidant behaviors, which can thwart widely held goals of improving equity and inclusion. Some instructors can move from a state of discomfort to actionability through the process of reflection. This leads us to suggest potential policies for universities to help the adoption and effective use of dashboards that include socio-demographic data. Our results show ways to help institutions reduce discomfort and move the onus of responsibility away from external factors (students or department/university trends) to course design that instructors can control.

5.4 Study 2

Building upon the findings of Study 1, we designed a larger, controlled experiment to replicate the observed discomfort effects and address certain limitations. In Study 1, participants were solely asked to reflect on their own course data. To gain a more comprehensive understanding of the factors influencing discomfort, Study 2 introduced a perspective manipulation and a new gap manipulation. Previous research on perspective-taking highlights the distinct effects of self-perspective and other-perspective on emotions and prejudice reduction [Vorauer and Sasaki, 2014, Jackson et al., 2006]. By incorporating a perspective manipulation, Study 2 aimed to explore whether the discomfort effects observed in Study 1 were influenced by participants' self-perspective or other-

perspective.

Furthermore, Study 1's reliance on real-course data limited our ability to control the size and type of performance gaps presented to instructors. As Quinn [2020] demonstrated in their study on the effects of news reporting, the presence of achievement gaps in news reporting can influence viewers' perceptions. To address this, Study 2 introduced a new gap manipulation, carefully designing visualizations to represent different scenarios:

- Racial Gap: A clear performance gap based on race (similar to the "achievement gap" framing used by Quinn).
- No Gap: A balanced performance distribution (similar to the "counter-stereotypical gap" used by Quinn).
- Non-Racial Gap: A performance gap based on a non-racial attribute (similar to the control group in Quinn's study).

By controlling the gap type and magnitude, Study 2 allowed for a more systematic examination of how these factors influence instructor discomfort. This approach provides a more nuanced understanding of the phenomenon, as it isolates the impact of the gap itself from other contextual factors.

To operationalize discomfort, Study 2 adopted an avoidance-based approach, aligning with existing research on social situations and information avoidance. Previous studies have demonstrated that discomfort can lead individuals to avoid outperformed peers [Exline et al., 2013] and potentially unwanted information [Sweeny et al., 2010, Golman et al., 2015]. In racial contexts, Apfelbaum et al. [2012] found that discomfort with race-related topics can

hinder information gathering and workplace diversity efforts. Study 2's behavioral measurement of avoidance allows participants to choose whether to continue viewing aggregated data or avoid such representations. This approach provides a more concrete and measurable assessment of discomfort, enabling us to directly examine its relationship with the experimental manipulations.

To guide the study, we formulated two hypotheses and one research question:

- **H1:** Educators who see data with a racial performance gap for their course are subsequently more likely to avoid additional data that is disaggregated by race relative to educators assigned to any of the other conditions (i) no racial gap in their course, (ii) racial gap in a peer's course, (iii) gap in a non-racial attribute in their course, or (iv) gap in a non-racial attribute in peer's course.
- **H2:** Educators who see data with a racial performance gap for their course rate the visualizations as (a) less usefulness, (b) lower ease of use, and (c) lower trust relative to educators assigned to any of the other conditions (i) no racial gap in their course, (ii) racial gap in a peer's course, (iii) gap in a non-racial attribute in their course, or (iv) gap in a non-racial attribute in peer's course.
- **RQ:** How do avoidance behavior and perceptions (usefulness, ease of use, trust) vary across the conditions where educators see (i) no racial gap in their course, (ii) a racial gap in a peer's course, (iii) a gap in a non-racial attribute in their course, and (iv) a gap in a non-racial attribute in peer's course.

5.4.1 Methods

Participants

A total of 550 white U.S. educators were recruited from Cloud Connect for the study. Participants were specifically targeted based on their teaching experience (K-12 or higher education) and demographic characteristics (white, U.S.-born). This population was selected based on previous research suggesting they might be more likely to experience discomfort when confronted with racial performance gaps [DiAngelo, 2018]. Due to technical issues with the survey, 73 participants were removed due to incomplete responses. 469 participants were included in the analysis. Participants received $2.00 for completing the 10-minute survey, advertised as seeking input on designing data-driven insights for improving teaching. The majority identified as women (59%), followed by men (39%), followed by non-binary/agender (1%). 1% of participants preferred not to answer. 46% of participants primarily teach in K12, followed by 38% teaching in Higher Education settings, and participants who selected Other (8%) and those who teach Vocational/Technical Education (8%).

Experimental Manipulation

The study employed a between-subjects design with five experimental conditions. Participants were randomly assigned to one of these conditions, which manipulated their perspective (self or other), the type of student attribute (race or non-race), and the presence or absence of a performance gap. The five experimental conditions are detailed below.

- Self/Race Gap: Participants are instructed to think about a course they taught and are provided with race-by-group data that shows a performance gap.
- Self/No Race Gap: Participants are instructed to think about a course they taught and are provided with race-by-group data that shows no performance gap.
- Self/Non-Race Gap: Participants are instructed to think about a course they taught and are provided with by-group data for another attribute (teacher, year in college) that shows a performance gap.
- Other/Race Gap: Participants are instructed to think about a course that another instructor taught and are provided with race-by-group data that shows a performance gap.
- Other/Non-Race Gap: Participants are instructed to think about a course that another instructor taught and are provided with by-group data for another attribute (teacher, year in college) that shows a performance gap.

Procedure

After participants consent to participate in the study, they are asked about their primary level of teaching. This information is used later in the survey to make the visualization more relatable if the participant has been assigned to a condition that displays a Non-Race gap. In order to strengthen the perspective manipulation, we asked participants to recall and write about a previous course they taught (participants in Self conditions) or a previous course that another instructor taught (participants in Other conditions). After describing a course,

they are told that they will be viewing course data to help inform course improvements and that they are instructed to consider the course they previously wrote about while viewing the data.

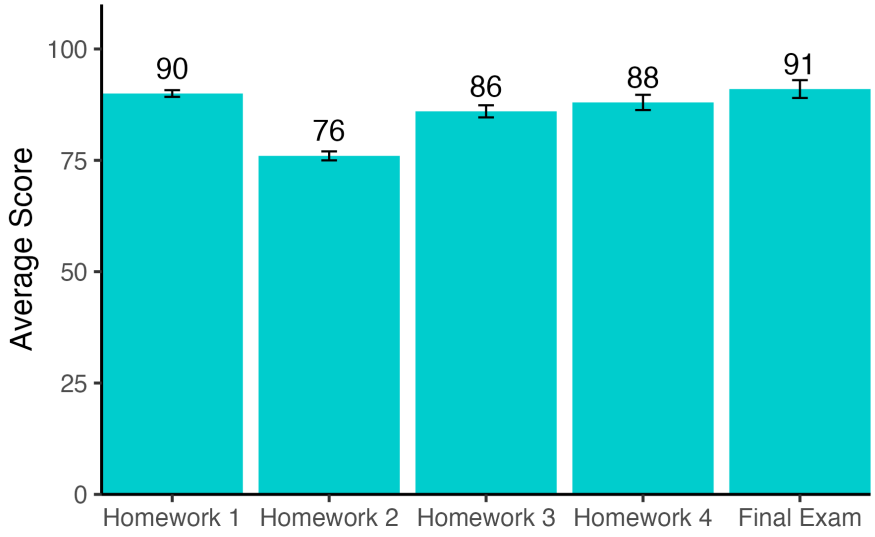


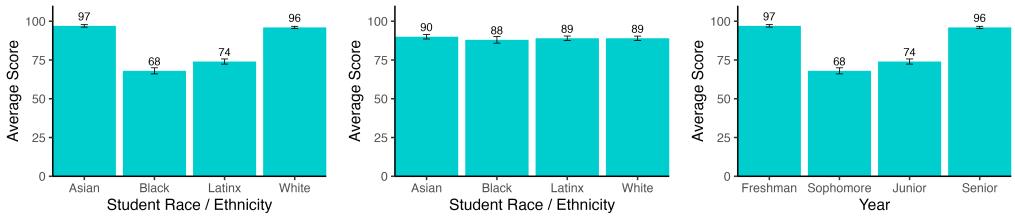
Figure 5.4: The overall visualization shown to all participants that shows student averages for Homework and Final Exam.

They are then presented with the first visualization, which shows the overall student performance on homework and a final exam. Regardless of the condition, all participants saw the same overall student performance visualization. Figure 5.4 is the first visualization presented to all participants. After they view the visualization, they are asked to choose whether they want more details about homework or the final exam scores. Depending on their choice, participants are then presented with a visualization showing homework scores or final exam scores grouped by a categorical variable that is randomly assigned. Participants in the No Race Gap conditions are shown average scores (homework or final exam), where each bar represents a different racial group, and there are no significant differences between the bars. Participants in the Race Gap conditions see a similar plot but the bars show White and Asian students signif-

icantly outperforming Black and Latinx students. Lastly, those in the Non-Race Gap condition see a visualization similar to the Race Gap condition, with two groups outperforming the other two groups; however, instead of the bars representing different racial groups, the bars correspond to a non-racial grouping relatable to the participant's teaching level. For example, higher education instructors in those conditions would be shown a visualization grouped by year in school (i.e., Freshman or Senior). High school and middle instructors would be shoen data grouped by the students' homeroom. An elementary school instructor would be shown data grouped by the previous year's teacher. Figure 5.5 shows the Race Gap condition. Participants' choice of homework or final score only affects the heading of the visualization; all values presented in the visualization are the same regardless of that choice. In addition, the gap size is the same for both Race Gap and Non-Race Gap conditions. After participants viewed their by-group visualizations, they were then asked to make another choice regarding which type of visualization they would like to see next about participation grades for this course. They are given the choice to see overall participation grades or participation grades broken down by their randomly assigned group (Race or Non-Race Group). After they make their selection, they are asked several questions about their perceptions of the visualizations (See Measures). Lastly, participants are shown the participation visualization they previously selected (Overall or By-Group).

Measures

We assessed participants' attitudes towards the data visualizations using the following pre-registered measures:



(a) Race Gap condition (b) No Race Gap (c) Non-Race Gap

Figure 5.5: By-group performance data provided to participants based on their randomly assigned condition

Self-Reported Measures

Performance Expectancy (PE). The visualizations effectively communicated the key insights about course performance. (Strongly agree to Strongly disagree)

Effort Expectancy (EE). The visualizations were clear and easy to interpret. (Strongly agree to Strongly disagree)

Trust. How much do you trust the data you viewed? (A great deal, A lot, A moderate amount, A little, Not at all)

Exploratory Question. Please describe the key insights you gained from the visualizations. (i.e., performance of different groups, the range of scores, scores of particular assignments)

Behavioral Measure

Avoidance behavior. Choice of visualization that is overall instead of disaggregated by group. Choosing the overall visualization indicates avoidance.

5.4.2 Analytic Approach

This study delves into the impact of data presentation on educators' avoidance behavior and their perceptions of data visualizations. The data was collected through a survey administered to educators participating in a research project, and various statistical techniques were employed to test the proposed hypotheses and research question. Hypothesis 1 suggests that educators who encounter data showing a racial gap with a self-perspective are more likely to avoid further disaggregated data by race than other conditions. A generalized linear model was used to analyze this hypothesis and examine the relationship between participants' choice (avoidance) and their randomly assigned condition. Hypothesis 2 posits that educators who see data with a racial gap with a self-perspective rate the data visualizations lower in terms of usefulness, ease of use, and Trust than other conditions. We utilized a linear regression model with robust standard errors to analyze the relationships between Performance Expectancy (PE), Effort Expectancy (EE), and Trust with the experimental conditions. Additionally, time analysis was conducted using a linear regression model with robust standard errors to explore the relationship between the amount of time participants spent on the visualization pages and the experimental conditions.

5.4.3 Results

Avoidance Analysis

We first investigate whether participants exhibited any avoidance patterns by analyzing their choices regarding participation visualizations (Overall vs. by

	Dependent variable:			
	Choice - Overall	PE	EE	Trust
	(1)	(2)	(3)	(4)
Self/No Race Gap	0.236***	0.181	0.064	0.189
	(0.064)	(0.122)	(0.104)	(0.125)
Other/Non-Race Gap	−0.068	−0.101	−0.227*	0.090
	(0.077)	(0.148)	(0.126)	(0.151)
Self/Non-Race Gap	−0.004	−0.025	−0.139	0.380***
	(0.074)	(0.141)	(0.120)	(0.144)
Other/Race Gap	−0.096	−0.034	0.039	−0.121
	(0.065)	(0.125)	(0.106)	(0.127)
Constant	0.518***	3.768***	4.411***	3.277***
	(0.046)	(0.088)	(0.074)	(0.089)
Observations	469	469	469	469
R^2		0.011	0.017	0.030
Adjusted R^2		0.002	0.008	0.022
Log Likelihood	−323.770			
Akaike Inf. Crit.	657.540			
Residual Std. Error (df = 464)		0.928	0.786	0.944
F Statistic (df = 4; 464)		1.293	1.952	3.592***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5.2: Regression results for the behavioral (Choice) and attitudinal (PE, EE, Trust) measures. The reference group condition is Self/Race Gap.

group) across different conditions. We fitted a logistic regression model to compare the differences between participants in the Self/Race Gap condition and the other four conditions (see Table 5.2). Contrary to our hypothesis (H1), participants in the Self/Race Gap condition did not significantly prefer the overall visualization compared to any other condition. In contrast, participants in the Self/No Race Gap condition were significantly more likely to view the Overall Participation visualization, indicating a tendency to ”avoid” additional data

aggregated by race. This finding is further illustrated in Figure 5.6.

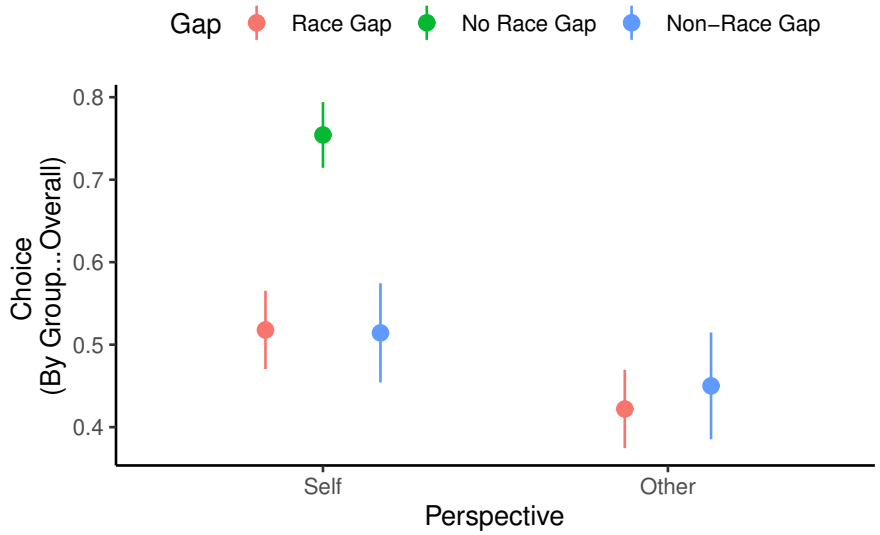


Figure 5.6: Participants selection of Overall vs By-Group visualization by condition. Choice values closer to 1 indicate avoidance by participants being more likely to choose to view Overall visualization.

Attitudinal Analysis

We then examined the differences in attitudinal measures between participants in the Self/Race Gap condition and the other four conditions, as summarized in Table 5.2. Contrary to our second hypothesis, we found no significant differences between the conditions in terms of PE. However, participants in the Self/Race Gap reported lower trust ratings compared to those in the Self/Non-Race Gap condition, and they had higher EE ratings when compared to participants in the Other/Non-Race Gap condition. Figure 5.7 illustrates these findings.

Regarding this study's research question (RQ), we identified only one other significant difference between the conditions. Similar to our second hypothe-

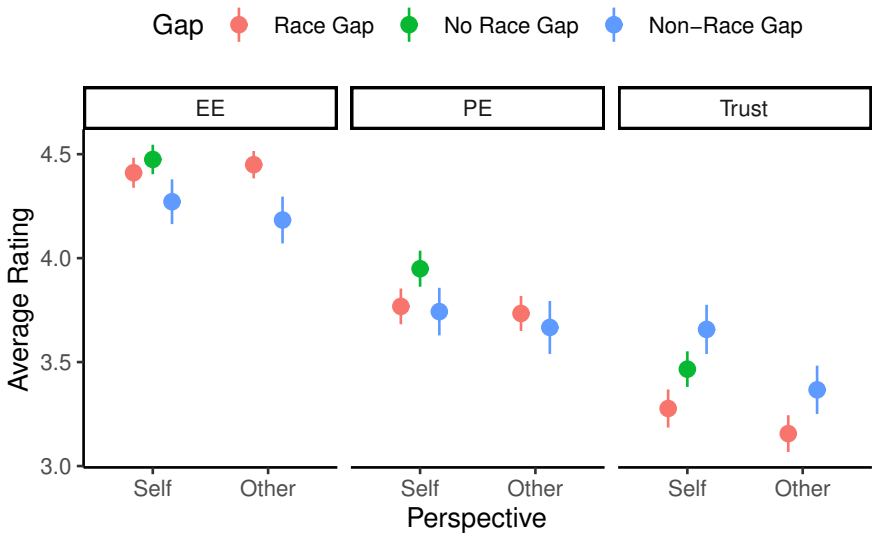


Figure 5.7: Attitudinal measures by condition. There are few significant differences between the conditions.

sis, participants in the Self/Non-Race Gap indicated significantly higher trust values compared to those in the Other/Race Gap condition ($b = 0.501, t_{464} = 3.389, p < 0.001$). Additionally, participants in the Self/No Race Gap also reported significantly higher trust values compared to those in the Other/Race Gap condition ($b = 0.310, t_{464} = 2.517, p = 0.012$). Overall, the combined results for both H2 and the RQ suggest that participants shown racial gap data—regardless of perspective—trusted the visualizations significantly less than those in the non-race gap or race data without a gap conditions ($b = -0.263, t_{464} = -3.024, p = 0.003$). In terms of EE, participants in the Other/Non-Race Gap condition reported significantly lower ratings compared to those in the Self/No Race Gap ($b = -0.291, t_{464} = -2.187, p = 0.029$) and Other/Race Gap ($b = -0.266, t_{464} = -2.039, p = 0.042$) conditions.

Given the variety of trust ratings, we conducted an exploratory analysis of the relationship between trust and participants' visualization choices across the five conditions. Table 5.3 presents the results of this regression analysis. Partici-

	Dependent variable:				
	choice-Overall				
	Self Race Gap	Self No Race Gap	Self Non-Race Gap	Other Race Gap	Other Non-Race Gap
Trust	−0.135***	−0.064	0.108*	−0.024	0.023
	(0.048)	(0.043)	(0.060)	(0.052)	(0.073)
Constant	0.959***	0.976***	0.118	0.496***	0.373
	(0.163)	(0.153)	(0.228)	(0.170)	(0.254)
Observations	112	118	70	109	60
Log Likelihood	−78.295	−67.865	−50.147	−78.664	−44.195
Akaike Inf. Crit.	160.591	139.730	104.294	161.328	92.390
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 5.3: Regression results by condition to examine the relationship between trust and the participants visualization choice.

pants in the Self/Race Gap condition exhibited the most significant relationship with trust, indicating that lower trust ratings were associated with a greater tendency to avoid by group visualizations (favoring Overall over By Group). Figure 5.8 further illustrates the trust relationship within the Self/Race Gap condition.

5.4.4 Discussion

In this study, we explored how educators perspective and gap presence affected their decisions on future visualizations and their perceptions of the visualizations. While there were not many significant differences between the conditions, there were a few surprising patterns that warranted more investigation.

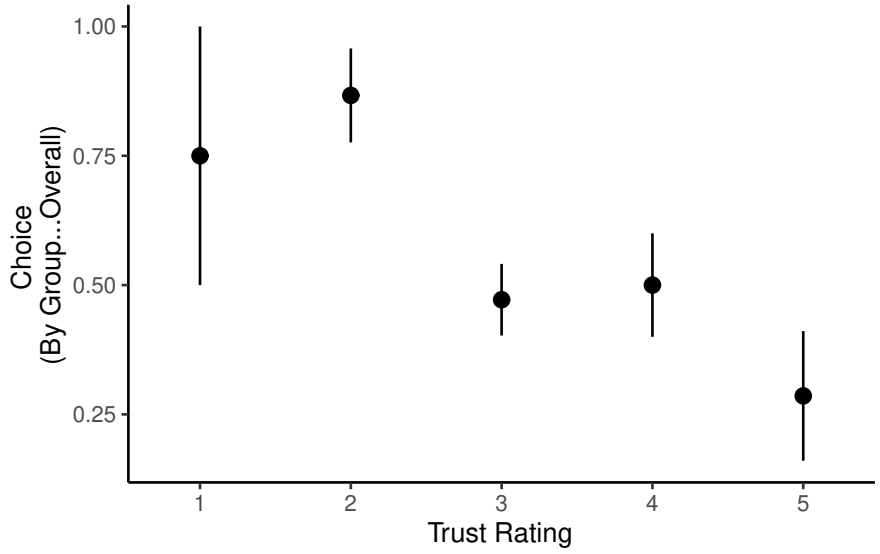


Figure 5.8: Visualization choice by trust rating for the Self/Race Gap Condition. Lower trust ratings were associated with a higher probability of avoiding the by-group visualization.

The Impact of Racial Gaps on Avoidance Behavior

Despite our initial hypothesis, we did not find clear evidence of discomfort among educators who were assigned a self-perspective and presented with racial gap data. Surprisingly, these educators were almost equally likely to choose to view additional data by group or overall. Interestingly, while not significant, a trend emerged where participants in the other-perspective condition, regardless of the gap condition, tended to prefer by-group visualizations. This finding aligns with previous research suggesting that individuals may experience less discomfort when considering someone else's performance rather than their own [Vorauer and Sasaki, 2014, Jackson et al., 2006].

However, a counterintuitive finding emerged: the group that, on average, chose to view overall visualizations was composed of educators who had a self-perspective and were presented with racial data without a gap. This finding

suggests that even when educators do not initially perceive a racial gap, they may still exhibit "avoidance" behaviors, potentially due to (1) a fear of discovering new negative information or (2) a belief that additional data would not provide meaningful insights. Individuals may subconsciously avoid information that could confirm their negative beliefs or expectations [Golman et al., 2015]. In this context, educators who were initially presented with a positive outcome may have feared that additional data could reveal underlying inequities or challenges. By avoiding further information, they may have sought to maintain a positive self-perception and avoid the discomfort associated with confronting negative realities.

Another potential explanation for the observed behavior is the concept of satisficing [Zach, 2005]. Instead of maximizing utility by exploring all available data, individuals may opt for satisfactory information [Roberts et al., 2019]. Future studies could incorporate a 'sufficient information' option to distinguish between those who genuinely lack interest in additional data and those who avoid potentially negative information. By considering these factors, we can gain a deeper understanding of the complex interplay between identity, perception, and decision-making in educational contexts. This knowledge can pave the way for the development of more effective data visualization tools and strategies, offering hope for a more equitable and inclusive educational landscape.

The Effect of Racial Gaps on Attitudinal Perceptions

Contrary to our hypothesis, there was not much separation between the experimental conditions on the key attitudinal measures (PE, EE, and Trust). Similar to Study 1, participants had high levels for the two primary constructs (PE and

EE) from the UTAUT framework. There were a few findings that merit additional discussion. First, participants who showed non-race gaps had the lowest ratings for ease of use (EE). This could be attributed to a more cognitive load when trying to understand the non-race attribute they are presented with. Usefulness was also primarily similar across the five conditions.

Given that participants in the Self/No Race Gap condition strongly preferred overall visualizations, it is intriguing that their attitudinal measures (PE, EE, and trust) were not significantly different from those in other conditions. This finding suggests that factors beyond these attitudinal measures may have influenced their visualization choices. One potential explanation is that participants in the Self/No Race Gap condition may have focused primarily on the overall visualization when providing their ratings, potentially neglecting the impact of the by-group visualization. Future research could employ a design to address this limitation, asking participants to provide ratings for each visualization individually. This design would allow for a more nuanced understanding of how different visualization types influence attitudes and decision-making.

One of the most significant findings of our study was the impact of racial gap data on trust. Educators who saw a race gap condition reported significantly lower trust ratings compared to the other conditions. This suggests that presenting individuals with data that highlights potential inequities can have a profound negative impact on their trust in the visualization. This finding underscores the importance of considering the potential implications of data presentation in our work.

To mitigate the negative impact of racial gap data on trust, future research should explore specific design strategies that can enhance trust and facilitate

understanding. Building on the work of Boyd Davis et al. [2021], researchers could investigate the use of clear and concise labeling, consistent color schemes, and contextual explanations to improve the overall user experience. By carefully considering these factors, we can create data visualizations that are not only informative but also emotionally resonant and trustworthy.

5.5 Conclusion

Our findings from Study 1 suggest that individuals may exhibit avoidance behaviors when confronted with data that could confirm negative beliefs or expectations. This finding is particularly evident when the data highlights potential inequities, such as racial disparities. In Study 2, we observed that participants presented with racial gap data reported lower levels of trust in the visualization. This finding further emphasizes the importance of addressing discomfort and promoting trust in data-driven decision-making.

To mitigate the negative impact of racial gap data and promote the effective use of data visualizations, we propose a modified UTAUT model that incorporates factors such as data source agreeableness and discomfort. By providing clear explanations, addressing concerns about data accuracy, and offering support for data interpretation, we can enhance trust and facilitate the adoption of data-driven practices. Additionally, the design of data visualizations should consider the emotional impact of the data and aim to minimize discomfort while maximizing understanding.

By understanding the psychological and social factors that influence individuals' responses to data, we can develop more effective strategies for using

data to drive positive change in education.

CHAPTER 6

THE SOCIOTECHNICAL PRACTICES NEEDED TO LEVERAGE DATA TO IMPROVE EDI IN EDUCATION

In the last four chapters, I have covered a broad overview of educational data-use for improving EDI. I have uncovered the limited research into setting up and evaluating Learning Analytic Dashboards for EDI efforts, despite the spike in creating these types of dashboards following the incidents of 2020. I have proposed metrics to better capture issues of course equality and inclusiveness. I have shown that against initial research findings, educators do not show algorithm aversion when presented with data or technology using complex algorithms. Lastly, I have identified the issue of race disaggregation as a bigger problem for data use than issues related to algorithm aversion. While the results from most of the studies show promise for future research and practice into leveraging data for inclusivity and equitable education, and indicate that data/AI literacy may not be the roadblock to data-driven decision making, another issue arose surrounding the need to address EDI literacy and readiness to fully use data for improving EDI. In the following section, I will discuss key takeaways for researchers and practitioners interested in embarking on data-driven decision-making for EDI.

6.1 EDI Data-driven Decision-Making Roadblocks

Using Ziskin and Young [2023] five barriers to as an outline, this section unpacks the research in the prior chapters to help explain current roadblocks to using EDI for data.

6.1.1 Addressing Performativity Misalignment

The concept of performativity misalignment, as discussed in previous chapters, is a significant barrier to educators' effective use of data for equity and inclusion. This phenomenon occurs when educators prioritize other performance metrics over equity metrics, leading to a disengagement with disaggregated data [Ziskin and Young, 2023]. As demonstrated in Chapter 3, educators without specific EDI metrics often struggle to make informed decisions related to equity and inclusion. Even with a primary objective to identify courses needing improvement, those with a misalignment of priorities, such as a focus on traditional metrics, may exhibit lower confidence and trust in their decisions. Similar to the findings in Chapter 5, educators may fear that data will be used punitively, hindering their willingness to engage with it. This misalignment can arise when the organization prioritizes traditional evaluation metrics over EDI initiatives, creating a culture of disdain for equity-related data.

Addressing this barrier is crucial for educational organizations to support educators in improving EDI. Performativity misalignment can lead to cynicism among educators who see no value in equity-focused initiatives if they are not prioritized and rewarded. It can also create a culture of resistance to data use if measures are perceived as solely for reporting and accountability. Furthermore, educators may feel ill-equipped to tackle EDI problems without proper resources and support. To overcome performativity misalignment, organizations must prioritize equity metrics, provide adequate resources, and foster a culture that values data-driven decision-making for EDI. By aligning incentives and expectations, educational institutions can empower educators to effectively use data to create more equitable and inclusive learning environments.

6.1.2 Addressing Collaborators' Resistance and Data Legitimacy

A significant barrier to using data for equity and inclusion (EDI) is resistance from colleagues who question the focus on sociodemographics or the relevance of the data [Ziskin and Young, 2023]. While previous research suggested algorithm aversion as a primary obstacle, our findings in Chapter 4 indicate that this may not be the primary concern. Instead, the resistance often centers on the legitimacy and relevance of the data itself. When confronted with data that reveals racial or ethnic disparities, educators may question the significance of the gaps, raise concerns about data completeness, or attribute performance differences to individual factors such as preparedness or motivation. These justifications, while potentially valid, are often selectively invoked to avoid confronting uncomfortable realities (see Chapter 5).

As demonstrated in Chapters 5, educators are more likely to question the significance of racial gaps and raise concerns about data completeness when confronted with unfavorable data. This tendency to resist data legitimacy is evident in both Study 1 and Study 2. Educators who encountered racial gaps in their course data were more likely to question the data's validity and avoid further exploration as demonstrated in Study 1. While, Study 2 indicated trust in the data was also lower when racial disparities were present.

This resistance to data legitimacy is a crucial issue that has been underresearched in education. Unlike algorithm aversion, which may have diminished in importance, the tendency to question the legitimacy of EDI data persists. To overcome this barrier, it is essential to address concerns about data

validity, transparency, and relevance. By providing clear explanations, contextualizing data, and fostering collaboration, educational institutions can enhance trust in data-driven decision-making and promote the use of data for EDI initiatives.

6.1.3 Addressing Non-Equity-Minded Frameworks

Existing evaluation frameworks may not be sufficiently equipped to identify or address racial inequities in student learning [Ziskin and Young, 2023]. As discussed in Chapter 3, we proposed two novel metrics centered on equality and inclusiveness to address these limitations. Our findings highlight the shortcomings of traditional metrics, which often conceal issues related to equity, diversity, and inclusion (EDI). These metrics frequently overlook the experiences of underrepresented student groups, who are disproportionately affected by inequities.

To mitigate these challenges, we designed our new metrics to incorporate weighting mechanisms that ensure small sample sizes do not obscure important trends. Additionally, recognizing the multifaceted nature of EDI, we developed two distinct metrics: course equality, which focuses on performance disparities, and course inclusiveness, which addresses issues of composition and belonging. The gamification-resistant design of these metrics further safeguards against superficial improvements that may mask underlying inequities. For instance, increasing course inclusiveness by enrolling a diverse student body without providing adequate support could lead to a spike in course equality, indicating a lack of equitable outcomes.

This issue is also connected to the broader landscape of educational data visualizations, as explored in Chapter 2. While there has been significant rhetoric surrounding the need to improve educational climates for underrepresented students following the events of 2020, research on designing visualizations to address these inequities has been limited. Consequently, many universities have developed ineffective dashboards that either fail to drive meaningful change or inadvertently contribute to the narrative of reverse discrimination.

6.1.4 Addressing Inequitable Processes

The process of collecting, analyzing, and interpreting educational data can inadvertently perpetuate existing racial biases, hindering efforts to address inequities [Ziskin and Young, 2023]. This issue is particularly relevant in the context of data-driven decision-making. The increasing reliance on data in education has raised concerns among students and instructors regarding surveillance ethics and the potential for biased outcomes.

Our research, while not explicitly focused on surveillance ethics, encountered educators who expressed concerns about the large-scale data collection ecosystem and its implications for privacy and equity. The growing use of data has also sparked debates about data ownership and the right of individuals to opt out of data collection. In Chapter 2, we observed educators who "disconnected" their courses from data collection due to concerns about evaluation metrics. While this may address individual concerns, it can also undermine institutional efforts to improve teaching and learning.

Chapter 5 further revealed the complexities of data sharing among educa-

tors. While many instructors expressed interest in benchmarking their performance against peers, they were reluctant to share individual course data due to privacy concerns. This suggests that while educators desire information about their peers' practices, they also harbor concerns about potential negative consequences associated with poor performance data.

Addressing these challenges requires a delicate balance between privacy concerns and the need for data-driven decision-making. While surveillance concerns are valid, particularly for marginalized groups [Birhane et al., 2022], opting out may not be a viable solution in educational contexts. In fact, it could exacerbate inequities if those who opt out are disproportionately not from marginalized groups. To ensure fairness, evaluation metrics must be applied consistently across all educators, regardless of their participation in data collection.

6.1.5 Addressing Lack of Structure

Institutions often lack the necessary structures to support educators in effectively using data to develop and implement equity-focused interventions. This deficiency was evident throughout the previous chapters, from the absence of research support for creating EDI dashboards (see Chapter 2) to the lack of confidence and trust in data-driven decision-making when no support is given (see Chapter 3). Additionally, educators expressed confusion about how to use data to evaluate and implement EDI interventions (see Chapter 5).

Organizational climate plays a crucial role in facilitating or hindering the use of data. As outlined in the Unified Theory of Acceptance and Use of Technol-

ogy (UTAUT), the "Facilitating Conditions" construct highlights the importance of providing educators with the necessary support and knowledge to leverage data effectively. While our study found that instructors generally found the data usable and easy to use, other factors, such as a lack of support, prevented some from fully utilizing it. Many educators expressed a desire for peer support without fear of judgment or consequences. This underscores the need for institutions to establish organizational supports to guide educators in using data and to address the potential challenges associated with failed implementations.

6.2 Key Recommendations

6.2.1 EDI Data Literacy

While the findings from Chapter 4 suggest a potential decline in AI literacy as a barrier to data-driven decision-making, Chapter 5 highlights that there still is a gap between data and action. To address this, future research and practice should prioritize the development of EDI data literacy. Educators are increasingly expected to use data to inform their decisions, often without adequate preparation in critical inquiry skills. This can lead to misinterpreting data as objective and neglecting structural factors that influence student outcomes [Bertrand and Marsh, 2015, Datnow and Park, 2018]. Educators face significant challenges in navigating these complex issues in the current sociopolitical climate, characterized by competing discourses on equity and individual rights.

To effectively use data for equity, educators must develop a critical data-

driven decision-making (CDDDM) mindset [Dodman et al., 2023]. This mindset involves not only the ability to gather, analyze, and interpret data but also the capacity to critically examine systemic and institutional factors that contribute to inequitable outcomes. A key component of CDDDM is Data Use for Equity (DUE), which combines culturally relevant data literacy (CRDL) and equity literacy. CRDL emphasizes the importance of understanding the cultural context of data and applying a holistic perspective to data analysis. Equity literacy, on the other hand, involves recognizing and addressing systemic biases and power imbalances. Educators can use data to identify and challenge inequities and implement evidence-based interventions by developing these skills.

Dodman et al.'s [2023] iterative approach engages educators in the skills needed to develop DUE via the School and Classroom Equity Audit (SCEA), a process that allows professional development to be designed to address opportunity gaps identified in its results specifically. To uncover inequities within a school as they relate to various identities, an SCEA collects demographic data to be disaggregated (e.g., race, gender, ability, sexual identity, language) as well as details of achievement, discipline, extracurriculars, and staffing. The equity audit aims to explore the existence of disproportionalities, analyze the differences, and then implement and monitor a solution and its outcomes. The equity audit offers educators an active approach to increasing their data understanding and skill while enhancing their awareness of equity issues and empowering them to act as change agents. Engagement with an SCEA has shown evidence of triggering new or increased sensemaking and noticing from educators and primed them to start hard conversations about equity by using data. This action-oriented approach to professional development provides educators with crucial skills needed to participate in increasingly data-driven educational

environments.

6.2.2 Personalized Instructor Feedback for Generative AI

The findings from Chapter 4 underscore a significant trend: Educators are increasingly adopting AI tools in their teaching methods. This shift not only underscores the potential of generative AI but also presents a unique opportunity to develop innovative solutions aimed at addressing the critical issues of EDI within the educational landscape. The potential of generative AI to revolutionize traditional teaching methods is already being realized in education. Recent studies have explored how generative AI can significantly enhance various aspects of the educational experience, such as providing personalized instruction tailored to individual learning needs, automating mundane administrative tasks, and improving accessibility for students with diverse needs [Ruiz et al., 2024]. When used thoughtfully and strategically, generative AI has the potential to act as a collaborative tool, enriching teaching practices and enhancing overall learning experiences [Kshetri, 2023].

Furthermore, the use of generative AI not only opens up new avenues for advancing EDI initiatives across diverse sectors but also underscores the pressing need for its responsible implementation. Recent applications of generative AI have included initiatives focused on combating systemic racism [Gabriel et al., 2024], creating inclusive role-play simulations to foster empathy and understanding [Holtham, 2023], and promoting equity within STEM teams to ensure diverse participation and representation (Nixon et al., 2024). To ensure these tools are used responsibly, researchers have proposed various frame-

works aimed at establishing DEI safeguards, particularly in chatbots [Abdelhalim et al., 2024]. Such frameworks can serve as a foundation for expanding the use of generative AI in educational settings, helping to design tools that provide educators with personalized feedback regarding their EDI practices. By integrating the metrics developed in Chapter 3 along with insights gained from Chapter 5, educational tools can assist teachers in identifying specific areas for improvement and providing tailored recommendations for action. This targeted approach could significantly contribute to fostering a culture of continuous learning and professional growth among educators.

However, it is crucial to approach the deployment of generative AI with caution. Potential challenges must be addressed, including issues related to bias, the spread of misinformation, and the possible erosion of critical thinking skills among students. To navigate these risks effectively, researchers and educators need to collaborate to develop comprehensive guidelines and best practices for the ethical and effective use of AI technology in educational settings. By prioritizing ethical considerations and promoting awareness of the limitations of these tools, the educational community can work towards maximizing the benefits of generative AI while minimizing its risks. Ultimately, by embracing the potential of generative AI, the education sector can strive toward transformative advancements that promote equity, diversity, and inclusion, ensuring a more inclusive and equitable learning environment for all students.

CHAPTER 7

CONCLUSION

When I embarked on this research, the educational landscape was undergoing significant transformation. A global pandemic, a renewed focus on racial justice, and rapid technological advancements were reshaping the field. While these challenges presented obstacles, they also created unique opportunities for exploration. For instance, the emergence of large language models (LLMs) like ChatGPT and Gemini transformed perceptions of AI tools within a short period, instilling a sense of hope and optimism for the future of education. While Study 1 in chapter 3 was conducted before this shift, Study 2 captured this evolving sentiment. This rapid change highlights the potential of new AI tools in education.

Similarly, the initial momentum for improving EDI initiatives following the events of 2020 began to wane. The rise of anti-DEI legislation and attacks on affirmative action created a challenging environment for these efforts. Chapter 5 sheds light on some factors contributing to this shift, particularly the superficial nature of many EDI initiatives that did not include appropriate support structures for educators. This research underscores the importance of centering the needs of historically marginalized students and avoiding tokenistic approaches that help majority groups feel better but may inadvertently exacerbate inequities.

While the challenges outlined in this dissertation may seem daunting, I remain optimistic about the future of education. Understanding the complex interplay of social, technological, and institutional factors, we can develop more effective data-driven solutions to address EDI issues. The growing apprecia-

tion for AI tools and a focus on equity present an opportunity to leverage technology to personalize instructional feedback and create more inclusive educational environments. Ultimately, this research calls for a nuanced approach to data-driven decision-making in education. By acknowledging the social and emotional dimensions of data use, we can develop tools that not only identify inequities but also empower educators to take action.

BIBLIOGRAPHY

Esraa Abdelhalim, Kemi Salawu Anazodo, Nazha Gali, and Karen Robson. A framework of diversity, equity, and inclusion safeguards for chatbots. *Business Horizons*, 67(5):487–498, September 2024. ISSN 00076813. doi: 10.1016/j.bushor.2024.03.003. 00012.

Stephen J. Aguilar. Guidelines and tools for promoting digital equity. *Information and Learning Sciences*, 121(5/6):285–299, June 2020. ISSN 2398-5348, 2398-5348. doi: 10.1108/ILS-04-2020-0084.

Stephen J Aguilar, Stuart A Karabenick, Stephanie D Teasley, and Clare Baek. Associations between learning analytics dashboard exposure and motivation and self-regulated learning. *Computers & Education*, 162:104085, March 2021. ISSN 03601315. doi: 10.1016/j.compedu.2020.104085.

Sara Ahmed. *Living a feminist life*. Duke University Press, Durham, 2017. ISBN 978-0-8223-6304-0.

June Ahn, Fabio Campos, Maria Hays, and Daniela DiGiacombo. Designing in context: Reaching beyond usability in learning analytics dashboard design. *Journal of Learning Analytics*, 6(2):70–85, August 2019. ISSN 19297750. doi: 10.18608/jla.2019.62.5.

Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, December 1991. ISSN 07495978. doi: 10.1016/0749-5978(91)90020-T.

Clement Chimezie Aladi. It higher education teachers and trust in ai-enabled ed-tech: Implications for adoption of ai in higher education. In *Proceedings of the 2024 Computers and People Research Conference*, page 1–16, Murfreesboro TN

USA, May 2024. ACM. ISBN 9798400704772. doi: 10.1145/3632634.3655852. URL https://dl.acm.org/doi/10.1145/3632634.3655852.

Naif Radi Aljohani, Ali Daud, Rabeeh Ayaz Abbasi, Jalal S Alowibdi, Mohammad Basheri, and Muhammad Ahtisham Aslam. An integrated framework for course adapted student learning analytics dashboard. *Computers in Human Behavior*, 92:679–690, March 2019. ISSN 07475632. doi: 10.1016/j.chb.2018.03.035.

Ishari Amarasinghe, Davinia Hernandez-Leo, Konstantinos Michos, and Milica Vujovic. An actionable orchestration dashboard to enhance collaboration in the classroom. *IEEE Transactions on Learning Technologies*, 13(4):662–675, October 2020. ISSN 1939-1382. doi: 10.1109/TLT.2020.3028597.

Evan P. Apfelbaum, Samuel R. Sommers, and Michael I. Norton. Seeing race and seeming racist? evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology*, 95(4):918–932, October 2008. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0011990.

Evan P. Apfelbaum, Michael I. Norton, and Samuel R. Sommers. Racial color blindness: Emergence, practice, and implications. *Current Directions in Psychological Science*, 21(3):205–209, June 2012. ISSN 0963-7214, 1467-8721. doi: 10.1177/0963721411434980. 00572.

Amara Atif, Deborah Richards, Danny Liu, and Ayse Aysin Bilgin. Perceived benefits and barriers of a prototype early alert system to detect engagement and support 'at-risk' students: The teacher perspective. *Computers & Education*, 156:103954, October 2020. ISSN 03601315. doi: 10.1016/j.compedu.2020.103954.

Debra Austin. Leadership lapse: Laundering systemic bias through student evaluations. *Vill. L. Rev.*, 65:995, 2020.

Musa Adekunle Ayanwale, Owolabi Paul Adelana, and Tolulope Timothy Odufuwa. Exploring steam teachers' trust in ai-based educational technologies: a structural equation modelling approach. *Discover Education*, 3(1):44, April 2024. ISSN 2731-5525. doi: 10.1007/s44217-024-00092-z.

Ryan S. Baker, Lief Esbenshade, Jonathan Vitale, and Shamya Karumbaiah. Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining*, 15(22):22–52, June 2023. ISSN 2157-2100. doi: 10.5281/zenodo.7702628.

Vaclav Bayer, Paul Mulholland, Martin Hlosta, Tracie Farrell, Christothea Herodotou, and Miriam Fernandez. Co-creating an equality diversity and inclusion learning analytics dashboard for addressing awarding gaps in higher education. *British Journal of Educational Technology*, 55(5):2058–2074, September 2024. ISSN 0007-1013, 1467-8535. doi: 10.1111/bjet.13509.

Estela Mara Bensimon. The diversity scorecard: A learning approach to institutional change. *Change: The Magazine of Higher Learning*, 36(1):44–52, January 2004. ISSN 0009-1383. doi: 10.1080/00091380409605083.

Melanie Bertrand and Julie A. Marsh. Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, 52(5):861–893, October 2015. ISSN 0002-8312, 1935-1011. doi: 10.3102/0002831215599251. 00300.

Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Trans-*

parency, page 948–958, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533157. URL https://dl.acm.org/doi/10.1145/3531146.3533157. 00091.

Blackboard. About us, 2021a. URL https://www.blackboard.com/en-uk/about-us.

Blackboard. Dashboard, 2021b. URL https://help.blackboard.com/Web{_}Community{_}Manager/Parent{_}Community{_}Member/Get{_}Started/Dashboard.

Oni J. Blackstock, Jessica E. Isom, and Rupinder K. Legha. Health care is the new battlefront for anti-DEI attacks. *PLOS Global Public Health*, 4(4):e0003131, April 2024. ISSN 2767-3375. doi: 10.1371/journal.pgph.0003131. URL https://dx.plos.org/10.1371/journal.pgph.0003131.

Paulo Blikstein and Marcelo Worsley. Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2):220–238, September 2016. ISSN 1929-7750. doi: 10.18608/jla.2016.32.11.

William Bloemer, Scott Day, and Karen Swan. Gap analysis: An innovative look at gateway courses and student retention. *Online Learning*, 21(3):5–14, 2017.

Robert Bodily and Katrien Verbert. Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, 10(4):405–418, October 2017. ISSN 1939-1382. doi: 10.1109/TLT.2017.2740172.

Robert Bodily, Tarah K Ikahihifo, Benjamin Mackley, and Charles R Graham. The design, development, and implementation of student-facing learning an-

alytics dashboards. *Journal of Computing in Higher Education*, 30(3):572–598, December 2018. ISSN 1042-1726. doi: 10.1007/s12528-018-9186-0.

Anne Boring, Kellie Ottoboni, and Philip B. Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 0(0), 2016. ISSN 2199-1006. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1. URL https://scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Lisa Bowleg. When black + lesbian + woman ≠ black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles*, 59(5–6):312–325, September 2008. ISSN 0360-0025, 1573-2762. doi: 10.1007/s11199-008-9400-z.

Stephen Boyd Davis, Olivia Vane, and Florian Kräutli. Can i believe what i see? data visualization and trust in the humanities. *Interdisciplinary Science Reviews*, 46(4):522–546, October 2021. ISSN 0308-0188, 1743-2790. doi: 10.1080/03080188.2021.1872874. 00027.

Tom Broos, Laurie Peeters, Katrien Verbert, Carolien Van Soom, Greet Langie, and Tinne De Laet. *Dashboard for Actionable Feedback on Learning Skills: Scalability and Usefulness*, page 229–241. Springer International Publishing, Cham, 2017a. doi: 10.1007/978-3-319-58515-4_18. URL http://link.springer.com/10.1007/978-3-319-58515-4{_}18.

Tom Broos, Katrien Verbert, Greet Langie, Carolien Van Soom, and Tinne De Laet. Small data as a conversation starter for learning analytics. *Journal of Research in Innovative Teaching & Learning*, 10(2):94–106, July 2017b. ISSN 2397-7604. doi: 10.1108/JRIT-05-2017-0010.

Tom Broos, Maarten Pinxten, Margaux Delporte, Katrien Verbert, and Tinne De Laet. Learning dashboards at scale: early warning and overall first year experience. *Assessment & Evaluation in Higher Education*, 45(6):855–874, August 2020. ISSN 0260-2938. doi: 10.1080/02602938.2019.1689546.

Michael Brown. Seeing students at scale: how faculty in large lecture courses act upon learning analytics dashboard data. *Teaching in Higher Education*, 25 (4):384–400, May 2020. ISSN 1356-2517. doi: 10.1080/13562517.2019.1698540.

Sylvain Bruni, Jessica Shenberger-Trujillo, Tim Clark, Isabel Erickson, Tatiana Toumbeva, and Sarah Meyer. Developing a human-centered corporate metric for inclusion, diversity, equity, & antiracism in a small business: A case study and lessons learned. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1):1451–1456, September 2023. ISSN 1071-1813, 2169-5067. doi: 10.1177/21695067231192523.

Simon Buckingham Shum, Rebecca Ferguson, and Roberto Martinez-Maldonado. Human-centred learning analytics. *Journal of Learning Analytics*, 6(2), July 2019. ISSN 1929-7750. doi: 10.18608/jla.2019.62.1. URL `https://learning-analytics.info/index.php/JLA/article/view/6627`.

David V. Budescu and Mia Budescu. How to measure diversity when you must. *Psychological Methods*, 17(2):215–227, 2012. ISSN 1939-1463, 1082-989X. doi: 10.1037/a0027129.

Linda Castañeda and Neil Selwyn. More than tools? making sense of the ongoing digitizations of higher education. *International Journal of Educational Technology in Higher Education*, 15(1):22, December 2018. ISSN 2365-9440. doi: 10.1186/s41239-018-0109-y.

Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, October 2019. ISSN 0022-2437, 1547-7193. doi: 10.1177/0022243719851788.

Rachelle Chadwick. On the politics of discomfort. *Feminist Theory*, 22 (4):556–574, December 2021. ISSN 1464-7001, 1741-2773. doi: 10.1177/1464700120987379.

Blerta Abazi Chaushi, Besnik Selimi, Agron Chaushi, and Marika Apostolova. *Explainable Artificial Intelligence in Education: A Comprehensive Review*, volume 1902 of *Communications in Computer and Information Science*, page 48–71. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-44066-3. doi: 10.1007/978-3-031-44067-0_3. URL https://link.springer.com/10.1007/978-3-031-44067-0_3.

Baiyun Chen, Stephen Sivo, Ryan Seilhamer, Amy Sugar, and Jin Mao. User acceptance of mobile technology: A campus-wide implementation of blackboard's mobile™ learn application. *Journal of Educational Computing Research*, 49(3):327–343, October 2013. ISSN 0735-6331, 1541-4140. doi: 10.2190/EC.49.3.c.

Shengnan Chen, Qifang Liu, and Bin He. A generative ai-based teaching material system using a human-in-the-loop model. In *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*, page 1–8, Wuhan, China, November 2023. IEEE. ISBN 9798350342895. doi: 10.1109/IEIR59294.2023.10391244. URL https://ieeexplore.ieee.org/document/10391244/.

Lingwei Cheng and Alexandra Chouldechova. Overcoming algorithm aversion: A comparison between process and outcome control. In *Proceedings of*

the 2023 CHI Conference on Human Factors in Computing Systems, page 1–27, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581253. URL https://dl.acm.org/doi/10.1145/3544548.3581253.

Seongyune Choi, Yeonju Jang, and Hyeoncheol Kim. Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human–Computer Interaction*, 39(4): 910–922, February 2023. ISSN 1044-7318, 1532-7590. doi: 10.1080/10447318.2022.2049145.

Ruth Cobos, Silvia Gil, Angel Lareo, and Francisco A Vargas. Open-dlas. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, page 265–268, New York, NY, USA, April 2016. ACM. ISBN 978-1-4503-3726-7. doi: 10.1145/2876034.2893430. URL https://dl.acm.org/doi/10.1145/2876034.2893430.

Brett Russell Coleman, Erin Beattie, Alina Raetz, and Kevin Wang. "Sorry It Took a Pandemic and Multiple Murders to Get Here": The Impact of Twin Pandemics on EDI in a Predominantly White School District. *Multicultural Perspectives*, 24(2):75–89, April 2022. ISSN 1521-0960, 1532-7892. doi: 10.1080/15210960.2022.2067856. URL https://www.tandfonline.com/doi/full/10.1080/15210960.2022.2067856.

Patricia Hill Collins. Intersectionality's definitional dilemmas. *Annual Review of Sociology*, 41(1):1–20, August 2015. ISSN 0360-0572. doi: 10.1146/annurev-soc-073014-112142.

Nicholas B. Colvard, C. Edward Watson, and Hyojin Park. The impact of open educational resources on various student success metrics. *International Journal*

of Teaching and Learning in Higher Education, 30(2):262–276, 2018. ISSN 1812-9129.

Mutlu Cukurova, Xin Miao, and Richard Brooker. *Adoption of Artificial Intelligence in Schools: Unveiling Factors Influencing Teachers' Engagement*, volume 13916 of *Lecture Notes in Computer Science*, page 151–163. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-36271-2. doi: 10.1007/978-3-031-36272-9_13. URL https://link.springer.com/10.1007/978-3-031-36272-9_13.

Miao Dai, Jui-Long Hung, Xu Du, Hengtao Tang, and Hao Li. Knowledge tracing: A review of available technologies. *Journal of Educational Technology Development and Exchange*, 14(2):1–20, 2021. ISSN 19418035. doi: 10.18785/jetde.1402.01.

Rafaan Daliri-Ngametua, Ian Hardy, and Sue Creagh. Data, performativity and the erosion of trust in teachers. *Cambridge Journal of Education*, 52(3):391–407, May 2022. ISSN 0305-764X, 1469-3577. doi: 10.1080/0305764x.2021.2002811.

Amanda Datnow and Vicki Park. Opening or closing doors for students? equity and data use in schools. *Journal of Educational Change*, 19(2):131–152, May 2018. ISSN 1389-2843, 1573-1812. doi: 10.1007/s10833-018-9323-6. 00188.

Yossi Ben David, Avi Segal, and Ya'akov (Kobi) Gal. Sequencing educational content in classrooms using bayesian knowledge tracing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, page 354–363, Edinburgh, United Kingdom, 2016. ACM Press. ISBN 978-1-4503-4190-5. doi: 10.1145/2883851.2883885. URL http://dl.acm.org/citation.cfm?doid=2883851.2883885.

Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319, September 1989. ISSN 02767783. doi: 10.2307/249008.

Robin DiAngelo. *White fragility: Why it's so hard for white people to talk about racism*. Seal Press, 2018. 00002.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114, 2015.

Stephanie L. Dodman, Elizabeth K. DeMulder, Jenice L. View, Stacia M. Stribling, and Rebecca Brusseau. "i knew it was a problem before, but did i really?": Engaging teachers in data use for equity. *Journal of Educational Change*, 24(4):995–1023, November 2023. ISSN 1389-2843, 1573-1812. doi: 10.1007/s10833-022-09477-z. 00007.

Hua Du, Yanchao Sun, Haozhe Jiang, A. Y. M. Atiquil Islam, and Xiaoqing Gu. Exploring the effects of ai literacy in teacher learning: an empirical study. *Humanities and Social Sciences Communications*, 11(1), May 2024.

Vanessa Echeverria, Roberto Martinez-Maldonado, Simon Buckingham Shum, Katherine Chiluiza, Roger Granda, and Cristina Conati. Exploratory versus explanatory visual learning analytics: Driving teachers' attention through educational data storytelling. *Journal of Learning Analytics*, 5(3), November 2018a. ISSN 1929-7750. doi: 10.18608/jla.2018.53.6. URL `https://learning-analytics.info/index.php/JLA/article/view/6114`.

Vanessa Echeverria, Roberto Martinez-Maldonado, Roger Granda, Katherine Chiluiza, Cristina Conati, and Simon Buckingham Shum. Driving data story-

telling from learning design. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, page 131–140, New York, NY, USA, March 2018b. ACM. ISBN 978-1-4503-6400-3. doi: 10.1145/3170358.3170380. URL https://dl.acm.org/doi/10.1145/3170358.3170380.

Juliana Elisa Raffaghelli. Is Data Literacy a Catalyst of Social Justice? A Response from Nine Data Literacy Initiatives in Higher Education. *Education Sciences*, 10(9):233, September 2020. ISSN 2227-7102. doi: 10.3390/educsci10090233. URL https://www.mdpi.com/2227-7102/10/9/233.

Julie Juola Exline, Anne L. Zell, and Marci Lobel. Sidestepping awkward encounters: avoidance as a response to outperformance-related discomfort. *Journal of Applied Social Psychology*, 43(4):706–720, April 2013. ISSN 0021-9029, 1559-1816. doi: 10.1111/j.1559-1816.2013.01047.x. 00028.

James Steven Fairweather. The mythologies of faculty productivity: Implications for institutional policy and decision making. *The Journal of Higher Education*, 73(1):26–48, 2002. ISSN 1538-4640. doi: 10.1353/jhe.2002.0006.

Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92, March 2006. ISSN 1609-4069, 1609-4069. doi: 10.1177/160940690600500107.

Alexandra N. Fisher, Danu Anthony Stinson, and Anastasija Kalajdzic. Unpacking backlash: Individual and contextual moderators of bias against female professors. *Basic and Applied Social Psychology*, 41(5):305–325, September 2019. ISSN 0197-3533, 1532-4834. doi: 10.1080/01973533.2019.1652178.

USC Center for Urban Education. Equity scorecard, 2021. URL https://cue.usc.edu/tools/the-equity-scorecard/.

Ed Foster and Rebecca Siddle. The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education*, 45(6):842–854, August 2020. ISSN 0260-2938. doi: 10.1080/02602938.2019.1682118.

Saadia Gabriel, Jessy Xinyi Han, Eric Liu, Isha Puri, Wonyoung So, Fotini Christia, Munzer Dahleh, Catherine D'Ignazio, Marzyeh Ghassemi, Peko Hosoi, and Devavrat Shah. Advancing equality: Harnessing generative ai to combat systemic racism. *An MIT Exploration of Generative AI*, March 2024. doi: 10.21428/e4baedd9.7dc53bbf. URL https://mit-genai.pubpub.org/pub/1ake7rfu. 00000.

Russell Golman, David Hagmann, and George Loewenstein. Information avoidance. *SSRN Electronic Journal*, 2015. ISSN 1556-5068. doi: 10.2139/ssrn.2633226. URL http://www.ssrn.com/abstract=2633226. 01012.

Jennifer Grace, Felix Simieou, Renée E. Lastrapes, and John Decman. Confronting the racism boogeyman: Educational leaders make meaning of the impact of George Floyd. *Education, Citizenship and Social Justice*, 19 (1):124–138, March 2024. ISSN 1746-1979, 1746-1987. doi: 10.1177/17461979221123014. URL http://journals.sagepub.com/doi/10.1177/17461979221123014.

Kimberly A. Griffin and Samuel D. Museus. Application of mixed-methods approaches to higher education and intersectional analyses. *New Directions for Institutional Research*, 2011(151):15–26, September 2011. ISSN 02710579. doi: 10.1002/ir.396.

Kimberly A Griffin, Jessica C Bennett, and Jessica Harris. Analyzing gender differences in black faculty marginalization through a sequential mixed-methods design. *New Directions for Institutional Research*, 2011(151):45–61, September 2011. ISSN 02710579. doi: 10.1002/ir.398.

Julio Guerra, Margarita Ortiz-Rojas, Miguel Angel Zúñiga-Prieto, Eliana Scheihing, Alberto Jiménez, Tom Broos, Tinne De Laet, and Katrien Verbert. Adaptation and evaluation of a learning analytics dashboard to improve academic support at three latin american universities. *British Journal of Educational Technology*, 51(4):973–1001, July 2020. ISSN 0007-1013. doi: 10.1111/bjet.12950.

Hongwen Guo, Matthew Johnson, Kadriye Ercikan, Luis Saldivia, and Michelle Worthington. Large-scale assessments for learning: A human-centred ai approach to contextualizing test performance. *Journal of Learning Analytics*, 11 (2):229–245, August 2024. ISSN 1929-7750. doi: 10.18608/jla.2024.8007.

Francisco Gutiérrez, Karsten Seipp, Xavier Ochoa, Katherine Chiluiza, Tinne De Laet, and Katrien Verbert. Lada: A learning analytics dashboard for academic advising. *Computers in Human Behavior*, 107:105826, June 2020. ISSN 07475632. doi: 10.1016/j.chb.2018.12.004.

David C. Haak, Janneke HilleRisLambers, Emile Pitre, and Scott Freeman. Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, 332(6034):1213–1216, June 2011. doi: 10.1126/science.1204820.

Jeongyun Han, Kwan Hoon Kim, Wonjong Rhee, and Young Hoan Cho. Learning analytics dashboards for adaptive support in face-to-face collaborative argumentation. *Computers & Education*, 163:104041, April 2021. ISSN 03601315. doi: 10.1016/j.compedu.2020.104041.

Judith M. Harackiewicz and Stacy J. Priniski. Improving student outcomes in higher education: The science of targeted intervention. *Annual Review of Psychology*, 69(1):409–435, January 2018. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev-psych-122216-011725.

Shaun R. Harper and Sylvia Hurtado. Nine themes in campus racial climates and implications for institutional transformation. *New Directions for Student Services*, 2007(120):7–24, 2007. ISSN 01647970, 15360695. doi: 10.1002/ss.254.

Elmer Harris, Rosemarie Allen, and Dorothy Shapland Rodriguez. Transformational learning practices: Embedding Equity Social Studies Methods for Preservice Teachers. *Black History Bulletin*, 87(1):6–11, March 2024. ISSN 2153-4810. doi: 10.1353/bhb.2024.a923021. URL https://muse.jhu.edu/article/923021.

Frank Harris and Estela Mara Bensimon. The equity scorecard: A collaborative approach to assess and respond to racial/ethnic disparities in student outcomes. *New Directions for Student Services*, 2007(120):77–84, 2007. ISSN 01647970. doi: 10.1002/ss.259.

Jennifer Heath and Eeva Leinonen. *An Institution Wide Approach to Learning Analytics*, page 73–87. IGI Global, Hershey, PA, 2016. URL http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-4666-9983-0.ch003.

Jan Hellings and Carla Haelermans. The effect of providing learning analytics on student behaviour and performance in programming: a randomised controlled experiment. *Higher Education*, 2020. ISSN 1573-174X. doi: 10.1007/s10734-020-00560-z. URL https://doi.org/10.1007/s10734-020-00560-z.

Christothea Herodotou, Bart Rienties, Avinash Boroowa, Zdenek Zdrahal, and Martin Hlosta. A large-scale implementation of predictive learning analytics in higher education: the teachers' role and perspective. *Educational Technology Research and Development*, 67(5):1273–1306, October 2019. ISSN 1042-1629, 1556-6501. doi: 10.1007/s11423-019-09685-0.

Christothea Herodotou, Bart Rienties, Martin Hlosta, Avinash Boroowa, Chrysoula Mangafa, and Zdenek Zdrahal. The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study. *The Internet and Higher Education*, 45:100725, April 2020. ISSN 10967516. doi: 10.1016/j.iheduc.2020.100725.

Christothea Herodotou, Claire Maguire, Martin Hlosta, and Paul Mulholland. Predictive learning analytics and university teachers: Usage and perceptions three years post implementation. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, page 68–78, Arlington TX USA, March 2023. ACM. ISBN 978-1-4503-9865-7. doi: 10.1145/3576050.3576061. URL https://dl.acm.org/doi/10.1145/3576050.3576061.

Yann Hicke. Knowledge tracing challenge: Optimal activity sequencing for students. 2023. doi: 10.48550/ARXIV.2311.14707. URL https://arxiv.org/abs/2311.14707.

Isabel Hilliger, Mar Pérez-Sanagustín, Ronald Pérez-Álvarez, Valeria Henríquez, Julio Guerra, Miguel Ángel Zuñiga-Prieto, Margarita Ortiz-Rojas, Yi-Shan Tsai, Dragan Gasevic, Pedro J. Muñoz-Merino, Tom Broos, and Tinne De Laet. *Leadership and Maturity: How Do They Affect Learning Analytics Adoption in Latin America?*,

page 305–326. 2020. doi: 10.1007/978-3-030-47392-1_16. URL http://link.springer.com/10.1007/978-3-030-47392-1{_}16.

Kenneth Holstein, Bruce M. McLaren, and Vincent 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*, page 257–266, Vancouver British Columbia Canada, March 2017. ACM. ISBN 978-1-4503-4870-6. doi: 10.1145/3027385.3027451. URL https://dl.acm.org/doi/10.1145/3027385.3027451.

Clive Holtham. Deploying generative ai to draft a roleplay simulation of difficult conversations about inclusivity. *Irish Journal of Technology Enhanced Learning*, 7(2):146–157, December 2023. ISSN 2009-972X. doi: 10.22554/ijtel.v7i2.127. 00000.

Matthew T. Hora, Jana Bouwma-Gearhart, and Hyoung Joon Park. Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning. *The Review of Higher Education*, 40(3):391–426, 2017. ISSN 1090-7009. doi: 10.1353/rhe.2017.0013.

Joel A Howell, Lynne D Roberts, Kristen Seaman, and David C Gibson. Are we on our way to becoming a "helicopter university"? academics' views on learning analytics. *Technology, Knowledge and Learning*, 23(1):1–20, April 2018. ISSN 2211-1662. doi: 10.1007/s10758-017-9329-9.

Katharine Elizabeth Hubbard. Institution level awarding gap metrics for identifying educational inequity: useful tools or reductive distractions? *Higher Education*, April 2024. ISSN 0018-1560, 1573-174X. doi: 10.1007/s10734-024-01216-y. URL https://link.springer.com/10.1007/s10734-024-01216-y.

IES. About ies: Connecting research, policy and practice, 2021. URL https://ies.ed.gov/aboutus/.

Instructure. Our company story, 2021a. URL https://www.instructure.com/about/our-story.

Instructure. How do i use the dashboard as a student?, 2021b. URL https://community.canvaslms.com/t5/Student-Guide/How-do-I-use-the-Dashboard-as-a-student/ta-p/512{#}:{%}7B{~}{%}7D:text=TheDashboardisthefirst,DashboardlinkinGlobalNavigation.

Leanna Ireland. Who errs? algorithm aversion, the source of judicial error, and public support for self-help behaviors. *Journal of Crime and Justice*, 43(2): 174–192, March 2020. ISSN 0735-648X, 2158-9119. doi: 10.1080/0735648X.2019.1655781.

Philip L. Jackson, Eric Brunet, Andrew N. Meltzoff, and Jean Decety. Empathy examined through the neural mechanisms involved in imagining how i feel versus how you feel pain. *Neuropsychologia*, 44(5):752–761, January 2006. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2005.07.015. 01207.

Ioana Jivet, Maren Scheffel, Marcus Specht, and Hendrik Drachsler. License to evaluate. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, page 31–40, New York, NY, USA, March 2018. ACM. ISBN 978-1-4503-6400-3. doi: 10.1145/3170358.3170421. URL https://dl.acm.org/doi/10.1145/3170358.3170421.

David A Joyner, May Carlon, Jeffrey Cross, Eduardo Corpeño, Rocael Hernández Rizzardini, Oscar Rodas, Dhawal Shah, Manoel Cortes-Mendez,

Thomas Staubitz, and José A Ruipérez-Valiente. Global learning @ scale. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, page 229–232, New York, NY, USA, August 2020. ACM. ISBN 978-1-4503-7951-9. doi: 10.1145/3386527.3405956. URL https://dl.acm.org/doi/10.1145/3386527.3405956.

Yuze Kang. Algorithm aversion and self-driving cars:. Zhuhai, China, 2022. doi: 10.2991/aebmr.k.220307.077. URL https://www.atlantis-press.com/article/125971967.

Esther Kaufmann. Algorithm appreciation or aversion? comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence*, 2:100028, 2021. ISSN 2666920X. doi: 10.1016/j.caeai.2021.100028.

K. Kelly, Yan Wang, Tamisha Thompson, and N. Heffernan. Defining mastery: Knowledge tracing versus n- consecutive correct responses. In *Educational Data Mining*, 2015.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. Evaluating the diversity, equity and inclusion of nlp technology: A case study for indian languages. (arXiv:2205.12676), April 2023. URL http://arxiv.org/abs/2205.12676.

Jeonghyun Kim, Il-Hyun Jo, and Yeonjeong Park. Effects of learning analytics dashboard: analyzing the relations among dashboard utilization, satisfaction, and learning achievement. *Asia Pacific Education Review*, 17(1):13–24, March 2016. ISSN 1598-1037. doi: 10.1007/s12564-015-9403-8.

Joe L Kincheloe. *Critical Constructivism Primer*, volume 2. Peter Lang, Bern, Switzerland, 2005.

René F. Kizilcec. To advance ai use in education, focus on understanding educators. *International Journal of Artificial Intelligence in Education*, 34(1):12–19, March 2024. ISSN 1560-4292, 1560-4306. doi: 10.1007/s40593-023-00351-4.

Carrie Klein, Jaime Lester, Thien Nguyen, Abigail Justen, Huzefa Rangwala, and Aditya Johri. Student sensemaking of learning analytics dashboard interventions in higher education. *Journal of Educational Technology Systems*, 48(1): 130–154, September 2019. ISSN 0047-2395. doi: 10.1177/0047239519859854.

Sean Kross and Philip J Guo. Students, systems, and interactions. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, page 1–10, New York, NY, USA, June 2018. ACM. ISBN 978-1-4503-5886-6. doi: 10.1145/3231644.3231662. URL `https://dl.acm.org/doi/10.1145/3231644.3231662`.

Nir Kshetri. The future of education: Generative artificial intelligence's collaborative role with teachers. *IT Professional*, 25(6):8–12, November 2023. ISSN 1520-9202, 1941-045X. doi: 10.1109/MITP.2023.3333070. 00000.

Ryan C LaBrie, Gerhard H Steinke, Xiangmin Li, and Joseph A Cazier. Big data analytics sentiment: Us-china reaction to data collection by business and government. *Technological Forecasting and Social Change*, 130:45–55, May 2018. ISSN 00401625. doi: 10.1016/j.techfore.2017.06.029.

Philipp Leitner and Martin Ebner. *Development of a Dashboard for Learning Analytics in Higher Education*, page 293–301. Springer International

Publishing, Cham, 2017. URL http://link.springer.com/10.1007/978-3-319-58515-4{_}23.

Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. Moral machines or tyranny of the majority? a systematic review on predictive bias in education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, page 499–508, Arlington TX USA, March 2023. ACM. ISBN 978-1-4503-9865-7. doi: 10.1145/3576050.3576119. URL https://dl.acm.org/doi/10.1145/3576050.3576119.

Qiujie Li, Yeonji Jung, and Alyssa Friend Wise. *Beyond First Encounters with Analytics: Questions, Techniques and Challenges in Instructors' Sensemaking*, page 344–353. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 978-1-4503-8935-8. URL https://doi.org/10.1145/3448139.3448172.

Lisa-Angelique Lim, Shane Dawson, Srecko Joksimovic, and Dragan Gašević. Exploring students' sensemaking of learning analytics dashboards. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, page 250–259, New York, NY, USA, March 2019. ACM. ISBN 978-1-4503-6256-6. doi: 10.1145/3303772.3303804. URL https://dl.acm.org/doi/10.1145/3303772.3303804.

Lisa-Angelique Lim, Shane Dawson, Dragan Gašević, Srecko Joksimović, Abelardo Pardo, Anthea Fudge, and Sheridan Gentili. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. *Assessment & Evaluation in Higher*

Education, 0(0):1–21, June 2020. ISSN 0260-2938. doi: 10.1080/02602938.2020.1782831.

Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, March 2019. ISSN 07495978. doi: 10.1016/j.obhdp.2018.12.005.

Brian S. Lowery, Eric D. Knowles, and Miguel M. Unzueta. Framing inequity safely: Whites' motivated perceptions of racial privilege. *Personality and Social Psychology Bulletin*, 33(9):1237–1250, September 2007. ISSN 0146-1672, 1552-7433. doi: 10.1177/0146167207303016.

Jeffrey Thomas Ludwig. A new mathematical metric for inclusive excellence in teaching applied before and during the covid-19 era. *International Journal of Education*, 13(2):1, May 2021. ISSN 1948-5476. doi: 10.5296/ije.v13i2.18466.

Julie C Luecke. The gender facilitative school: Advocating authenticity for gender expansive children in pre-adolescence. *Improving Schools*, 21(3):269–284, November 2018. ISSN 1365-4802, 1475-7583. doi: 10.1177/1365480218791881.

Hasan Mahmud and Najmul Islam. *The Role of Algorithm and Task Familiarity in Algorithm Aversion: An Empirical Study*, volume 14316 of *Lecture Notes in Computer Science*, page 3–13. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-50039-8. doi: 10.1007/978-3-031-50040-4_1. URL `https://link.springer.com/10.1007/978-3-031-50040-4_1`.

Nikola Marangunić and Andrina Granić. Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information So-*

ciety, 14(1):81–95, March 2015. ISSN 1615-5289, 1615-5297. doi: 10.1007/s10209-014-0348-1.

Wannisa Matcha, Noraayu Ahmad Uzir, Dragan Gasevic, and Abelardo Pardo. A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, 13(2):226–245, April 2020. ISSN 1939-1382. doi: 10.1109/TLT.2019.2916802.

Kieran Mathieson, Eileen Peacock, and Wynne W. Chin. Extending the technology acceptance model: the influence of perceived user resources. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 32(3):86–112, July 2001. ISSN 0095-0033, 1532-0936. doi: 10.1145/506724.506730.

Rebecca L. Matz, Benjamin P. Koester, Stefano Fiorini, Galina Grom, Linda Shepard, Charles G. Stangor, Brad Weiner, and Timothy A. McKay. Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, 3(4):233285841774375, October 2017. ISSN 2332-8584, 2332-8584. doi: 10.1177/2332858417743754.

Matthew J. Mayhew and Sonia Deluca. Fernández. Pedagogical practices that contribute to social justice outcomes. *The Review of Higher Education*, 31(1):55–80, 2007. ISSN 1090-7009. doi: 10.1353/rhe.2007.0055.

Meyliana, Henry A E Widjaja, and Stephen W Santoso. University dashboard: An implementation of executive dashboard to university. In *2014 2nd International Conference on Information and Communication Technology (ICoICT)*, page 282–287, Bandung, Indonesia, May 2014. IEEE. ISBN 978-1-4799-3581-9. doi: 10.1109/ICoICT.2014.6914080. URL `http://ieeexplore.ieee.org/document/6914080/`.

Martijn Millecamp, Francisco Gutiérrez, Sven Charleer, Katrien Verbert, and Tinne De Laet. A qualitative evaluation of a learning dashboard to support advisor-student dialogues. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, page 56–60, New York, NY, USA, March 2018. ACM. ISBN 978-1-4503-6400-3. doi: 10.1145/3170358.3170417. URL https://dl.acm.org/doi/10.1145/3170358.3170417.

Charles W. Mills. *The Racial Contract*. Cornell University Press, December 2019. ISBN 978-0-8014-7135-3. doi: 10.7591/9780801471353. URL https://www.degruyter.com/document/doi/10.7591/9780801471353/html.

H. Richard Milner. Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher*, 36(7):388–400, October 2007. ISSN 0013-189X. doi: 10.3102/0013189X07309471.

Donald Mitchell, Tiffany Steele, Jakia Marie, and Kathryn Timm. Learning race and racism while learning: Experiences of international students pursuing higher education in the midwestern united states. *AERA Open*, 3 (3):233285841772040, July 2017. ISSN 2332-8584, 2332-8584. doi: 10.1177/2332858417720402.

Akira Miyake, Lauren E. Kost-Smith, Noah D. Finkelstein, Steven J. Pollock, Geoffrey L. Cohen, and Tiffany A. Ito. Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330(6008): 1234–1237, November 2010. doi: 10.1126/science.1195996.

Shalina Nair, José E. Rodríguez, Samantha Elwood, Elisabeth Wilson, Annamalai Ramanathan, Debra Stulberg, Belinda Vail, Kristen Rundell, and C. J. Peek. Departmental metrics to guide equity, diversity, and inclusion for aca-

demic family medicine departments. *Family Medicine*, 56(6):362–366, June 2024. ISSN 0742-3225, 1938-3800. doi: 10.22454/FamMed.2024.865619.

Tanya Nazaretsky, Mutlu Cukurova, Moriah Ariely, and Giora Alexandron. Confirmation bias and trust: human factors that influence teachers' attitudes towards ai-based educational technology. In *CEUR Workshop Proceedings*, volume 3042, 2021.

Tanya Nazaretsky, Mutlu Cukurova, and Giora Alexandron. An instrument for measuring teachers' trust in ai-based educational technology. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, page 56–66, Online USA, March 2022. ACM. ISBN 978-1-4503-9573-1. doi: 10.1145/3506860.3506866. URL https://dl.acm.org/doi/10.1145/3506860.3506866.

Jeroen Ooge, Shotallo Kato, and Katrien Verbert. Explaining recommendations in e-learning: Effects on adolescents' trust. In *27th International Conference on Intelligent User Interfaces*, page 93–105, Helsinki Finland, March 2022. ACM. ISBN 978-1-4503-9144-3. doi: 10.1145/3490099.3511140. URL https://dl.acm.org/doi/10.1145/3490099.3511140.

Eric L. Oslund, Amy M. Elleman, and Kelli Wallace. Factors related to data-based decision-making: Examining experience, professional development, and the mediating effect of confidence on teacher graph literacy. *Journal of Learning Disabilities*, 54(4):243–255, July 2021. ISSN 0022-2194, 1538-4780. doi: 10.1177/0022219420972187.

Grace Pai. Beyond demographic identifiers and classification: A search for alternative approaches to quantitatively evaluating educational equity. *Current Issues in Comparative Education*, 25(1):121–126, 2023.

Guy Paré, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2):183–199, March 2015. ISSN 03787206. doi: 10.1016/j.im.2014.08.008.

Leigh Patel. Desiring diversity and backlash: White property rights in higher education. *The Urban Review*, 47(4):657–675, 2015.

Lori D Patton, Berenice Sánchez, Jacqueline Mac, and D-L Stewart. An inconvenient truth about "progress": An analysis of the promises and perils of research on campus diversity initiatives. *The Review of Higher Education*, 42(5): 173–198, 2019. ISSN 1090-7009. doi: 10.1353/rhe.2019.0049.

Elizabethe Payne and Melissa Smith. Power, emotion, and privilege: "discomfort" as resistance to transgender student affirmation. *Teachers College Record: The Voice of Scholarship in Education*, 124(8):43–65, August 2022. ISSN 0161-4681, 1467-9620. doi: 10.1177/01614681221121521.

Agoritsa Polyzou, Maria Kalantzi, and George Karypis. Faireo: User group fairness for equality of opportunity in course recommendation. 2021. doi: 10.48550/ARXIV.2109.05931. URL `https://arxiv.org/abs/2109.05931`.

Paul Prinsloo and Sharon Slade. Educational triage in open distance learning: Walking a moral tightrope. *The International Review of Research in Open and Distributed Learning*, 15(4), August 2014. ISSN 1492-3831. doi: 10.19173/irrodl.v15i4.1881. URL `http://www.irrodl.org/index.php/irrodl/article/view/1881`.

Fen Qin, Kai Li, and Jianyuan Yan. Understanding user trust in artificial intelligence-based educational systems: Evidence from china. *British Journal*

of Educational Technology, 51(5):1693–1710, September 2020. ISSN 0007-1013, 1467-8535. doi: 10.1111/bjet.12994.

David M. Quinn. Experimental effects of "achievement gap" news reporting on viewers' racial stereotypes, inequality explanations, and inequality prioritization. *Educational Researcher*, 49(7):482–492, October 2020. ISSN 0013-189X, 1935-102X. doi: 10.3102/0013189X20932469. 00084.

Ahmad Rahal and Mohamed Zainuba. Improving students' performance in quantitative courses: The case of academic motivation and predictive analytics. *The International Journal of Management Education*, 14(1):8–17, March 2016. ISSN 14728117. doi: 10.1016/j.ijme.2015.11.003.

Caroline Roberts, Emily Gilbert, Nick Allum, and Léïla Eisner. Research synthesis. *Public Opinion Quarterly*, 83(3):598–626, November 2019. ISSN 0033-362X, 1537-5331. doi: 10.1093/poq/nfz035. 00086.

Lynne D. Roberts, Joel A. Howell, and Kristen Seaman. Give me a customizable dashboard: Personalized learning analytics dashboards in higher education. *Technology, Knowledge and Learning*, 22(3):317–333, October 2017. ISSN 2211-1670. doi: 10.1007/s10758-017-9316-1.

Rachel Roegman. Central office foci and principal data use: A comparative study of equity-focused practice in six districts. *Education Policy Analysis Archives*, 28:181, December 2020. ISSN 1068-2341. doi: 10.14507/epaa.28.5304.

Pati Ruiz, Alessandra Rangel, and Merijke Coenraad. Using generative ai to support pk-12 teaching and learning: Developing sample lessons and more. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, page 1800–1801, Portland OR USA, March 2024. ACM. ISBN

9798400704246. doi: 10.1145/3626253.3635522. URL https://dl.acm.org/doi/10.1145/3626253.3635522.

Ronny Scherer, Fazilat Siddiq, and Jo Tondeur. The technology acceptance model (tam): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education*, 128:13–35, January 2019. ISSN 03601315. doi: 10.1016/j.compedu.2018.09.009.

Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. "there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1616–1628, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533218. URL https://dl.acm.org/doi/10.1145/3531146.3533218.

Judith Schoonenboom. Using an adapted, task-level technology acceptance model to explain why instructors in higher education intend to use some learning management system tools more than others. *Computers & Education*, 71:247–256, February 2014. ISSN 03601315. doi: 10.1016/j.compedu.2013.09.016.

Beat A. Schwendimann, Maria Jesus Rodriguez-Triana, Andrii Vozniuk, Luis P. Prieto, Mina Shirvani Boroujeni, Adrian Holzer, Denis Gillet, and Pierre Dillenbourg. Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, 10(1): 30–41, January 2017. ISSN 1939-1382. doi: 10.1109/TLT.2016.2599522.

Neil Selwyn. Re-imagining 'learning analytics' ... a case for starting again? *The*

Internet and Higher Education, 46:100745, July 2020. ISSN 10967516. doi: 10.1016/j.iheduc.2020.100745.

Enakshi Sengupta, Patrick Blessinger, Jaimie Hoffman, and Mandla Makhanya. *Introduction to Strategies for Fostering Inclusive Classrooms in Higher Education*, volume 16, page 3–16. Emerald Publishing Limited, February 2019. ISBN 978-1-78756-061-1. URL https://www.emerald.com/insight/content/doi/10.1108/S2055-364120190000016005/full/html.

Simon Buckingham Shum. Should predictive models of student outcome be "colour-blind"?, 2020. URL https://simon.buckinghamshum.net/2020/07/should-predictive-models-of-student-outcome-be-colour-blind/.

Jay Sloan-Lynch and Robert Morse. Equity-forward learning analytics: Designing a dashboard to support marginalized student success. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, page 1–11, Kyoto Japan, March 2024. ACM. ISBN 9798400716188. doi: 10.1145/3636555.3636844. URL https://dl.acm.org/doi/10.1145/3636555.3636844.

SoLAR. From solar executive committee: Statement of support and call for action, 2020. URL https://www.solaresearch.org/2020/06/statement-of-support-and-call-for-action/.

Jessaca Spybrook, Ran Shi, and Benjamin Kelcey. Progress in the past decade: an examination of the precision of cluster randomized trials funded by the u.s. institute of education sciences. *International Journal of Research & Method in Education*, 39(3):255–267, July 2016. ISSN 1743-727X. doi: 10.1080/1743727X.2016.1150454.

Kaiwen Sun, Abraham H Mhaidli, Sonakshi Watel, Christopher A Brooks, and Florian Schaub. It's my data! tensions among stakeholders of a learning analytics dashboard. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, page 1–14, New York, New York, USA, 2019. ACM Press. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300824. URL http://dl.acm.org/citation.cfm?doid=3290605.3300824.

Karen Swan, William Bloemer, Leonard Bogle, and Scott Day. Digging deeper into the data: The role of gateway courses in online student retention. *Online Learning*, 22(4), December 2018. ISSN 2472-5730, 2472-5749. doi: 10.24059/olj.v22i4.1515. URL https://olj.onlinelearningconsortium.org/index.php/olj/article/view/1515.

Kate Sweeny, Darya Melnyk, Wendi Miller, and James A. Shepperd. Information avoidance: Who, what, when, and why. *Review of General Psychology*, 14(4): 340–353, December 2010. ISSN 1089-2680, 1939-1552. doi: 10.1037/a0021288.00744.

Pengcheng Tang, Xuan Liu, Yao Hong, and Shuwang Yang. Moving beyond economic criteria: Exploring the social impact of green innovation from the stakeholder management perspective. *Corporate Social Responsibility and Environmental Management*, 30(3):1042–1052, May 2023. ISSN 1535-3958, 1535-3966. doi: 10.1002/csr.2401.

Beverly Daniel Tatum. The complexity of identity:"who am i?". *Readings for diversity and social justice*, 2:5–8, 2000.

Z. W. Taylor, Jase Kugiya, Chelseaia Charran, and Joshua Childs. Building Equitable Education Datasets for Developing Nations: Equity-Minded Data

Collection and Disaggregation to Improve Schools, Districts, and Communities. *Education Sciences*, 13(4):348, March 2023. ISSN 2227-7102. doi: 10.3390/educsci13040348. URL https://www.mdpi.com/2227-7102/13/4/348.

Rachel L. Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for ai. *Patterns*, 3(5):100476, May 2022. ISSN 26663899. doi: 10.1016/j.patter.2022.100476.

Dawit Negussie Tolossa, Dr. Jabe Bekele Hirgo, Dr. Bhavesh A. Prabhakar, and Yohannes Negussie. Advancing equity and inclusion in education: A bibliometric analysis. *International Journal of Research Publication and Reviews*, 4(9): 3399–3404, October 2023.

Noor Toraif, Neha Gondal, Pujan Paudel, and Alison Frisellaa. From colorblind to systemic racism: Emergence of a rhetorical shift in higher education discourse in response to the murder of George Floyd. *PLOS ONE*, 18(8): e0289545, August 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0289545. URL https://dx.plos.org/10.1371/journal.pone.0289545.

Kimberly A Truong and Kay Martinez. From dei to jedi, April 2021. URL https://www.diverseeducation.com/opinion/article/15109001/from-dei-to-jedi.

Ofir Turel and Shivam Kalhan. Prejudiced against the machine? implicit associations and the transience of algorithm aversion. *MIS Quarterly*, 47(4): 1369–1394, December 2023. ISSN 02767783, 21629730. doi: 10.25300/MISQ/2022/17961.

Timothy Tuti, Chris Paton, and Niall Winters. Learning to represent healthcare providers knowledge of neonatal emergency care. In *Proceedings of the Tenth*

International Conference on Learning Analytics & Knowledge, page 320–329, New York, NY, USA, March 2020. ACM. ISBN 978-1-4503-7712-6. doi: 10.1145/3375462.3375479. URL https://dl.acm.org/doi/10.1145/3375462.3375479.

Mark G. Urtel. Assessing academic performance between traditional and distance education course formats. *Journal of Educational Technology & Society*, 11 (1):322–330, 2008. ISSN 1176-3647.

Bob Uttl, Carmela A. White, and Daniela Wong Gonzalez. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54:22–42, September 2017. ISSN 0191491X. doi: 10.1016/j.stueduc.2016.08.007.

Ben Van Dusen and Jayson Nissen. Associations between learning assistants, passing introductory physics, and equity: A quantitative critical race theory investigation. *Physical Review Physics Education Research*, 16(1):010117, April 2020. ISSN 2469-9896. doi: 10.1103/PhysRevPhysEducRes.16.010117.

Venkatesh, Morris, Davis, and Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425, 2003. ISSN 02767783. doi: 10.2307/30036540.

Katrien Verbert, Erik Duval, Joris Klerkx, Sten Govaerts, and José Luis Santos. Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10):1500–1509, October 2013a. ISSN 0002-7642. doi: 10.1177/0002764213479363.

Katrien Verbert, Sten Govaerts, Erik Duval, Jose Luis Santos, Frans Van Assche, Gonzalo Parra, and Joris Klerkx. Learning dashboards: an overview and

future research opportunities. *Personal and Ubiquitous Computing*, November 2013b. ISSN 1617-4909, 1617-4917. doi: 10.1007/s00779-013-0751-2. URL http://link.springer.com/10.1007/s00779-013-0751-2.

Olga Viberg, Mutlu Cukurova, Yael Feldman-Maggor, Giora Alexandron, Shizuka Shirai, Susumu Kanemune, Barbara Wasson, Cathrine Tømte, Daniel Spikol, Marcelo Milrad, Raquel Coelho, and René F. Kizilcec. What explains teachers' trust of ai in education across six countries? 2023. doi: 10.48550/ARXIV.2312.01627. URL https://arxiv.org/abs/2312.01627.

Camilo Vieira, Paul Parsons, and Vetria Byrd. Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122:119–135, July 2018. ISSN 03601315. doi: 10.1016/j.compedu.2018.03.018.

Jacquie D. Vorauer and Stacey J. Sasaki. Distinct effects of imagine-other versus imagine-self perspective-taking on prejudice reduction. *Social Cognition*, 32(2):130–147, April 2014. ISSN 0278-016X. doi: 10.1521/soco.2014.32.2.130. 00051.

Edward Vul, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive Science*, 38 (4):599–637, May 2014. ISSN 0364-0213, 1551-6709. doi: 10.1111/cogs.12101.

Sherri L. Wallace, Angela K. Lewis, and Marcus D. Allen. The state of the literature on student evaluations of teaching and an exploratory analysis of written comments: Who benefits most? *College Teaching*, 67(1):1–14, January 2019. ISSN 8756-7555, 1930-8299. doi: 10.1080/87567555.2018.1483317.

Deliang Wang, Cunling Bian, and Gaowei Chen. Using explainable ai to unravel classroom dialogue analysis: Effects of explanations on teachers'

trust, technology acceptance and cognitive load. *British Journal of Educational Technology*, page bjet.13466, April 2024. ISSN 0007-1013, 1467-8535. doi: 10.1111/bjet.13466.

Kimberly Williamson and Rene Kizilcec. A review of learning analytics dashboard research in higher education: Implications for justice, equity, diversity, and inclusion. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, page 260–270, Online USA, March 2022. ACM. ISBN 978-1-4503-9573-1. doi: 10.1145/3506860.3506900. URL https://dl.acm.org/doi/10.1145/3506860.3506900.

Kimberly Williamson and René F Kizilcec. Effects of algorithmic transparency in bayesian knowledge tracing on trust and perceived accuracy. *International Educational Data Mining Society*, 2021.

Alyssa Friend Wise and Yeonji Jung. Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, 6 (2):53–69, July 2019. ISSN 1929-7750. doi: 10.18608/jla.2019.62.4.

Benjamin Xie. How data can support equity in computing education. *XRDS: Crossroads, The ACM Magazine for Students*, 27(2):48–52, December 2020. ISSN 1528-4972, 1528-4980. doi: 10.1145/3433136. URL https://dl.acm.org/doi/10.1145/3433136.

Lingrui Xu, Zachary A. Pardos, and Anirudh Pai. Convincing the expert: Reducing algorithm aversion in administrative higher education decision-making. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, page 215–225, Copenhagen Denmark, July 2023. ACM. ISBN 9798400700255. doi: 10.1145/3573051.3593378. URL https://dl.acm.org/doi/10.1145/3573051.3593378.

Jie Yang, Seth DeVore, Dona Hewagallage, Paul Miller, Qing X. Ryan, and John Stewart. Using machine learning to identify the most at-risk students in physics classes. *Physical Review Physics Education Research*, 16(2):020130, October 2020. ISSN 2469-9896. doi: 10.1103/PhysRevPhysEducRes.16.020130.

David S. Yeager and Gregory M. Walton. Social-psychological interventions in education: They're not magic. *Review of Educational Research*, 81(2):267–301, June 2011. ISSN 0034-6543, 1935-1046. doi: 10.3102/0034654311405999.

Sangseok You, Cathy Liu Yang, and Xitong Li. Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, 39(2):336–365, April 2022. ISSN 0742-1222, 1557-928X. doi: 10.1080/07421222.2022.2063553.

Renzhe Yu, Hansol Lee, and René F. Kizilcec. Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, page 91–100, Virtual Event Germany, June 2021. ACM. ISBN 978-1-4503-8215-1. doi: 10.1145/3430895.3460139. URL https://dl.acm.org/doi/10.1145/3430895.3460139.

Lisl Zach. When is "enough" enough? modeling the information-seeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology*, 56(1):23–35, January 2005. ISSN 1532-2882, 1532-2890. doi: 10.1002/asi.20092.

Michalinos Zembylas. 'pedagogy of discomfort' and its ethical implications: the tensions of ethical violence in social justice education. *Ethics and Education*, 10 (2):163–174, May 2015. ISSN 1744-9642, 1744-9650. doi: 10.1080/17449642.2015.1039274.

Michalinos Zembylas. *Affect, race, and white discomfort in schooling: Decolonial strategies for 'pedagogies of discomfort'*, pages 86–104. Routledge, 2020.

Qiao Zhang and Christopher J Maclellan. Going online: A simulated student approach for evaluating knowledge tracing in the context of mastery learning. *International Educational Data Mining Society*, 2021.

Juan Zheng, Lingyun Huang, Shan Li, Susanne P Lajoie, Yuxin Chen, and Cindy E Hmelo-Silver. Self-regulation and emotion matter: A case study of instructor interactions with a learning analytics dashboard. *Computers & Education*, 161:104061, February 2021. ISSN 03601315. doi: 10.1016/j.compedu.2020.104061.

Aayesha Zia, Jalal Nouri, Muhammad Afzaal, Yongchao Wu, Xiu Li, and Rebecka Weegar. A step towards improving knowledge tracing. In *2021 International Conference on Advanced Learning Technologies (ICALT)*, page 38–39, Tartu, Estonia, July 2021. IEEE. ISBN 978-1-66544-106-3. doi: 10.1109/ICALT52272.2021.00019. URL `https://ieeexplore.ieee.org/document/9499728/`.

Mary B. Ziskin and Pamela Cross Young. Using assessment data to advance equity: Five things that get in the way. *Intersection: A Journal at the Intersection of Assessment and Learning*, 4(2), May 2023. ISSN 2688-7207. doi: 10.61669/001c.75482. URL `https://aalhe.scholasticahq.com/article/75482-using-assessment-data-to-advance-equity-five-things-that-get-`
00002.