2021

The Upstream Sources Of Bias: Investigating Theory, Design, And Methods Shaping Adaptive Learning Systems

Shamya Chodumada Karumbaiah
University of Pennsylvania

The Upstream Sources Of Bias: Investigating Theory, Design, And Methods Shaping Adaptive Learning Systems

Abstract

Adaptive systems in education need to ensure population validity to meet the needs of all students for an equitable outcome. Recent research highlights how these systems encode societal biases leading to discriminatory behaviors towards specific student subpopulations. However, the focus has mostly been on investigating bias in predictive modeling, particularly its downstream stages like model development and evaluation. My dissertation work hypothesizes that the upstream sources (i.e., theory, design, training data collection method) in the development of adaptive systems also contribute to the bias in these systems, highlighting the need for a nuanced approach to conducting fairness research. By empirically analyzing student data previously collected from various virtual learning environments, I investigate demographic disparities in three cases representative of the aspects that shape technological advancements in education: 1) non-conformance of data to a widely-accepted theoretical model of emotion, 2) differing implications of technology design on student outcomes, and 3) varying effectiveness of methodological improvements in annotated data collection. In doing so, I challenge implicit assumptions of generalizability in theory, design, and methods and provide an evidence-based commentary on future research and design practices in adaptive and artificially intelligent educational systems surrounding how we consider diversity in our investigations.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Education

First Advisor

Ryan S. Baker

Keywords

Adaptive Systems, Algorithmic Fairness, Education, Generalizability, Group Differences, Student Demographics

Subject Categories

Instructional Media Design

THE UPSTREAM SOURCES OF BIAS: INVESTIGATING THEORY, DESIGN, AND METHODS SHAPING ADAPTIVE LEARNING SYSTEMS

Shamya Chodumada Karumbaiah

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Dr. Ryan S. Baker

Associate Professor of Education

Graduate Group Chairperson

Dr. J Matthew Hartley

Professor of Education

Dissertation Committee

Dr. Rand Quinn, Associate Professor of Education

Dr. René Kizilcec, Assistant Professor of Information Science, Cornell University

THE UPSTREAM SOURCES OF BIAS: INVESTIGATING THEORY, DESIGN, AND METHODS SHAPING ADAPTIVE LEARNING SYSTEMS

Dedicated to

my mother, Radha, for having the utmost faith in educating girls,

my father, Karumbaiah, for enduring 16-hour workdays so I could have a dream,

Deepak, for keeping me sane as I wrote this through the pandemic,

and my ancestors - Guru Karana - for leading my way.

ACKNOWLEDGMENT

Study #2: My thanks to my coauthors Ryan S. Baker, Jaclyn Ocumpaugh, and Alexandra Andres. My thanks to Penn Center for Learning Analytics for funding this research. I would also like to thank Dr. Nigel Bosch, Dr. Luc Paquette and to Dr. Anthony Botelho for their early inputs while developing this work. I would further like to express my gratitude to Dr. Anna Fisher, Dr. Anthony Botelho, Dr. Douglas DiStefano, Dr. James Lester, Dr. Jennifer Sabourin, Juan Miguel Andres-Bray, Dr. Karrie E. Godwin, Dr. Ma. Mercedes Rodrigo, Dr. Scott McQuiggan and Dr. Thea Faye Guia for sharing the affect datasets used in this analysis.

Study #3: My thanks to my coauthors Jaclyn Ocumpaugh and Ryan S. Baker. My thanks to the NSF (Cyberlearning Award #1623730) for sponsoring this project, and my thanks to Matthew Labrum and Wanjing-Anya Ma for their support in data preparation.

Study #4: My thanks to my coauthors Andrew Lan, Sachit Nagpal, Ryan S. Baker, Anthony Botelho, and Neil Heffernan. My thanks to the NSF (Division of Information & Intelligent Systems awards IIS-1917545 and IIS-1917713) for sponsoring the project that funded this work, and my thanks to Brian Zylich for his support with the servers.

ABSTRACT

THE UPSTREAM SOURCES OF BIAS: INVESTIGATING THEORY, DESIGN, AND METHODS SHAPING ADAPTIVE LEARNING SYSTEMS

Shamya Chodumada Karumbaiah

Ryan S. Baker

Adaptive systems in education need to ensure population validity to meet the needs of all students for an equitable outcome. Recent research highlights how these systems encode societal biases leading to discriminatory behaviors towards specific student subpopulations. However, the focus has mostly been on investigating bias in predictive modeling, particularly its downstream stages like model development and evaluation. My dissertation work hypothesizes that the upstream sources (i.e., theory, design, training data collection method) in the development of adaptive systems also contribute to the bias in these systems, highlighting the need for a nuanced approach to conducting fairness research. By empirically analyzing student data previously collected from various virtual learning environments, I investigate demographic disparities in three cases representative of the aspects that shape technological advancements in education: 1) non-conformance of data to a widely-accepted theoretical model of emotion, 2) differing implications of technology design on student outcomes, and 3) varying effectiveness of methodological improvements in annotated data collection. In doing so, I challenge implicit assumptions of generalizability in theory, design, and methods and provide an evidence-based commentary on future research and design practices in adaptive and artificially intelligent educational systems surrounding how we consider diversity in our investigations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Adaptive and artificially intelligent systems are being increasingly adopted in education (VanLehn, 2011; Luckin et al., 2016). Formal public education is also proliferated with virtual learning environments (Koedinger et al., 1997). Consider the example of Cognitive Tutor (CT; Ritter et al., 2007), now named Mathia, an intelligent tutoring system that is designed to provide instruction and real-time adaptive feedback to students learning high school mathematics. CT utilizes a data-driven approach to model students' evolving knowledge as they progress in their subject matter learning (Koedinger & Corbett, 2006). Originally designed for the purposes of researching human cognition in the early 1980s (Anderson et al., 1995), CT has found widespread use in schools across the United States. For instance, the Pittsburg Advanced Cognitive Tutor (PACT) center reports that CT was being used in 1400 schools in 2003, and the estimated current usage is over half a million students. Although designed and developed in the United States, CT has also been used in other countries such as Chile (Ogan et al., 2015)). In 2006, CT Algebra was reported to have doubled students' end-of-course problem-solving test scores (Koedinger & Corbett, 2006). However, a 2014 study concluded that CT had a mixed effect on high school students' mathematics achievement - complex positive effects for CT Algebra but negative effects for CT Geometry (Pane et al., 2014). Like CT, several other tutoring systems such as ALEKS (~600K students), Inq-ITS (~100K students), MathGarden, and Alef NexGen are being used widely in classrooms. As these systems continue to be used by diverse

student populations across the United States and the globe, they need to ensure that their pedagogical decisions meet the needs of all students for an equitable outcome.

Recent research highlights how adaptive systems encode societal biases leading to discriminatory behaviors towards specific subpopulations (Baker & Hawn, 2021; Kizilcec & Lee, 2020). This is in direct contradiction to the popular view of data-driven decision-making as fair and objective (Lee, 2018). Several recent studies across varied domains (e.g., criminal justice, medicine, education) have demonstrated bias in systems using computer-driven algorithms to make different decisions for different individuals. This includes high-stakes decisions such as predicting recidivism (Angwin et al., 2016), administering anesthesia (O'Reilly-Shah et al., 2020), and hiring (Garcia, 2016). Adaptive systems in education incorporate different kinds of algorithmic decision-making - data-driven or rule-based, explainable or inscrutable, descriptive or predictive. For instance, intelligent tutoring systems like Learnta use an algorithmic model of student knowledge to predict a student's current skill level and decide the difficulty level of the content they will see next (Baker et al., 2020). Bias in these systems can lead to inequitable student outcomes, causing harm to specific student subpopulations (Doroudi & Brunskill, 2019; Holstein & Doroudi, 2021). For instance, if a tutoring system is biased against female students, it may systematically deliver content below the students' skill level, leading to missed learning opportunities and suboptimal experience - potentially lowering achievement and engagement in female students in the long run. Several recent research projects have examined bias in algorithmic systems in education (see review in Baker & Hawn, 2021). However, the focus has mostly been on investigating bias in predictive

modeling, particularly its downstream stages like model development and evaluation (Kizilcec & Lee, 2020). By considering three upstream components separately (e.g., theory, design, training data collection method), this dissertation aims to bring to attention those aspects of the design and development of adaptive systems that are often obscured in the evaluation of bias in downstream products (e.g., an algorithmic model in an adaptive system). As elaborated in a later section, studying upstream bias is important not only because such bias could manifest itself in specific ways downstream (e.g., suboptimal choice of variables in a predictive model based on a biased theory), but also because it could lead to direct discriminatory behaviors in an adaptive system (e.g., biased affective intervention).

My dissertation work hypothesizes that the upstream components (e.g., theory, design, training data collection method) in the development of adaptive systems also contribute to the bias in these systems, highlighting the need for a nuanced approach in conducting bias research. Using previously collected student data from various virtual learning environments, I conduct empirical analyses of demographic disparities in three cases that are representative of the aspects that shape technological advancements in adaptive systems in education: 1) non-conformance of data to a widely-accepted theoretical model, 2) differing implications of technology design on student outcomes, and 3) varying effectiveness of methodological improvements in annotated data collection. In the remainder of this chapter, I present a detailed overview of the overarching theme of the dissertation research - identification of the upstream sources of bias. First, I will briefly introduce bias pertaining to adaptive systems in education. Then, I provide a detailed

theoretical foundation for bias, the harms caused by biased systems, and the current discourse on the origins of bias. After summarizing the empirical evidence of bias reported in educational algorithmic systems so far, I argue for the need to move upstream to identify other sources of bias. I conclude with a discussion on the potential origins of upstream biases. Finally, I discuss the purpose of the three studies conducted in this dissertation, connecting them back to gaps identified in the literature and how they each contribute to the identification of upstream biases.

1.1 Bias in Adaptive Learning Systems

In this dissertation, I will focus only on fully automated adaptive systems designed to provide personalized instruction and immediate feedback to students (Corbett et al., 1997; Shute & Psotka, 1994). The general goal of these systems is to optimize student learning with one-to-one instruction which is often hard to achieve in traditional schooling. Several forms of adaptive systems exist today - from fully online (e.g., Database Place) to blended learning environments (e.g., ASSISTments), from multiple-choice and fill-in-the-blank problem answering (e.g., Imagine Learning's Reasoning Mind) to learning games (e.g., Zoombinis) to open-ended learning (e.g., MiGen), and from sketching-based (e.g., Physics Playground) to dialogue-based tutors (e.g., AutoTutor). They have been built on several platforms, including standalone, web, or mobile applications, virtual reality systems, and wearable technology. Likewise, they have been designed for diverse subject matters such as mathematics (e.g., Cognitive Tutor Algebra), biology (e.g., MetaTutor), literacy(e.g., Writing Pal), and vocational training (e.g., TransfrVR). I chose to investigate these systems for the following reasons. First, they often involve models of complex educational

constructs (e.g., models of student affect, cognition, and self-regulated learning behaviors). Second, they utilize rich student data (e.g., fine-grained interaction data) to make real-time decisions that impact students' experiences closely. Third, the two complexities above may likely introduce complex biases which may need a more nuanced approach to identify and mitigate. Lastly, despite the wide usage of these systems, biases in these systems have not yet been studied thoroughly.

The distinguishing feature of these adaptive systems compared to traditional computer-aided instruction is their ability to adapt based on the perceived student needs (Self, 1999). Although originally designed to adapt instruction by inferring students' knowledge (Anderson et al., 1995), these learning environments soon started adapting their pedagogical actions based on other aspects of the student experience such as behavior, affect, and motivation (Desmarais and Baker, 2012). In some cases, the system behavior is driven by a set of rules - often derived from small-scale experimental studies (e.g., when and how to provide hints; Aleven et al., 2016). In other cases where the educational construct is more complex (e.g., students' changing subject matter knowledge), these systems need an accurate computational model of the construct to adapt meaningfully (e.g., choosing the next appropriate math problem based on a student's current skill level). Such models are often data-driven, i.e., they are built to understand the patterns in the past data to make future decisions. Nevertheless, in most cases, the result is automated decision-making with little to no human intervention (there are some exceptions, such as Learnta and SQL-Tutor, which leaves the final decision for teachers or students to make).

One of the central assumptions in such rule-driven or data-driven algorithmic systems is that their decisions (e.g., who gets a hint) are inherently objective and fair (Lee, 2018). However, several recent studies have reported bias in automated decision-making - both in education (Yudelson et al., 2014; Kai et al., 2017, Ocumpaugh et al., 2014) and other domains more broadly, including medicine (O'Reilly-Shah et al., 2020), hiring (Garcia, 2016), and criminal justice (Angwin et al., 2016). In addition, algorithmic systems in education often serve a diverse student population, with a single system catering to students from different grade levels (e.g., Alef NexGen serves students from elementary through high school), countries (e.g., Cognitive Tutor used in the United States and Chile), and socioeconomic and cultural backgrounds. This further complicates the issue of bias as system behavior now needs to be non-discriminatory across all the student populations it serves. Bias in adaptive educational systems is a serious issue as it has the potential to harm students (O'Neil, 2016). Consider a system that adapts the learning content and student interaction based on the student's affect. Now, assume that the identification of student affect is biased against female students. For instance, perhaps it consistently misidentifies boredom in female students and intervenes less effectively than for male students. Such a bias will likely result in poorer educational experiences for female students and potentially also lead to loss of motivation and interest in the subject matter. Worst, this could also become a self-fulfilling prophecy wherein the incorrect affective interventions for female students may lead to actual boredom, further confirming the system's bias, thus perpetuating the system's behavior. In this fashion, algorithmic systems have the potential to obscure the root cause of the bias and automate its effect - making it harder to objectively

identify and fix it. Hence, identifying and mitigating bias in adaptive educational systems has become an important area of research in recent years (Baker & Hawn, 2021; Holstein et al., 2019; Holstein & Doroudi, 2021; Kizilcec & Lee, 2020).

Empirical evidence on bias in adaptive systems. Studying bias in adaptive educational systems is not a new topic. Among the empirical studies of (non-algorithmic) bias, it has been more common to study the biased effectiveness of the adaptive system as a whole as compared to identifying the sources of the bias. A few studies have examined the demographic differences in the effectiveness of adaptive system design (e.g., Arroyo et al., 2013; Finkelstein et al., 2013; Ogan et al., 2015). However, other upstream sources such as the theories driving these design choices or the data collection methods used for decision-making have not been empirically studied in detail. Here are some relevant past research that examined differences in key educational constructs (e.g., learning, affect, help-seeking, and off-task behaviors) in adaptive systems based on students' demographic categories such as gender, disabilities, nationality, and other sociocultural factors.

First, there is some evidence that the differences in the design of adaptive systems may explain some of the demographic differences observed in educational outcomes. Arroyo and colleagues (2013) conducted a series of four studies over ten years to examine the gender differences in the effectiveness of pedagogical agents and hints in a virtual math tutoring system. These studies suggested that using female characters as pedagogical agents was beneficial for female students, while male students exhibited worse outcomes in the presence of female characters - in terms of both learning and affective experience. Another randomized control study with the same tutor reported that the affective feedback

delivered by the pedagogical agents was more effective for low achievers (including students with learning disabilities) as compared to high achievers (Woolf et al., 2010).

Second, cultural mismatches in student behaviors and classroom practices have been reported in adaptive system studies as possible explanations for demographic differences. When studying the relationship between help-seeking and learning in three different countries (Costa Rica, the Philippines, and the United States), Ogan and colleagues (2015) observed differences in classroom practices that may contradict assumptions about effective help-seeking, as measured through student behaviors logged by an online tutor. For instance, students in Costa Rica displayed collaboration outside the tutor leading to help usage. This study highlights the need to consider differences in cultural values and practices when an adaptive system is used in a country that is different from where it was designed, developed, and tested (Blanchard, 2012). Another study (Ogan et al., 2012) conducted in three Latin American countries (Brazil, Mexico, and Costa Rica) also found intensive and ongoing collaborative use of an adaptive system (Cognitive Tutor) that was developed in the United States for individual use. Students also walked away from the computer from time to time, potentially leading to significant missing data in the interaction log. In another study comparing differences in nationalities, Rodrigo and colleagues (2013) conducted a set of three studies to investigate the behavioral differences in the use of an intelligent tutoring system based on students' nationality. They reported consistently less off-task behavior in students in the Philippines as compared to the students in the United States after controlling for curriculum and study methods.

Third, a few articles (mostly non-empirical) have suggested how research methods may lead to biases in adaptive systems. Researchers in the Philippines have shared the unique challenges surrounding infrastructure, logistics, institutional support, and student communication when conducting educational technology research in their context (Andres et al., 2015). They highlight the disconnect in implementing methods developed in the western context to a low-income global south country. More broadly, Blanchard (2015) analyzed the international representation in one of the primary research communities studying adaptive systems in education, i.e., Artificial Intelligence in Education (AIED). The analysis suggested "limited cultural diversity" (p. 225) in the student population investigated in AIED research and differing needs of students from non-WEIRD (Western Educated Industrialized Rich and Democratic) contexts.

Finally, some studies have also focused on overcoming bias in adaptive systems by adapting their design to students' needs. For instance, Finkelstein and colleagues (2013) observed significantly better performance of third graders in science when the adaptive system used a similar dialect as the native tongue of the students (African American Vernacular English; AAVE). Another promising new research has also attempted to adapt the content of a math learning system (ActiveMath) based on the students' cultural context (Melis et al., 2019). It is yet to be tested more broadly. As discussed so far, the past empirical studies on non-algorithmic bias in adaptive systems have viewed bias as a design issue without much deliberation on other upstream sources such as methods or theory (except a few conceptual papers). In the next section, I present the more recent advances

in the study of bias in adaptive systems and its origins, which also seem to be lacking in efforts to move upstream to identify and mitigate bias.

1.2 Theoretical Foundation: Bias and its Origins

In this section, I will first present some background literature on bias and a brief overview of studies that have investigated bias in algorithmic systems in education. I will discuss the current understanding of the origin of bias to then (in the next section) present arguments for why its upstream origins also need to be considered.

What is bias? Despite the recent spike in the research on algorithmic bias, there is no consensus on a single definition of bias (Crawford, 2017). Furthermore, the often overlapping conceptualizations of bias and the harms they cause have themselves been considered a potential limitation that needs to be addressed as the work to mitigate it emerges (Blodgett et al., 2020). Lack of clarity in the description of bias has been argued to hinder the progress in conversations about "what kinds of system behaviors are harmful, in what ways, to whom, and why" (Blodgett et al., 2020, p. 2). Here I present some definitions of bias in the current literature. A popular use of the term bias is to refer to the difference in the accuracy of a computational model (or a lack of generalizability thereof) for different social groups - in some definitions, this explicitly relates to disadvantaged groups (Mitchell et al., 2021) but not in others (Gardner et al., 2019). Some researchers prefer to use alternate terms such as unfairness (Mehrabi et al., 2019), discrimination (Friedman & Nissenbaum, 1996), or demographic disparity (Barcoas et al., 2019) while restricting the technical use of the term bias to indicate a systematic statistical error

(Barocas et al., 2019). It is relatively less popular to view bias from an individual point of view as compared to the group definitions mentioned so far (Dwork et al., 2012). However, a small subset of research has investigated ways to attain individual fairness, wherein "people who are 'similar' with respect to the [prediction] task receive similar outcomes" (Binns, 2020, p. 1). In this dissertation, I investigate biases originating at the group level and define biases as the possible "sources of downstream harm" - in the same sense as used by Suresh and Guttag (2020, p. 1). However, instead of restricting the sources of harm to the stages of a machine learning pipeline, as Suresh and Guttag (2020) do, I expand the scope of inquiry to broader aspects of the adaptive system that can similarly lead to "societally unfavorable outcome[s]" (p. 2).

What harms do biased systems cause? Downstream harm in the case of adaptive systems I study in this dissertation refers specifically to inequitable student outcomes - not just restricted to learning outcomes but also their experience more broadly, including motivation, self-identity, and affect. To understand how biases can lead to such harms, it may help to examine the consequences of bias through the lens of allocative and representational harms (Crawford, 2017). Consider the previous example of a tutoring system biased against female students in its knowledge estimation. In this case, allocative harm for this student group refers to the missed learning opportunity (content offered below the actual skill level). A more well-known example of allocative harm in education is the bias in standardized testing leading to denied college admissions (Dorans, 2010; Santelices & Wilson, 2010). Similarly, bias against certain linguistic and ethnic minorities in automated essay scoring may lead to lower scores for those groups in high-stakes exams

like the GRE (Madnani et al., 2017). On the other hand, representational harm refers to discriminatory depictions of certain subgroups. For instance, African American English is sometimes tagged as hate speech in discussion forum posts (Sap et al., 2019). Representational harm can also be more subtle sometimes. Historical bias may lead to certain groups being advantaged over the other - making the representation of those groups more favorable than others. If we then use demographics as predictor variables in a model, the model is likely to make decisions based on the group membership of the student - leading to such biases being perpetuated further. For example, students belonging to a certain subgroup may be systematically predicted to have lower grades (Wolff et al., 2013). Including demographic variables while modeling is not uncommon in education. In the recent review of papers in educational data mining, Paquette and colleagues (2020) found that around half of the papers use demographic features in their models.

Demographic categories studied. As synthesized by Baker and Hawn (2021), the majority of the empirical studies on bias in educational algorithmic systems have focused primarily on the more frequently measured categories of race, gender, and nationality. They attribute this trend partly to the perceived importance of these demographic categories and highlight the need to also consider other categories studied in a fewer number of studies like native language and dialect, disabilities, urbanicity, parental educational background, socioeconomic status, international students, and military-connected status. To identify which categories need to be considered in the investigation of bias, Baker and Hawn (2021) propose a "non-malleability test" - similar to the "immutability" consideration of

identifying protected classes (Soundarajan & Clausen, 2018) - which examines the "the degree to which an individual can move in and out of the suggested class."

In this dissertation, I will be investigating bias with respect to nationality in one study, urbanicity in two studies and race/ethnicity, economic status, English as a second language, special education, and charter school in another study. The choice of urbanicity as a demographic variable is driven by reasons similar to those outlined by Ocumpaugh and colleagues (2014). This includes 1) racial, ethnic, and economic differences in student populations across urbanicities (e.g., affluent White students in suburban schools as compared to non-White students with poorer economic backgrounds in urban schools), 2) reported correlations between educational outcomes and urbanicity (Graham and Provost, 2012), 3) the use of urbanicity as a demographic in education research more broadly (Campbell, 1989; Hu, 2003), and 4) common exclusion of rural or urban schools in research studies due to practical constraints around budget, recruitment, accessibility, and time. Also, due to limited data availability, I currently use a broad categorization for race/ethnicity (e.g., Asian), which has been criticized for oversimplifying these groups (e.g., combining several distinct communities such as Sri Lankan, Korean, and Vietnamese) primarily on political lines rather than cultural identities (Strmic-Pawl et al., 2018). This is a limitation.

Empirical studies on algorithmic bias. Here, I provide a brief summary of the empirical evidence from recent studies on bias in predictive models in education. As noted earlier, most of this research has focused on race, gender, and nationality. Due to limited research on other student subpopulations, I will summarize results only for these traditionally

measured categories and recommend readers to refer to the review by Baker and Hawn (2021) for a more comprehensive list. Kai and colleagues (2017) studied models predicting student retention in an online college program and found that some algorithms were more equitable than others when gender and race (African American and White) differences were examined. Hu and Rangwala (2020) also examined race and gender differences in at-risk prediction (for failing a course) and found that the models did worse for African American students and male students. Anderson and colleagues (2019) studied differences in college graduation prediction and observed higher false-positive rates for White students and higher false-negative rates for both Latino and male students. Lee and Kizilcec (2020) developed an equity-corrected course grade prediction algorithm and found that this version did better with underrepresented racial/ethnic groups and with male students as compared to the unmodified algorithm. Yu and colleagues (2020) found that including student demographics as a predictor variable in a success prediction model for undergraduates led to inaccurate predictions for female students and for some racial groups. Bridgeman and colleagues (2009, 2012) studied automated essay scoring algorithms and found that they were more accurate for male students and assigned higher scores for Chinese and Korean students. Finally, Ogan and colleagues (2015) studied cross-country validity (Costa Rica, Philippines, and United States) of student learning models and found that the models built using data from the same country were more accurate than the data from the other countries. Of these, only the study conducted by Ogan and colleagues investigated bias in a model used in adaptive learning systems.

Overall, there is evidence that it is likely problematic to assume population validity in algorithmic systems in education. More work is needed to understand student subpopulation differences further - both within the core categories (of race, gender, and nationality) and more broadly. First, there are no studies on non-binary gender identities or any other categories of LGBTQ identities (Baker & Hawn, 2021). Second, only one study examined data on indigenous students (Anderson et al., 2019) and reported that the model was highly unstable for indigenous students who were significantly underrepresented in the data. Third, the research so far has emerged primarily from research groups in the United States which indicates that there is a serious lack of voice from other contexts in this discussion. Fourth, some of the categorizations may be oversimplified or politically influenced (e.g., with race as discussed earlier). Finally, as noted earlier, more work is needed on categories other than race, gender, and nationality that are potentially important in understanding students' experience and learning.

The current discourse on the origins of bias. Unlike statistical bias originating from a systematic measurement error, "societal bias" - which is the focus of this dissertation - is often attributed to discriminatory social structures that get embedded in the data (Mitchell et al., 2021). The examination of the origin of societal bias has been focused primarily on the different stages of the machine learning pipeline - the process beginning at data collection and continuing to model development and application (e.g., Suresh and Guttag, 2020). The key early work on identifying the origins of algorithmic bias is that of Friedman and Nissenbaum (1996), who mapped the biases to the context in which they arise, resulting in a corresponding sequential map of the origins - from historical bias to the ones

introduced during data collection and later by the use in the real world. Such a sequential framing has been instrumental in the present-day conceptualizations of the machine learning process related origin of bias - both in finer-grained (Suresh and Guttag, 2020) and broader framings (Barocas et al., 2019; Kizilcec and Lee, 2020; Mehrabi et al., 2019). Below I present one of the comprehensive frameworks on the sources of bias.

Figure 1.1 presents the framework proposed by Suresh and Guttag (2020), connecting the sources of bias to different stages of the machine learning pipeline. The first three sources of bias are related to the data collection stage. First, historical bias originates from data that represents the world as-is, including the biases that exist historically. For example, historical biases in college admission decisions (Rosser, 1987; Clark et al., 2009) may be replicated by an algorithmic system used to inform future admission decisions. Second, underrepresentation of certain student subpopulations in the data causes representation bias, leading to lower predictive performance for that group. For example, recent research has reported less accurate college graduation predictions for minority groups in higher education, such as indigenous students (Anderson et al., 2019). Third, measurement bias occurs from variables lacking construct validity - either entirely or for specific subpopulations. For example, African American English has been reported to be annotated incorrectly as hate speech in discussion forum posts (Sap et al., 2019).

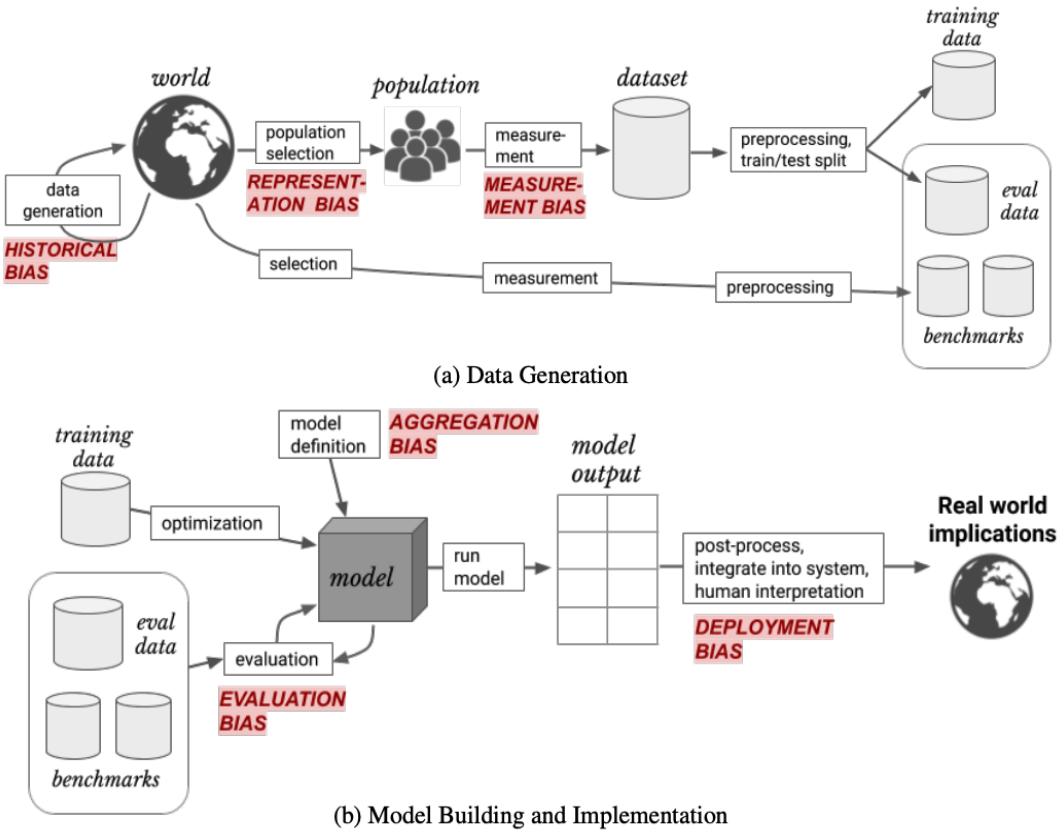


Figure 1.1 The framework on the sources of bias proposed by Suresh and Guttag (2020).

The next three sources of bias are related to model development and its use in the real world. First, aggregation bias results from combining data from different student subpopulations who are to be ideally considered separately due to inherent differences in them. For example, lower affect detection performance has been reported when rural, urban, and suburban students are combined together as compared to the individual models trained separately (Ocumpaugh et al., 2014). The challenge, however, is that it is often difficult to identify specific student populations (Baker et al., 2019), let alone those populations that need to be considered separately and those that could be aggregated safely. Second, evaluation bias occurs due to the mismatch in the population used to assess the

model and the actual target population. It is usually difficult to identify evaluation bias as it is not a common practice to report the population on which the models were assessed (there is also a problematic lack of reporting population information of training data; see discussion in Paquette et al., 2020). Third, deployment bias results from using models in ways they were not originally designed for. For example, at-risk predictions may be used automatically to customize future course selection, while it was designed to be used just as a heuristic by the course instructor (potentially along with other information) to identify students who may need personal attention.

1.3 Upstream Sources of Bias

As presented so far, the empirical studies and the overall discussion on the sources of bias in adaptive educational systems have focused mainly on the development and evaluation of the algorithmic (mostly predictive) models in them (Baker & Hawn, 2021). The investigations in this dissertation are based on the central hypothesis that there are other important upstream factors that contribute to the biased decision-making in adaptive systems. The current emphasis on the research in the downstream locations of predictive modeling - while being an important step in the right direction - will limit the progress towards fairness and equity in adaptive educational systems if we don't broaden our search. By considering three upstream components separately (e.g., theory, design, training data collection method), this dissertation aims to bring to attention those aspects of the design and development of adaptive systems that are often obscured in the evaluation of bias in downstream products (e.g., an algorithmic model in an adaptive system). Studying upstream bias is important not only because such bias could manifest itself in specific

downstream components (e.g., suboptimal choice of variables in a predictive model based on a biased theory), but also because it could lead to direct discriminatory behaviors in an adaptive system (e.g., biased affective intervention). Hence, in this dissertation, I investigate three upstream components for potential sources of bias - the theories shaping the adaptive systems, design choices of these systems, and data collection methods adopted. Next, I present some reasons for choosing these three upstream factors for the three studies.

The Influence of Theory, Design, and Methods on Adaptive Decisions. There are several ways in which theory, design, and data collection methods are known to influence decision-making in adaptive systems. Moreover, theories may also impact design and data collection. And design may also impact data collection. For instance, an upstream bias in theories can manifest itself in a biased design (e.g., culturally irrelevant rewards as extrinsic motivators) or a biased data collection method (e.g., choice of culturally irrelevant affective states to collect data for). It can also influence the decision-making of an adaptive system more directly (e.g., theory-driven rules of affective interventions that are culturally insensitive) or indirectly induce bias through algorithmic models used in them (e.g., influencing a biased choice of predictors for boredom detection). In the latter cases, upstream biases (e.g., theoretical bias) could be a root cause of the downstream biases introduced at different stages of algorithmic modeling. With examples from the design and development of virtual learning environments, I illustrate the different roles that these three facets play.

The development of adaptive and artificially intelligent systems in education has accelerated as the amounts of learning-related digital data generated by these systems increased rapidly (Siemens & Baker, 2012) - both in terms of the number of students and the fine-grained data collected per student. With the exception of some earlier cases of algorithmic interventions in education which were designed based on specific theories (e.g., Cognitive Tutors were built on John Anderson's ACT-R theory of human cognition), the algorithmic design has mostly been data-driven, involving models that automatically learn patterns (that may or may not be interpretable) from large amounts of historical data. However, recognizing the important (often implicit) role theories play, there has been a recent call to pay more attention to theories when dealing with educational data (see special issue by Wise & Shaffer, 2015). I now present a few ways in which theories continue to shape several aspects of adaptive systems.

First, theories drive the assumptions around the conceptualization of educational constructs. Consider the highly-cited theoretical model of affect dynamics proposed by D'Mello and Graessar (2012) that hypothesized the likely transitions between different affective states prominent during a student's deep learning of a subject matter. For example, this model states that students are more likely to transition from engaged concentration to confusion when faced with an impasse, while it is less likely that they transition from engaged concentration directly to boredom. This theoretical model has been cited to justify the selection of two multi-state transition patterns between four affective states - starting from engaged concentration to confusion to frustration to boredom and vice versa - when

conducting studies of affect sequences in digital learning environments (Andres et al., 2019; Ocumpaugh et al., 2020).

Second, theory influences the design of variables used in predictive modeling. Consider an algorithmic model of carelessness (e.g., Hershkovitz et al., 2013) that includes affective states as the predictor variables in it. The theoretical model from the example above will likely shape the choice of representing a student's affective experience in relation to carelessness with the same four states or sequences, even though the original study may not have focused on carelessness. As Wise and Shaffer (2015) explain, theory plays a key role in directing our attention to the variables that can be assumed to be more important. This is especially relevant in the current era of big data, when fine-grained digital logs provide many options to choose from. However, it is very likely that a widely accepted theory will be used in contexts for which it was not originally developed - both in terms of the construct being modeled and the student subpopulations it will be used for. As theory closely shapes the assumptions surrounding modeling, theoretical biases could become the root cause of the downstream biases introduced at different stages of modeling.

Third, theory influences the interpretation of student behaviors in the data collected and also the design of educational interventions in the adaptive systems. For example, our basic understanding of how to assess students' affective experiences is highly influenced by theories like Csikszentmihalyi's "zone of flow" (Csikszentmihalyi, 1990). Several emotion theories assume that there is positive and negative affect, which then determines how we interpret the data collected on the student experience (e.g., boredom as bad and flow as good) (Csikszentmihalyi, 1990; Kort et al., 2001). Moreover, the design of affective

interventions in an adaptive system may also drive its decisions based on these assumptions. There are examples of automated interventions that try to get students into the positive affective states and avoid the negative affective states (e.g., Padron-Rivera et al., 2018). In contrast to this popular assumption, a recent review of the empirical studies on the relationship between students' affect and their learning outcomes suggests that most affective states have mixed relationships with learning (Karumbaiah et al., 2021). Moreover, whether an affective state is conducive for learning or not is also context-dependent. For example, studies conducted in authentic settings like classrooms had significantly more negative relationships between confusion and learning outcomes as compared to laboratory studies (Karumbaiah et al., 2021).

Design influences the quality of student experience and learning directly and also shapes what data gets collected and how. Consider the design of hints in virtual learning environments. Arroyo et al. (2000) found that the effectiveness of different hint designs varied by gender. Specifically, girls benefited more from highly interactive hints, while boys did better with less interactive hints. This work matches findings in other learning contexts, which has shown both that there may be racial and gendered interactions influencing differences in help-seeking behaviors and that these different behaviors may explain subsequent achievement patterns (Ryan et al., 2009). In addition to impacting student experience more directly, design may also influence data collection. In the previous example, hint usage in a virtual learning environment is often used as a proxy for students' help-seeking tendencies (e.g., Aleven et al., 2016). Hint design can influence students' perception of help-seeking opportunities as affecting their sense of competence and

autonomy. Based on how hints are designed in a system (automatic versus on-demand, ease of access to bottom out hints, goal feedback versus content feedback), the measurements of students' help-seeking behaviors may vary in meaning.

Data collection methods closely shape the measurement of the educational constructs. Going back to affect data collection, let's compare the three methods employed in past studies - field observations, self-reports, and automated detection. First, field observations by expert coders will likely generate coarser-grained affect data for each student (e.g., one observation every three minutes). In addition, cultural differences between the annotators and students have been reported to interfere with the quality of affect annotations (Okur et al., 2018). Second, student self-reports may suffer from reliability issues (Richardson, 2004), especially if students perceive certain affective states as being positive and others as negative or if the teacher is involved in the data collection efforts. In some studies, students are asked to report their affect retrospectively. In this case, reliability issues may arise from students' ability to recall discrete emotions they experienced during their learning when asked to reflect on them after the completion of the learning activity. Recent work suggests that there is considerable variation in students' retrospective memory of their own affect, even shortly thereafter (Rebolledo-Mendez et al., 2021). Finally, using an automated detector of affect enables the collection of finer-grained data for a much larger sample. Obtaining a large sample is not feasible with the other methods, all of which require considerable human labor. However, general-purpose automated detectors may not generalize well to diverse student populations. If the detector involves facial expression

analysis, there are also concerns about the poor performance of facial recognition for female students and students with darker skin tones (Lohr, 2018).

As theories, design, and methods continue to shape future research and innovation in adaptive systems in education, solely fixing downstream predictive models for biases will not be sufficient. Upstream components need to be checked to make sure that the bias embedded in them doesn't form a basis for bias in downstream stages of algorithmic systems. Letting an algorithmic model perpetuate an upstream bias in its discriminatory decision-making may further harm the sensitive group, leading to outcomes that falsely confirm the bias. For example, intervening incorrectly on boredom for a certain student population based on a biased theory, design, or affect data collection may lead to an actual increase in boredom in that population due to ineffective interruptions. Identifying and mitigating upstream biases is an important step towards preventing such feedback loops that may amplify the harm. In this dissertation, I explore upstream biases that manifest directly as differing rates of theoretical conformance, the validity of design assumptions, and methodological effectiveness in the empirical evidence for different student subpopulations. However, as discussed above, it is possible that the biases in theories could also have an indirect effect, wherein the theoretical bias could be the root cause of the downstream biases observed in different stages of modeling (see Figure 1.1). This is likely to be the result of the unacknowledged influences that theories have on several aspects of data and modeling practices as described above (e.g., choice of variables, measurement, interpretation of results). Thus, the identification of theoretical bias will likely be an

arduous process involving researchers and designers questioning several of their assumptions driven by implicit theoretical lenses.

Origins of Upstream Bias. Biases in upstream sources like theory, design, and data collection methods could originate from some of the similar mechanisms described earlier. Theories and design choices are often informed by experiments that could be affected by historical, representation, and measurement biases (terms borrowed from Suresh and Guttag, 2020). First, forming theories and design choices by observing the world as it exists (as opposed to as it should be) may lead to a theory that reflects the historical biases in the world. Consider educational interventions aimed at improving achievement gaps in students in the United States (e.g., writing exercises to promote self-affirmation in Borman et al., 2016). It can be argued that this framing of achievement gaps is biased and ignores the historical, economic, sociopolitical, and moral debt owed to Black, Latina/o, and Native American students whose identities were treated as barriers in the educational system (Ladson-Billings, 2006). The debt, Ladson-Billings argues, is not about what we might mitigate, but what is owed after years of undeserving them. Such historical biases, if embedded in theory, design, or methods, get perpetuated by downstream applications, such as the design of an adaptive system or a predictive model that then automates the biased decision-making at a potentially larger scale (e.g., biased college admissions negatively impacting groups already underrepresented in higher education).

Next, representation bias can invalidate the assumptions in a theory, design choice, or data collection method, especially for student subpopulations not represented in the experiments informing these assumptions. This is not uncommon - in fact, it is likely for small-scale

experiments to not contain students from all the target subpopulations. Representation bias occurs due to several reasons. First, most educational research studies tend to be conducted in western countries with adaptive systems developed by designers in the west (Blanchard, 2012). The study results and the designed adaptive systems then get used in non-western contexts (especially the global south) - oftentimes with little to no research on the generalizability of the design to the new population, despite evidence that the systems are used by students in very different ways in those countries (Ogan et al., 2012). Consider the running example of affect detection. An affect-aware tutoring system built for western students may not identify or respond optimally to affective experiences of students in a country like Sri Lanka. Culture is known to influence variation in beliefs and personal dispositions towards emotional expression and moderation (Tsai & Levenson, 1997; Uchida et al., 2009) and the frequency and emergence of certain affective states (Kitayama et al., 2000). It is not too long ago that the universality of the basic emotions (e.g., Ekman's six basic emotions of anger, fear, disgust, sadness, joy, and surprise were assumed to be innate and cross-cultural; Ekman, 1971) was demonstrated to be invalid for its lack of consideration of sociocultural factors influencing affective experiences (Elfenbein, 2002).

Second, even within the western context, due to practical constraints of research projects (e.g., budget, recruitment, accessibility, and time), some small-scale experiments tend to recruit their participants from a convenience sample - like undergraduate, middle-class students in the United States (see Kimble, 1987 for discussion) - who are likely to exhibit significant differences in their behavior than other subpopulations (Henrich et al., 2010). In the case of affect, for example, age is known to influence people's emotional

expressivity (Dunn & Brown, 1994; Gross et al., 1997) and inhibition (Cole, 1986). The inferences about students' affective experiences from an experiment with undergraduate students are likely not to hold with younger students in elementary, middle, or high school for whom several of the adaptive systems in education are developed. With education data, there are other added logistical constraints such as the cost of reaching remote rural schools or difficulty with bureaucratic overheads that make it harder to collect data in some schools than others (Baker et al., 2019).

Third, even when there is access to larger, more diverse datasets - especially digital log data from virtual learning environments - it is often harder to collect student demographics data due to concerns over student privacy. This further limits investigation on demographic differences in the theoretical and design assumptions made in the adaptive systems. Finally, there is also a lack of representation among the designers of these adaptive systems within national boundaries. Institutionalized and unconscious bias and social and cultural distance between educational technology designers and those they seek to serve (especially low-income and minority groups) are the two common sources of failure for the equitable deployment of new technologies (Reich & Ito, 2017). Technology developers' lack of awareness of sociocultural contexts and the needs of different student subgroups can lead to unfortunate consequences. Reich and Ito (2017) emphasize that measuring differences in how various subgroups experience and benefit from adaptive systems in education will be a crucial component of our deepening understanding of these technologies and in addressing the inequalities that emerge.

Lastly, measurement bias can occur due to issues in data collection methods. Measuring complex and sometimes subjective educational constructs need caution to ensure that the measurements are reliable for different student subpopulations. Here are some ways in which measurement bias gets introduced. Going back to the example of student affect, coder bias due to cross-cultural affect coding (observers of one culture coding affect of students from another culture) may systematically introduce errors in affect measurements for students whose culture doesn't match with the coders (Okur et al., 2018). This is also true for other forms of affect measurements, such as automated detection through physiological signals like facial expression. For example, automated facial recognition has been reported to perform poorly for individuals with darker skin tones and for females (Lohr, 2018). This technological limitation may introduce systematic biases in the measurement of affect for female students of color. Previously, I discussed the role of student demographics in affective experiences. A general-purpose affect detector (e.g., McDuff et al., 2016) trained using data from a certain context (e.g., adults in non-learning related situations) is likely to produce less accurate measurements of affect in adaptive educational systems if used as-is. More generally, there could be other reliability issues that may differ between student subpopulations. In the case of self-reports of affective states, students of some ages or cultures may hesitate more or less to report negative affect in fear of consequences. In this dissertation, I will be focusing on illustrating the existence of bias in theories, design choices, and data collection methods that shape adaptive systems in education. Locating the exact origin of these upstream biases is an arduous but important task that is out of scope for this dissertation work.

1.4 Purpose of the Studies

In this section, I briefly present the three studies conducted as part of this dissertation research connecting each of them to the broader purpose. I briefly discuss how each study contributes to the gap identified in the previous sections. The discussion sections of the following three chapters present a more detailed implication to research and practice.

Study #1. Non-conformance of data to a widely-accepted theoretical model of emotion. The first upstream component investigated in this dissertation is a theory, specifically a popular theoretical model of affect dynamics. Affect dynamics is a field of research that studies how students' affect develops and manifests over time (Kuppens, 2015). The most commonly-cited model in this context, put forward by D'Mello and Graesser (2012), postulates that a specific set of affect transitions will be particularly prominent. It has been one of the most influential theoretical frameworks in affect dynamics research. However, few empirical studies have matched that model's predictions (see review in Karumbaiah et al., 2018), an issue which this study investigates. Further investigation of the literature reveals that at least some of the differences in the literature may be culturally driven. The studies that do show some evidence for the model were all conducted in the United States with undergraduate populations, but other student populations seem to show more variance in their transition patterns.

To better understand the validity of this method and its scope of applicability, I reanalyzed and synthesized previously collected data from diverse learning contexts. I first address some methodological concerns that had not been previously discussed in the literature,

presenting how various edge cases should be treated. Next, I present mathematical evidence that several past studies applied the transition metric incorrectly - leading to invalid conclusions of statistical significance - and provide a corrected method. Using this corrected analysis method, I reanalyze ten past affect datasets collected in diverse contexts and synthesize the results, determining that the findings do not match the most popular theoretical model of affect dynamics. More importantly, the results in this study suggest that affective patterns seem to differ based on the country in which the research was conducted (US versus Philippines). The results limit the scope of applicability of the theoretical model and highlight the need to focus on cultural factors in future affect dynamics research.

Study #2. Differing implications of technology design on student outcomes. The next upstream component investigated in this dissertation is the technological design of an adaptive system. Adaptive system designers need to ensure population validity as they attempt to meet the individual needs of all students. There is often access to larger and more diverse samples of student data to test replication of design assumptions across broad demographic contexts in adaptive systems as compared to either the small-scale experiments or the larger convenience samples often seen in experimental psychology studies of learning (see discussion in Kimble, 1987). However, the source of typical educational data in these systems (i.e., interaction log data) and concerns related to student privacy often limit the opportunity to collect demographic variables from individual students – the sample is diverse, but the researcher does not know how that diversity is realized in individual learners. Yet considerable research shows that demographic factors

are often related to differences in educational outcomes (Childs, 2017). In order to ensure equitable student outcomes, the research community should make greater efforts to develop new methods for addressing this shortcoming.

Recent work has sought to address this issue by investigating publicly-available, school-level differences in demographics, which can be useful when individual-level variation may be difficult or impossible to acquire data for (Wang & Beck, 2013; Pardos & Heffernan, 2010, Yudelson et al., 2014). In this study, I use this approach to better understand the role of social factors in students' self-regulated learning and motivation-related behaviors (as observed in an adaptive system) - behaviors whose effectiveness appears to be highly variable between groups, making a general-purpose design ineffective. I demonstrate that school-level demographics can be significantly associated with the relationships between students' help-seeking behavior, motivation, and outcomes (math performance and math self-concept). I do so in the context of Reasoning Mind (Khachatryan et al., 2014), an intelligent tutoring system for elementary mathematics. By studying the conditions under which these relationships vary across different demographic contexts, this study challenges implicit assumptions of generalizability of design choices and provides an evidence-based commentary on future research practices in the community surrounding how we consider diversity in our field's investigations.

Study #3. Varying effectiveness of methodological improvements in annotated data collection. The final upstream component investigated in this dissertation is a data collection methodology. Despite the abundance of data generated from students' activities in adaptive systems in education, the use of predictive modeling in these systems is limited

by the availability of labeled (annotated) data, which can be difficult to collect for complex educational constructs. A subfield of machine learning called Active Learning (AL) has been explored to improve the data labeling efficiency (Yang et al., 2018). AL trains a model and uses it, in parallel, to choose the next data sample to get labeled from a human expert. Due to the complexity of educational constructs and data, AL has suffered from the cold-start problem where the model does not have access to sufficient data yet to choose the best next sample to learn from. In this study, I explore the use of past data to warm start the AL training process.

More importantly, I examine the implications of differing contexts (urbanicity) in which the past data was collected. To this end, I use authentic affect labels collected through human observations in middle school mathematics classrooms to simulate the development of AL-based detectors of engaged concentration. I experimented with two AL methods (uncertainty sampling, L-MMSE) and random sampling for data selection. The results suggest that using past data to warm start AL training could be effective for some methods based on the target population's urbanicity. I show that mismatches in the urbanicity of the past data (and possibly other demographic dimensions) could be detrimental to effective model training in some cases. I provide recommendations on the data selection method and the quantity of past data to use when warm starting AL training in the urban and suburban schools. This study suggests that it may be problematic to assume uniform effectiveness of new data collection methods on different student populations, highlighting the need to conduct studies on demographic disparities in the benefits of methodological innovations.

Overview of Chapters

There are five chapters in this dissertation. Chapter 1 presents the overarching research problem along with the relevant background and theoretical foundation on bias in adaptive systems in education. After discussing the significance of this research, I summarize the purpose of the three studies, connecting each of them to the gaps identified in the literature. Chapter 2 presents the first study that investigates the non-conformance of empirical evidence to the widely accepted theoretical model of affect dynamics. Chapter 3 presents the second study that examines the differing implications of help-seeking and motivation design on student outcomes. Chapter 4 presents the third study that explores the impact of data contexts in the varying success of methodological innovation in optimizing annotated data collection. Chapter 5 presents a general discussion outlining the contributions of this dissertation and potential future directions towards an equitable design and development of adaptive learning systems.

Statement of Contribution

Chapter 1: Single-authored

Chapter 2: First-authored publication in which I led the conceptualization, methodology, data gathering, software, formal analysis, and writing (original draft and revisions)

Chapter 3: First-authored publication in which I led the conceptualization, methodology, data gathering, formal analysis, and writing (original draft and revisions)

Chapter 4: First-authored publication in which I led the conceptualization, methodology, formal analysis, and writing (original draft and revisions)

Chapter 5: Single-authored

CHAPTER 2

NON-CONFORMANCE OF DATA TO A WIDELY-ACCEPTED THEORETICAL MODEL OF EMOTION

Karumbaiah, S., Baker, R. S., Ocumpaugh, J., & Andres, A. (2021). A re-analysis and synthesis of data on affect dynamics in learning. *IEEE Transactions on Affective Computing*.

Abstract. Affect dynamics, the study of how affect develops and manifests over time, has become a popular area of research in affective computing for learning. In this paper, we first provide a detailed analysis of prior affect dynamics studies, elaborating both their findings and the contextual and methodological differences between these studies. We then address methodological concerns that have not been previously addressed in the literature, discussing how various edge cases should be treated. Next, we present mathematical evidence that several past studies applied the transition metric (L) incorrectly - leading to invalid conclusions of statistical significance - and provide a corrected method. Using this corrected analysis method, we reanalyze ten past affect datasets collected in diverse contexts and synthesize the results, determining that the findings do not match the most popular theoretical model of affect dynamics. Instead, our results highlight the need to focus on cultural factors in future affect dynamics research.

2.1 Introduction

Student affect in intelligent tutors and other types of adaptive and artificially intelligent educational systems has been shown to correlate with a range of other important constructs, including self-efficacy [1], analytical reasoning [2], motivation [3], and learning [4, 5]. Several research studies in the past decade have focused on building good quality affect detectors using physical and physiological sensors [6, 7, 8, 9], and interaction log data [10, 11, 12, 13]. Affect-sensitive interventions have been designed to improve student engagement [14], learning gains [5, 15, 16], and overall experience [17]. Developing effective real-time interventions depend on understanding how affect develops and manifests over time, an area of research termed *affect dynamics* (i.e. [18]), with a large body of research examining how students transition from one affective state to the next during learning activities (i.e., 2, 3, 5, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30].

The most commonly-cited model of affect dynamics in this context, put forward by D'Mello and Graesser [5], postulates that a specific set of affect transitions will be particularly prominent, but few empirical studies have matched that model's predictions, an issue which this paper investigates. Research has shown that affect plays three primary roles in learning and education: signaling, evaluation, and modulation. These roles refer to the ability of affective states to draw attention to learning challenges [31], appraise learning [32], and guide cognitive focus [22, 31, 33, 34]. These roles play a key function within the model [5] of affect dynamics during learning, which hypothesizes transitions between the educationally-important affective states of engaged concentration, confusion, frustration, and boredom (e.g., Fig. 1).

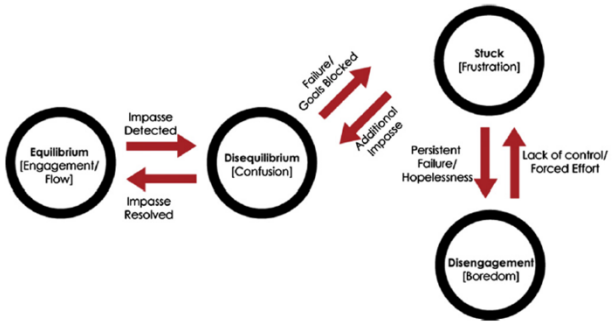


Figure 2.1 Theoretical model of affect dynamics by D'Mello and Graesser [5]

The primary model cited from the paper predicts that students who experience an impasse during the flow state will transition to a state of disequilibrium, which manifests itself as the affective state of confusion. If the student resolves this impasse, they are predicted to transition back to flow. If, however, the impasse is not resolved, students are hypothesized to become "stuck" (experienced as frustration). If the frustration persists, the model suggests the learner will disengage, transitioning to boredom. Two other links in this highly cited model (*frustration* → *confusion* and *boredom*→*frustration*) are also hypothesized as likely, but the justification for these transitions is not discussed as thoroughly. This same paper also presented two empirical lab studies to demonstrate support for the hypothesized transitions. While the results aligned with the majority of the transitions in the model, neither of the studies supported the transition of *frustration*→*confusion* and one of them failed to support *frustration*→*boredom*.

D'Mello and Graesser's model has been widely referenced (with around 400 citations as of this writing) by various research studies on affect dynamics, including many which have

used the ***L*** likelihood statistic advanced in [5, 23]. Described in detail in section 2.3, the ***L*** statistic compares a specific transition's frequency to the frequency that might be expected based on its originating and destination affective states and can be used in statistical significance testing to infer whether a transition is more likely than chance. However, empirical results across a range of learning environments have not aligned with the theoretical model's proposed affective transitions (see Section 2.2).

A number of factors may be contributing to the divergence between the theoretical model and these empirical results. These studies have varied in terms of population (from elementary school students to graduate students and from the US to the Philippines), the learning context (type of learning activity as well as laboratory versus classroom study), and both study methodology and analysis method. However, another key difference between D'Mello and Graesser [5] and other research is how the data are represented when a student remains in the same affective state across several observation points. In [5], only transitions between differing states were considered, whereas in many other studies (including earlier work by the same authors), a student remaining in the same affective state was considered to exhibit a self-transition that was included in calculations. More broadly, past affect dynamics studies have differed in exactly how they calculate affect transitions, particularly in how they treated the edge cases (see Section 2.3.2).

In this paper, we first provide a detailed analysis of the prior affect dynamics studies elaborating on the contextual and methodological differences in them. We then describe the steps involved in affect dynamics analysis using ***L*** with clarifications on the edge cases that have mostly been omitted from write-ups on how the prior affect dynamics studies

were conducted. Next, we present mathematical evidence that several past studies used the ***L*** statistic in ways that led to invalid conclusions of statistical significance and provided a correction to the interpretation of ***L*** statistic. Using a corrected analysis method, we re-analyze ten past affect datasets collected in diverse contexts and synthesize the results to find if there is empirical evidence for the D'Mello and Graesser's widely accepted model.

2.2 Prior Work on Affect Dynamics

To date, fifteen studies have used the ***L*** metric [23] to study the affect dynamics. The current study will focus primarily on the affective states included in the D'Mello and Graesser model (i.e., *boredom, engaged concentration, frustration, and confusion*), but as Table 2.1 summarizes, a range of other emotions have been included in these previously published papers (i.e., *anger, anxious, curiosity, delight, disgust, eureka, excitement, fear, happiness, neutral, sadness, and surprise*).

We use the term *engaged concentration* to refer to the affective state associated with flow [35], in line with the recommendations in [36], who noted that *flow* is a complex construct that goes beyond simply affective experience, also necessitating elements such as a perfect balance between challenge and ability. The reader should note that this state is alternately referred to in the affective dynamics literature as *flow*, *engagement, engaged concentration,* and *concentrating* due to different theoretical positions taken by the authors; however, the definitions used for this affective state are generally highly similar across papers.

Table 2.1 Affect States Studied in Previous Research on Affect Dynamics

Studies	BOR	ENG	DEL	FRU	SUR	NEU	CON	ANX	ANG	DIS	SAD	EUR	HAP	CUR	FEA	EXC
Andres & Rodrigo, 2014	x	x	x	x	x		x									
Botelho et al., 2018	x	x		x		x	x									
Baker, Rodrigo, & Xolocotzin, 2007	x	x	x	x	x	x	x									
Bosch & D'Mello, 2013	x	x		x			x									
Bosch, & D'Mello, 2017	x	x		x	x	x	x	x	x	x	x		x	x	x	
D'Mello & Graesser, 2012	x	x	x	x	x	x	x									
D'Mello et al., 2009	x			x	x	x	x	x	x	x	x	x	x	x		
D'Mello, Taylor, & Graesser, 2007	x	x	x	x	x		x									
D'Mello & Graesser, 2010	x	x	x	x	x		x									
Guia et al., 2011	x	x	x	x	x	x	x									
Guia et al., 2013	x	x	x	x	x	x	x									
McQuiggan et al., 2008; 2010	x	x	x	x			x	x	x		x				x	x
Ocumpaugh et al., 2017	x	x		x	x		x	x								
Rodrigo et al., 2008	x	x	x	x	x	x	x									
Rodrigo et al., 2011; 2012	x	x	x	x	x	x	x									

Categories studied in D'Mello & Graesser's Model are Highlighted in Gray. (BORed, ENGaged Concentration, DELight, FRUstration, SURprise, NEUtral, CONfused, ANXious, ANGer, DISgust, SADness, EUReka, CURious, FEAr, EXCited)

These studies have yielded a range of results. From the 15 studies considered, transitions that are both significantly more likely to occur than chance and align with the model of affect dynamics have been found predominantly in studies by D'Mello and his colleagues. *Engaged concentration→confusion* was reported in eight studies, including five by D'Mello et al. [4, 5, 21, 23, 37] as well as in studies by McQuiggan and colleagues [1, 26] and Ocumpaugh and colleagues [28]. However, fewer studies found support for other predicted transitions. *Confusion→engaged concentration* was reported in three D'Mello et al. studies [4, 5, 37] and in one study by Ocumpaugh et al. [28] and Botelho et al. [38]. Transitions of *confusion→frustration* (in [4, 5, 21, 37]) and *frustration→confusion* (in [4, 2, 21]) were reported exclusively in studies by D'Mello and his colleagues. *Frustration→boredom* was reported in D'Mello et al. studies [4, 5] and was marginally significant in one Rodrigo et al. study [3]. *Boredom→frustration* was reported in two studies by D'Mello et al. [5, 23] and in one study by Rodrigo and colleagues [30] and Botelho et al. [38].

2.2.1 Demographic Differences in Previous Work Examined

Across all of the hypothesized affective transitions, only one transition is seen in more than half of the studies, arguing that thus far, the theoretical predictions of this model are not being upheld. However, the 15 studies summarized in Table 2.2 differ noticeably in terms of the demographic characteristics of their samples, including age and the region where the research was conducted. It is possible that these differences may explain the inconsistencies in whether research supports the model.

Table 2.2 Summary of the Observed Methodological Differences Across 15 Studies on Affect Dynamics

	Region	Age	N	School/Grade Population	Learning System	Class v. Lab	Obs. Type/ Grain Size	Obs. Session	Self-trans	Aligned Transitions
Andres & Rodrigo, 2014	Quezon City, PH	13-16	60	Public school	Physics Playground	C	QFO	2hrs	Inc	0
Botelho et al., 2018	--	--	838	--	ASSISTments	C	Automated Detector. 20s	--	Exc	--
Baker et al., 2007	Manila, PH	14-19	36	High school	Inc. Machine	C	QFO ev. 60s	10min	Inc	0
Bosch & D'Mello, 2013	US	--	29	Undergrads	Unnamed	L	RJP on 100 fixed points	25min	Exc	3
Bosch, & D'Mello, 2017	Midwestern US	17-21	99	Undergrads	Unnamed	L	RJP on 100 fixed points	25min	Exc	5
D'Mello & Graesser, 2012	Southern US	--	28; 30	Undergrads	Auto-Tutor	L	RJP every 20s; fixed points	32min; 35min	Exc	4;5
D'Mello et al., 2007	Southern US	--	28	Undergrads	Auto-Tutor	L	RJP ev. 20s	32min	Inc	2
D'Mello et al., 2009	Southern US	--	41	Undergrads	Unnamed	L	RJP on fixed points	35min	Exc	1
D'Mello & Graesser, 2010	Southern US	--	28; 30	Undergrads	Auto-Tutor	L	RJP ev. 20s; fixed points	32min; 35min	Exc	3;3
Guia et al., 2011; 2013	Quezon City, PH	18-20	60	Undergrads	SQL Tutor	C	QFO ev. 200s	1hr	Inc	0
McQuiggan et al., 2008; 2010	US	21-60	35	Grad students	Crystal Island	L	SRI	35min	Inc	1
Ocumpaugh et al., 2017	New York, US	18-22	108	West Point	vMedic	C	QFO ev.122s	--	Inc	2
Rodrigo et al., 2008	Quezon City & Cavite Prov., PH	9-13	180	Private school	Ecolab	C	QFO	40min	Inc	1
Rodrigo et al., 2011; 2012	Quezon City, PH	43813	126	High school	Scatterplot Tutor	C	QFO ev. 200s	80min	Inc	1

* PH: Philippines, QFO: Qualitative field observation, RJP: Retrospective judgment protocol, SRI: self-report based on interactions, Inc: self transitions included, Exc: self transitions excluded.

All the studies by D'Mello and colleagues were conducted in the United States with undergraduate populations. By contrast, the studies by other researchers come from a wider range of demographics with students in middle school (private), high school (public and private), undergraduate programs, and graduate schools, from locations in the United States and in the Philippines. Differences in culture are known to influence variation in beliefs and personal dispositions towards emotional expression and moderation [39] and the frequency and emergence of certain affective states [40]. Likewise, age is known to

influence emotional expressivity [41, 42] and inhibition [43]. It is possible that some of the differences in results may be due to these factors; if so, this would suggest that D'Mello and Graesser's model may not be generalized across contexts.

2.2.2 Learning Settings

The studies were conducted across multiple instructional settings, including regular classroom environments and laboratory settings. Educational software used to investigate affective dynamics has covered a broad range of educational content, including mathematics [29, 30], biology [3, 26, 27], emergency medical practices [28], physics [19, 23, 44], computer literacy, and programming [4, 21, 24, 25, 37], and analytical problem solving [2]. The learning systems that have been used across these studies have also differed in terms of design. The Scatterplot Tutor, SQL-Tutor, AutoTutor and the other researcher-built learning environments used in studies conducted by D'Mello follow more linear designs wherein learners must complete problems before they are able to proceed. On the other hand, environments such as Physics Playground, Crystal Island, The Incredible Machine, vMedic, and Ecolab, are more open-ended systems that offer learners the opportunity to explore the range of possible solutions.

2.2.3 Data Collection Procedure, including Observation Grain-Size and Session Duration

Six of the 15 studies use the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP, [45]), a momentary time sampling method that uses a holistic coding practice to code for both affect and behavior. In this protocol, students are observed for up to 20 seconds in a round-robin manner throughout the given observation period to ensure uniform frequencies

of student observation. The protocol is enforced by an Android application known as the Human Affect Recording Tool (HART, [45]).

By contrast, D'Mello and his colleagues have used self-reporting methods, collecting affect data through retrospective judgment protocols which synchronize webcam video of students' faces to screen capture of the learning environment [2, 4, 5, 21, 23, 37]. McQuiggan and his colleagues also collected self-reported data, but used in-game dialogs to collect spontaneous reports rather than using a retrospective technique [26, 27].

Observation sessions in this research varied in length, ranging from 10 minutes [20] to 2 hours [19], potentially influencing the affect that emerges during observation. Prolonged exposure to similar tasks may produce fatigue or boredom [46], decreasing learner performance [47]. Correspondingly, it may also increase students' susceptibility to what [23] describes as vicious cycles of boredom, where learners are unable to transition to other affective states.

2.2.4 Differences in the Treatment of Self-Transitions Between Studies

All of the studies considered in this section analyze time-series data (e.g., the order of the occurrences of each affective state), but they have been inconsistent in their treatment of "self-transitions," which occur when a student remains in the same affective state over two (or more) consecutive observations. In more recent studies, D'Mello and colleagues have removed self-transitions during the data preparation stage [2, 4, 5, 21, 37], as have Botelho and colleagues [38]. For example, a sequence of *confusion, frustration, frustration, boredom* has one self-transition (from *frustration* to *frustration*), and would be modified into *confusion, frustration, boredom*.

However, this practice is not followed in all work. Nearly a dozen other studies conducted in this field do not report discarding self-transitions in their data processing [3, 24, 25, 26, 27, 28, 29, 44], including work by D'Mello and his colleagues (e.g. [23]). The choice of including or excluding self-transitions is likely based on the goal of the research study; including self- transitions may suppress non-self transitions. If some affective states are particularly persistent [36], including self-transitions in analysis helps better understand each state's persistence, but dilutes evidence for transitions between different affective states. In contrast, excluding self-transitions could be a better choice if the goal is to reveal a larger number of affective patterns that might otherwise be suppressed by the presence of persistent affective states. However, as we discuss in sections 2.3.3 and 2.3.4, this seemingly small step may have disproportionate effects on study outcomes, particularly in terms of this decision's impact on the interpretation of commonly used statistics.

2.2.5 Summary of Past Results

Tables 2.3 and 2.4 present the findings of the affect dynamics studies conducted in the past. Note that blank cells in Table 2.3 represent transitions that were either not studied or not reported in each paper. These transitions are also not counted in the calculations in Table 2.4. Also, Table 2.3 does not include [29] as this study only reported self-transitions. The studies that did not report discarding self-transitions predominantly reported null effects for the non-self transitions. (See Table 2.3, column 6.)

Table 2.3 Significance of the Transitions Reported in Previous Research

Studies	ENG_ENG	ENG_CON	ENG_FRU	ENG_BOR	CON_ENG	CON_CON	CON_FRU	CON_BOR	FRU_ENG	FRU_CON	FRU_FRU	FRU_BOR	BOR_ENG	BOR_CON	BOR_FRU	BOR_BOR
Andres & Rodrigo, 2014	+	Ø	Ø	-	Ø	Ø	Ø	Ø	-	Ø	+	Ø	-	-	Ø	+
Baker, Rodrigo, & Xolocotzin, 2007	+	-	Ø	-	Ø	+	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	+
Botelho et al., 2018		Ø	-	+	+		-	-	+	Ø		-	+	Ø	+	
Bosch & D'Mello, 2013		+			Ø		+			+		Ø	-		Ø	
Bosch, & D'Mello, 2017		+	Ø	Ø	+		+	Ø	Ø	+		+	+	Ø	Ø	
D'Mello & Graesser, 2012 [Study 1]		+	Ø	Ø	+		+	Ø	Ø	Ø		Ø	Ø	Ø	+	
D'Mello & Graesser, 2012 [Study 2]		+	Ø	Ø	+		+	Ø	+	Ø		+	Ø	Ø	+	
D'Mello et al., 2009										+						
D'Mello, Taylor, & Graesser, 2007	+	+	-	-	Ø	+	Ø	-	Ø	Ø	Ø	Ø	Ø	Ø	+	+
D'Mello & Graesser, 2010 [Study 1]		+			+		+									
D'Mello & Graesser, 2010 [Study 2]		+			+		+									
Guia et al., 2011	Ø	Ø	Ø	Ø	-	Ø	-	Ø					Ø	Ø	-	Ø
Guia et al., 2013	Ø	Ø	Ø	Ø	-	Ø	Ø	Ø	-	Ø	+	Ø	Ø	-	-	+
McQuiggan et al., 2008; 2010	Ø	+	-	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
Ocumpaugh et al., 2017	-	+	Ø	+	+	-	Ø	Ø	Ø	Ø	-	-	Ø	+	Ø	-
Rodrigo et al., 2008 [Control]	+	Ø	Ø	-	Ø	Ø	Ø	Ø	-	Ø	+	+	-	Ø	Ø	+
Rodrigo et al., 2008 [Experiment]	Ø	Ø	Ø	-	Ø	Ø	Ø	Ø	-	Ø	+	Ø	-	-	Ø	+
Rodrigo et al., 2011 [Control]	+					+					Ø					+
Rodrigo et al., 2011 [Experiment]	+					+					Ø					+
Rodrigo et al., 2012 [Control]	+	-	Ø	Ø	-	+	Ø	-	Ø	Ø	Ø	Ø	Ø	-	+	+
Rodrigo et al., 2012 [Experiment]	+	-	+	-	-	+	Ø	-	Ø	Ø	Ø	Ø	Ø	-	+	+

+ indicates a significant positive transition, - indicates a significant negative transition and Ø indicates an insignificant transition. Transitions not studied or not reported are left blank. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Only the states from D'Mello model are included - ENGaged concentration, CONfused, FRUstration, BORed. D'Mello studies and transitions studied in D'Mello & Graesser's model are highlighted in grey.

Table 2.4 Proportion of Studies That Reported Positive, Negative, and Null Values for the Non-Self Transitions

Transition	The proportion of studies reporting a significant **positive** L value		The proportion of studies reporting a significant **negative** L value		The proportion of studies reporting a **non-significant** L value	
	INC	EXC	INC	EXC	INC	EXC
ENG_CON	0.25	1.00	0.25	0.00	0.50	0.00
ENG_FRU	0.08	0.00	0.25	0.00	0.67	1.00
ENG_BOR	0.17	0.00	0.50	0.00	0.33	1.00
CON_ENG	0.17	0.83	0.33	0.00	0.50	0.17
CON_FRU	0.00	1.00	0.17	0.00	0.83	0.00
CON_BOR	0.00	0.00	0.33	0.00	0.67	1.00
FRU_ENG	0.09	0.33	0.36	0.00	0.54	0.67
FRU_CON	0.00	0.60	0.00	0.00	1.00	0.40
FRU_BOR	0.09	0.50	0.18	0.00	0.72	0.50
BOR_ENG	0.08	0.25	0.25	0.25	0.67	0.50
BOR_CON	0.08	0.00	0.42	0.00	0.50	1.00
BOR_FRU	0.33	0.50	0.17	0.00	0.50	0.50

INC - Studies that includes self-transitions total = 12); EXC - Studies that exclude self-transitions (total = 7). The transitions that were either not studied or not reported are not counted for that study. Transitions studied in D'Mello & Graesser's model are highlighted in grey.

The proportion of studies that report significant, positive likelihood for non-self transitions, and in particular for the transitions hypothesized by the D'Mello and Graesser model, is higher in the set where the self-transitions were discarded. This could imply that discarding self-transitions can result in higher conformance with the D'Mello and Graesser model. But, currently, all the studies (except one) in the set where self-transitions are discarded correspond to studies conducted by D'Mello and colleagues in lab settings involving undergraduate population from the United States and use retrospective affect judgments to collect affect data in observation sessions that are around 30 minutes long.

Previous work [48] removed self-transitions and reanalyzed data collected in a classroom setting in the Philippines, where 180 eighth and tenth graders used a learning game called Physics Playground [49] for 2 hours. In this study, affect data was collected through field observations using the BROMP protocol [45]. This paper found that excluding self-transitions increased the proportion of transitions that occur above chance, yet it did not lead to having a larger number of transitions that were more likely than chance and conformed to D'Mello and Graesser's theoretical model.

2.3 The Transition Metric *L*

In this section, we present the metric most commonly used in affect dynamics research, the ***L*** statistic [5], and provide our recommendations for how to handle the several special cases that occur in using this metric. Next, we discuss issues surrounding the underlying assumptions of this metric, provide mathematical evidence that the metric has been used in an invalid fashion in many past papers, and propose a correction for future use.

2.3.1 The L Statistic for Affect Dynamics

Given an affect sequence, the ***L*** statistic [23] calculates the likelihood that an affective state (*prev*) will transition to a subsequent (*next*) state, given the base rate of the next state occurring.

$$L(prev \rightarrow next) = \frac{P(next|prev) - P(next)}{1 - P(next)} \quad (1)$$

The expected probability for an affective state, *P(next),* is the percentage of times that the state occurred as a next state. Thus, the first affective state in a student's sequence will be excluded from this calculation since this state cannot take the role of a next state. For instance, for a state sequence AABB, the probability of the state A as the next state, P(A_next) is 0.33 (from ABB) instead of 0.5. Similarly, the calculation of the *prev* state excludes the last state in the sequence. The conditional probability, *P(next|prev)* is given by:

$$P(next \mid prev) = \frac{Count(prev \rightarrow next)}{Count(prev)} \quad (2)$$

where *Count(prev → next)* is the number of times the *prev* state transitioned to the next state, and *Count(prev)* is the number of times the state in *prev* occurred as the previous state. In the example sequence of AABB, Count(B_prev) is 1 (from AAB) instead of 2 as the last state in the sequence cannot be a *prev* state for any transition.

The value of ***L*** varies from -∞ to 1. D'Mello and Graesser [5] state that "the sign and the magnitude of ***L*** is intuitively understandable as the direction and size of the association." As has been expanded in subsequent papers [2, 4, 5, 19, 21, 24, 25, 26, 27, 29, 30, 37, 48,

50, 51], $\boldsymbol{L} = 0$ is treated as chance, while $\boldsymbol{L} > 0$ and $\boldsymbol{L} < 0$ are treated as transitions that are more likely or less likely (respectively) than chance.

To perform affect dynamics analysis across all students in an experiment, first the $\boldsymbol{L}$ value for each affect combination is calculated individually per student. Next, as [5, pg. 7] recommends, the researcher runs "one-sample [two-tailed] t-tests to test whether likelihoods were significantly greater than or equivalent to zero (no relationship between immediate and next state)," on the sample of individual student $\boldsymbol{L}$ values for each transition. Lastly, a Benjamini-Hochberg post-hoc correction procedure is used by some of the research groups conducting this type of analysis [3, 19, 29, 30, 38, 48, 51] to control for false-positive results since the set of hypotheses involves multiple comparisons – however, some early research papers by these groups omitted this step, and other research groups have not used any type of post-hoc correction at all.

2.3.2 Special Cases when Implementing L

There are several special cases in the calculation of $\boldsymbol{L}$ where there is no consensus in the literature on how to perform the calculation. [48] has recommended the following treatment:

1. When any affective state (A_n) being considered in a given study is not present for a given student's observation period:
   a. If transitions to A_n do not occur for that student, then $P(next) = 0$ and $P(next \mid prev) = 0$, and thus, $\boldsymbol{L} = 0$.

b. If transitions from A_n also do not occur, then we do not know what affective state would have followed A_n, and thus, $\boldsymbol{L}$ = undefined.

2. Following from case 1, if a student remains in a single affective state (A_s) throughout an observation period, the calculations differ based on whether or not the self-transitions are included.

   a. If self-transitions are included in the analyses, then:

      (1) Transitions from A_s to any other affective state (e.g., A_n) do not occur, and therefore, as in 1a, $\boldsymbol{L} = 0$ for any transition out of A_s.

      (2) Transitions to A_s from any other affective state (e.g., A_n) do not occur, and therefore, as in 1b, $\boldsymbol{L}$ = undefined.

   b. If self-transitions are discarded in the analyses, an affect sequence consisting of repeated observations of the same affective category is reduced to a single observation of that affective state. In this case, no transitions occur, and therefore $\boldsymbol{L}$ = undefined for all possible sequences being studied.

It is not always clear how these special cases are treated in the past published research. In this study, we follow [48]'s definition of $\boldsymbol{L}$ as outlined above.

2.3.3 The Case of Self-Transitions

One other special case that is not fully discussed in most of the literature is the case of self-transitions. In the majority of articles written by D'Mello's group and other groups' articles as well [i.e., 38], self-transitions (where the student remains in the same affective state for more than one step in a sequence) are removed. This straightforward procedure seems quite logical, but there is evidence that something may be wrong with the resulting calculations.

Notably, in [38], after removing self-transitions, all transitions into the affective state of engaged concentration were more likely than chance. As such, it may be worth examining the mathematical assumptions of this procedure. Specifically, while calculating the transition likelihood from the affective state of M_t (*prev*) to M_{t+1} (*next*), D'Mello explains that, "...if M_{t+1} and M_t are *independent* [emphasis added], then $Pr(M_{t+1}|M_t) = Pr(M_{t+1})$" [15]. However, removing self-transitions violates the assumption of independence between M_{t+1} and M_t, as M_{t+1}can now only take values other than M_t. For instance, if there are three states (A, B, C) and if $M_t = \mathrm{A}$, then M_{t+1} can only take the value of either B or C if self-transitions are not allowed. Hence, when self-transitions are excluded, $Pr(M_{t+1}|M_t) \neq Pr(M_{t+1})$ when M_t and M_{t+1} are independent.

Another sign of potential problems is found in [5], when that paper draws an analogy between $\boldsymbol{L}$ statistics and Cohen's kappa, saying, "The reader may note significant similarity to Cohen's kappa for agreement between raters and indeed the likelihood metric can be justified in a similar fashion." Although this analogy seems compelling based on the similarity of the equations between the two metrics, it is worth noting that there is a difference between the range of values the two statistics can take. While the value of $\boldsymbol{L}$ varies from -∞ to 1 [44], the value of Cohen's kappa varies from -1 to 1.

These findings raise the question: if a transition occurs at chance, and self-transitions are excluded, is the value of $\boldsymbol{L}$ still 0? We address this concern in section 2.3.4.

2.3.4 Correcting Chance L Value

For a state space with n affective states (n $>$ 2), there would be n^2 unique transitions if we include self-transitions, but only $n^2 - n$ unique transitions if we exclude self-transitions [50]. Thus, at chance, the expected probability is

$$P(next) = \frac{n}{n^2} = \frac{1}{n} \qquad \text{if self-transitions are included}$$

$$P(next) = \frac{n-1}{n^2-n} = \frac{1}{n} \qquad \text{if self-transitions are excluded}$$

However, at chance, the conditional probability is

$$P(next \mid prev) = \frac{1}{n} \qquad \text{if self-transitions are included}$$

$$P(next \mid prev) = \frac{1}{n-1} \qquad \text{if self-transitions are excluded}$$

Plugging these into the original equation of ***L*** (equation 1), the value of ***L*** at chance is

$$L = 0 \qquad \text{if self-transitions are included}$$

$$L = \frac{1}{(n-1)^2} \qquad \text{if self-transitions are excluded}$$

This finding shows that the ***L*** statistic must be interpreted differently depending on how many affective categories are being observed. Table 2.5 shows the values at chance, depending on how many affective states are being observed.

Table 2.5 The Value of L That Represents Chance, For Varying State Space

n	3	4	5	6	7	8
chance *L*	0.25	0.11	0.0625	0.04	0.0277	0.0204

As noted above, affect dynamics is most frequently studied in terms of four or five affective states. In such a setup (n = 5), the ***L*** value at chance is ***L***=0.0625. For the smallest

reasonable state space (n = 3), the ***L*** value at chance reaches 0.25. As the number of affective states observed increases, the impact of the difference between including and excluding self-transitions decreases (Table 2.5). This is particularly a problem if a statistical significance test is conducted that compares to a chance value of 0. Take, for instance, a case where three affective states are studied, and ***L*** is reliably 0.15 for a specific transition. In this case, a comparison to 0 may find that a transition occurs more often than chance when it actually occurs less often than chance.

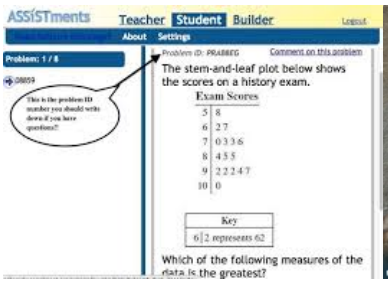
2.4 Re-analysis of Prior Data

2.4.1 Datasets used in this analysis

In order to collect affect datasets from diverse contexts, we reached out to authors of papers that have previously reported work on affect in learning and obtained 10 datasets. Figure 2.1 shows pictures of the eight online learning systems involved in several of these studies, and Table 2.6 provides an overview of these datasets, which are described in greater detail in the following subsections. Two of the datasets were collected in classroom studies with no learning system. In addition to providing information about the learning setting involved in each study, these sections (2.4.1.1-2.4.1.10) also outline the demographics of the participants in the study.

Nine of these datasets were produced using the BROMP protocol [44] to collect affect in classroom settings. BROMP is a momentary time sampling method where students are briefly observed by certified coders one after another, in repeated round-robin cycles. BROMP has been used by over 160 researchers and practitioners in seven countries for field observations.

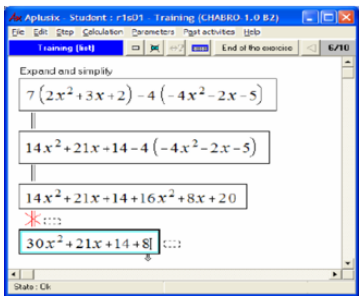
There are slight variations in how BROMP may be implemented in the different countries that are represented in this study. BROMP observations in the Philippines have historically used two coders making simultaneous observations on the same student. In contrast, BROMP coders in the United States record observations independently (after inter-rater checks are complete). In the former case, the observations from the different coders are merged to form a single affect sequence sorted by the time of observation.



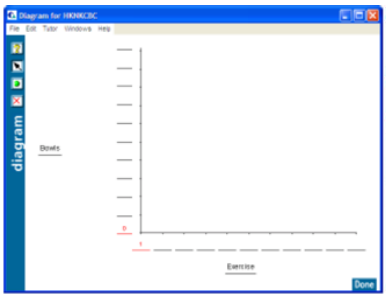
(a)

(b)

(c)

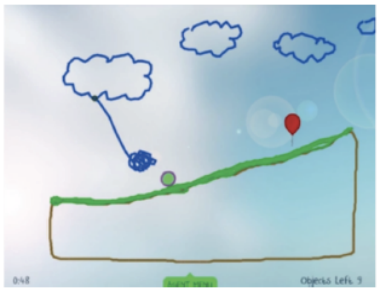


(d)

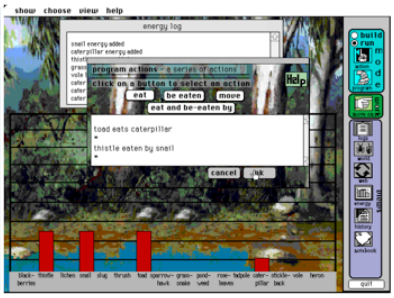


(e)

(f)



(g)



(h)

Figure 2.2 Snapshots of the different learning systems studied in this work. In parenthesis is the dataset number. From top left - a) ASSISTments (#1); b) vMedic (#4); c) Crystal Island (#5); d) Aplusix (DS#6); e) Scatterplot (#7); f) SQL-Tutor (#8); g) Physics Playground (#9); h) Ecolab (#10).

Table 2.6 Description of the 12 Datasets Reanalyzed In This Paper

Dataset#	Learning Sys/ Classroom	No. of Students	No. of Coders at Once	No. of Observations Collected	No. of Students Dropped	No. of Observations				
						ENG	CON	FRU	BOR	NA
1	Assistments	856	1	7673	167	5039	469	239	657	1269
2	Classroom	371	1	3793	58	2957	27	27	618	191
3	Classroom	18	1	5308	0	4012	38	2	925	331
4	vMedic	117	1	755	32	435	174	32	73	41
5	Crystal Island	35	0	592	6	249	77	35	19	212
6	Aplusix	140	2	3640	7	2641	494	101	126	278
7A	Scooter Control	61	2	2976	0	1293	1406	13	179	85
7B	Scooter Experiment	64	2	3072	0	1142	1605	8	204	113
8	SQL Tutor	29	2	1044	0	481	388	39	53	83
9	Physics Playground	241	2	62502	0	46040	3962	3469	2557	6474
10 A	Ecolab Control	90	2	4560	2	2855	638	235	596	236
10 B	Ecolab Experiment	90	2	4560	0	3164	621	131	420	224

2.4.1.1 Dataset #1- ASSISTments

Affect data from 856 students was collected in six schools across urban, suburban, and rural settings [3]. A total of 7,663 field observations were collected. ASSISTments (Figure

2.2 a) is a free web-based platform which is used by students in the classroom and at home to practice the learning content assigned by their teacher. It is designed to provide immediate feedback and on-demand hints and sequences of scaffolding to support students when they make errors.

2.4.1.2 Dataset #2- Elementary Classes

Godwin and colleagues [53] conducted observations across twenty-two classrooms that were selected from 5 charter schools located in or near a medium-sized Northeastern city in the United States of America. Students observed were between kindergarten and fourth grade. The average class size was 21 students (10 males, 11 females). The number of children observed per session ranged from 15 to 22 children. The observation sessions were staggered across three time periods, and a total of 128 observation sessions were conducted in their study. Each observation session lasted approximately one hour. The average number of observations per session was 346.13, and the average number of observations per student within a session was 19.27.

2.4.1.3 Dataset #3- Graduate Level Classes

DiStefano and colleagues [54] observed students in an introductory methods course in an urban graduate school of education in the Northeast, United States. There are approximately 25 full-time and 408 part-time Graduate School of Education students (81% female and 19% male; 53% White, 19% Black, 11% Asian, 16% unknown, & 1% Hawaiian). Students participating in the course were observed and coded across four different classroom conditions of class lecture, class discussion, small group work, and transition periods for one semester.

2.4.1.4 Dataset #4 - vMedic

Ocumpaugh and her colleagues [28] collected affect data from 108 West Point cadets (ages of 18-22) using vMedic (a.k.a. TC3Sim). vMedic (Figure 2.2b) is a virtual world developed for the US Army to provide training in combat medicine and battlefield doctrine around medical first response. Two BROMP-certified coders observed the trainees while they used vMedic for up to 25 minutes. The coders observed different trainees at a given time, coding for surprise and anxiety as well as more commonly-studied affective states. Each trainee was observed once every 122 seconds (std dev = 100.14), leading to a total of 756 observations.

2.4.1.5 Dataset #5 - Crystal Island

Affect data in crystal island environment (Figure 2.2 c) was collected by McQuiggan and colleagues [1, 26], where they observed 35 graduate students ranging in age from 21 to 60 (M = 24.4, SD = 6.41). This included 9 females and 26 males, and 60% were Asian (n = 21) and approximately 37% were Caucasian (n = 13). Participants interacted with crystal island for 35 minutes and self-reported their affective state via an in-game dialog from a selection of ten affective states (anger, anxiety, boredom, confusion, delight, excitement, fear, engaged concentration, frustration, and sadness). A total of 592 self-report of affective states was collected. Crystal Island is a game-based learning environment designed for middle-school students in the domains of microbiology and genetics to develop deeper understandings about scientific knowledge.

2.4.1.6 Dataset #6 - Aplusix

Rodrigo and colleagues [29] collected data from 140 high school students (ages 12-15; 83 females, 57 males) using Aplusix in 2008 within four schools within Metro Manila and one school in the Province of Cavite (Philippines). Students used Aplusix II [55], an algebra learning assistant that teaches students how to balance equations (Figure 2.2 d). In each session, ten students were observed for 45 minutes. Each student has 13 observations spaced three minutes apart for a total of 3640 observations across all students.

2.4.1.7 Dataset #7a and 7b – Scatterplot Tutor

Scatterplot Tutor data from 125 students (ages 12-14) was collected in 2008 from an urban high school in Quezon City in the Philippines during a project with both a control (7A) and an experimental (7B) condition [29, 30]. Scatterplot Tutor (Figure 2.2 e) is a Cognitive Tutor that teaches the generation and interpretation of scatter plots. Both datasets 7A and 7B were collected from a group of ten students over 80-minute learning sessions yielding 24 observations per student. The control group consisted of 61 students, yielding a total of 1,464 observations, while the experimental group consisted of 64 students, resulting in a total of 1,536 observations. In dataset 7B, the experimental condition, an interactive pedagogical agent named Scooter designed to reduce gaming the system behavior, was shown on-screen while the students interacted with the learning system. For the purpose of this study, data from the control group and the experimental group are treated as two separate datasets (7A and 7B) in order to account for any influences Scooter may have had on the emergence of affective states.

2.4.1.8 Dataset #8 – SQL-Tutor

Guia and colleagues [24] collected data on affect from 29 third-year undergraduate students from Ateneo de Manila University in the Philippines while using SQL-Tutor. SQL-Tutor [27] (Figure 2.2 f) is an intelligent tutor that is designed to teach Structured Query Language (SQL). The participants of this study were in a course that required knowledge in database programming, but none had previously used SQL-Tutor. The participants were randomly divided into three sections and were asked to use SQL-Tutor for 60 minutes. Each student was observed once per 200 seconds leading to a total of 1044 observations.

2.4.1.9 Dataset #9 - Physics Playground

The Physics playground data was collected in 2015 from 180 students: 120 8th-graders and 60 10th-graders from Baguio, Cebu, and Davao, in the Philippines [49]. Students spent 2 hours using Physics Playground (Figure 2.2 g), a learning environment that teaches qualitative physics to secondary students [56]. In this 2-dimensional game, students sketch different objects like pendulums, ramps, levers, and springboards to guide a ball to touch a balloon. Laws of physics apply to all the objects on the screen. Each student was observed approximately once per minute. On average, there were 135 observations per student, giving a total of 24,330 observations.

2.4.1.10 Dataset #10a and 10b - Ecolab

Rodrigo and colleagues [3] collected affect data from 180 students from two private, co-educational grade schools in the Philippines (ages 9-13) while they used the Ecolab learning system (Figure 2.2 h) to learn about food webs and chains. There were ten students per observation session, five in control (Ecolab) and five in experimental (M-Ecolab) condition. In M-Ecolab, the system was enhanced with an affective learning companion

who modified its demeanor based on automated assessments of the learner's degree of motivation. Students used the system for 40 minutes, and each student was observed for affect 12 times using BROMP.

In the current study, this dataset is split between the control (10A) and experiment conditions (10B) and used separately for the analysis. Each of the sub-datasets consists of 90 students and contains a total of 4560 observations across all students.

2.4.2 Affect Distribution Across Datasets

Table 2.7 Mean and Standard Deviation of the Proportions of the Affective States Across the Students In The 12 Datasets Reanalyzed in this Paper

Dataset#	Learning Sys. / Classroom	ENG	CON	FRU	BOR	NA
1	Assistments	0.628 (0.289)	0.068 (0.13)	0.03 (0.078)	0.097 (0.171)	0.177 (0.217)
2	Classroom	0.781 (0.176)	0.007 (0.028)	0.007 (0.028)	0.162 (0.143)	0.05 (0.101)
3	Classroom	0.756 (0.056)	0.007 (0.005)	0 (0.002)	0.174 (0.061)	0.062 (0.017)
4	vMedic	0.572 (0.315)	0.247 (0.286)	0.047 (0.103)	0.09 (0.157)	0.043 (0.102)
5	Crystal Island	0.453 (0.334)	0.118 (0.124)	0.058 (0.091)	0.03 (0.056)	0.342 (0.287)
6	Aplusix	0.726 (0.185)	0.136 (0.107)	0.028 (0.057)	0.035 (0.076)	0.076 (0.087)
7A	Scooter Control	0.437 (0.252)	0.469 (0.221)	0.004 (0.015)	0.061 (0.15)	0.028 (0.043)
7B	Scooter Experiment	0.372 (0.207)	0.522 (0.181)	0.003 (0.01)	0.066 (0.145)	0.037 (0.058)
8	SQL-tutor	0.461 (0.166)	0.372 (0.168)	0.037 (0.072)	0.051 (0.101)	0.08 (0.085)
9	Physics Playground	0.74 (0.147)	0.062 (0.062)	0.059 (0.071)	0.032 (0.062)	0.107 (0.101)
10A	Ecolab Control	0.624 (0.21)	0.136 (0.11)	0.054 (0.094)	0.142 (0.178)	0.044 (0.079)
10B	Ecolab Experiment	0.675 (0.188)	0.135 (0.117)	0.031 (0.071)	0.115 (0.16)	0.043 (0.073)

Standard deviations presented in paretheses.

The descriptive statistics on the distribution of the affective states across these states are given in Table 2.7. In order to be consistent, affective states other than the four theorized in the D'Mello and Graesser model have been converted to a not-applicable (N/A) label. Overall, across datasets, there was a relatively high incidence of *engaged concentration* followed by *confusion* and *boredom,* and there was a relatively low incidence of *frustration*.

2.5 Methods

This study reanalyzes the 12 datasets outlined in the previous section. Specifically, we standardize the treatment of transition types and edge cases that have been identified as sources of potential discrepancies in the results between studies. This allows us to compare the results for individual datasets to determine which show the most conformity to the D'Mello & Graesser model. Then we apply Stouffer's Z, a method that allows us to summarize results across multiple datasets.

2.5.1 Standardizing the Analysis of Transition Types and Edge Cases

As discussed in section 2.3, in studies that included self-transitions, the results for non-self transitions were less likely to be positive in direction and less likely to be statistically significant. While understanding the persistence of affective states might be important in practice (algorithms designed to trigger interventions, for instance), focusing on out-of-state transitions could be more important for a theoretical model of affect dynamics. As such, we have decided to exclude self-transitions in this work, and we have reanalyzed these datasets accordingly.

In addition, we have standardized our treatment of edge cases. Specifically, we have ensured that in cases where a student remains in the same affective state throughout the observation session, we discard the student from the analysis. The number of students who remained in a single affective state and were omitted from the analysis is given in Table 2.6.

2.5.2 Stouffer's Z to Summarize Significance Levels from Multiple Affect Datasets

In order to determine whether transitions are significantly more likely than chance across datasets, we combine p-values from the independent significance tests conducted on the multiple affect datasets, using Stouffer's Z [57], also known as the sum of Z's method, a classic method for summarizing significance values in the social sciences [58] where the datasets do not contain any of the same subjects (which we believe to be true of these datasets). For the k independent tests (k = number of affect datasets), Stouffer's Z is given by

$$\sum_{i=1}^{k} z(p_i)/\sqrt{k} \qquad (1)$$

where, p_i is the p-value from the ith affect dataset. This statistic can then be used in a Z statistical test. This is repeated for all the 12 non-self-transitions being studied in this paper. Since the $\boldsymbol{L}$ values can take both positive and negative values, we are using the two-tailed version of Stouffer's Z. For negative $\boldsymbol{L}$ values, the corresponding Z scores are converted to a negative value. This method looks across all tests to see what the aggregate evidence is in favor of there being a significant relationship. By the nature of this method (similarly to the more complex methods sometimes used in modern meta-analysis), one finding with very strong evidence can outweigh multiple null effects. Though Stouffer's Z is sometimes

used in meta-analysis, readers are cautioned against interpreting our study as a true meta-analysis – our study involves re-analysis of several datasets that we were able to obtain rather than a traditional meta-analysis, which functions solely from the information available in published papers and attempts to exhaustively survey all relevant papers.

2.6 Results

2.6.1 Analyses of Individual Datasets

Table 2.8 summarizes the results of the individual tests conducted on the 12 non-self-transitions in the 12 affect datasets, with a corrected ***L*** metric [50]. Across the possible 140 results (4 transitions had undefined ***L*** value), only 24 tests yielded transitions significantly more likely than chance, as compared to 59 tests that resulted in transitions significantly less likely than chance and 57 null results.

Of the 24 tests with significantly positive results, 15 belong to transitions in or out of *engaged concentration*. Correspondingly, the transitions out of *engaged concentration* have relatively few null results. In contrast, transitions out of frustration have the highest number of null results (25) and only one positive result. It is worth noting that, across the dataset overall, *engaged concentration* is the most common affective state, whereas *frustration* is most rare.

Across datasets, some studies seem to have more null values (e.g., vMedic, Crystal Island) than the others (particularly studies involving Physics Playground, ASSISTments). This may be an attribute of these systems, but it also may be due to the quantity of data. The studies with more null results were also the studies with smaller sample sizes or briefer duration of observations.

Table 2.8 Significance of the Transitions Tested in the Current Analysis for the Twelve Affect Datasets

Dataset#	Learning System / Classroom	ENG_CON	ENG_FRU	ENG_BOR	CON_ENG	CON_FRU	CON_BOR	FRU_ENG	FRU_CON	FRU_BOR	BOR_ENG	BOR_CON	BOR_FRU
1	ASSISTments	+	+	+	+	+	-	Ø	Ø	Ø	+	-	+
2	Classroom	+	+	+	Ø		Ø	Ø	-	Ø	+	-	-
3	Classroom	+	-	-	Ø	-	Ø				-	+	+
4	vMedic	Ø	-	Ø	Ø	Ø	-	Ø	Ø	Ø	Ø	Ø	Ø
5	Crystal Island	-	-	-	Ø	Ø	Ø	-	Ø	-	Ø	Ø	Ø
6	Aplusix	+	+	-	Ø	-	-	Ø	Ø	-	-	Ø	Ø
7A	Scooter Control	-	-	+	-	-	-	Ø	Ø	Ø	Ø	-	-
7B	Scooter Experiment	-	-	-	-	+	-	Ø	Ø	Ø	Ø	-	-
8	SQL-Tutor	Ø	-	-	Ø	-	-	Ø	Ø	Ø	Ø	-	-
9	Physics Playground	-	-	-	-	+	-	-	-	+	-	-	+
10A	Ecolab Control	+	-	-	Ø	-	Ø	-	Ø	Ø	-	Ø	Ø
10B	Ecolab Experiment	+	-	-	Ø	Ø	+	Ø	Ø	Ø	-	-	Ø
	Total +	6	3	3	1	3	1	0	0	1	2	1	3
	Total Ø	2	0	1	8	3	4	8	9	8	5	4	5
	Total -	4	9	8	3	5	7	3	2	2	5	7	4

+ indicates a significant positive transition, - indicates a significant negative transition and Ø indicates a null effect. Transitions never seen or with undefined L value are left blank. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

Looking across the 12 datasets, we find that many of the transitions postulated by [5] are not statistically significantly more likely than chance (and in fact are often less likely than chance), a finding noted in several of those earlier papers (see discussion in section 2.2). In fact, the only transition where a significant positive result is seen in a major of datasets is *engaged concentration -> confusion*. By contrast, other transitions are almost never statistically more likely than chance: 1/12 datasets for *confusion -> engaged concentration*, 1/12 datasets for *frustration -> confusion,* and 0/12 datasets for *frustration -> boredom*. It is also true that across all possible transitions (including the ones not in the theoretical model), no transition other than *engaged concentration -> confusion* has a majority of positive transitions.

2.6.2 Analyses Across Datasets

Since it is possible that the results of individual studies do not provide a sufficient sample to find a significant effect, it is important to test whether or not any of these transitions may be significant if the studies were aggregated. However, as Table 2.9 shows, aggregating across studies using Stouffer's Z does not increase the number of transitions that show statistically significant, positive effects. Instead, the only transition that is statistically significantly more likely than chance remains *engaged concentration -> confusion*. There are seven other transitions with a statistically significant result, but all of those have a negative Z score, indicating that they are statistically significantly less likely than chance. Among the other six transitions postulated in the [5] model, only one transition is significantly more likely than chance (*engaged concentration -> confusion*). Two of their transitions have a null result - *confusion -> engaged concentration* and *frustration ->*

boredom. Lastly, three transitions in the hypothesized model are significantly less likely than chance - *confusion -> frustration, frustration -> confusion,* and *boredom-> frustration*. It is worth noting that the original studies in [5] also had little to no support for the transitions *frustration -> confusion* and *frustration -> boredom*.

Table 2.9 Stouffer's Z and Combined p-values for The Twelve Non-Self-Transitions Studied In This Paper.

Transition	Stouffer's Z	Combined p
ENG_CON	**6.770**	**1.28e-11**
ENG_FRU	-10.878	1.46e-27
ENG_BOR	-12.296	9.40e-35
CON_ENG	-1.605	0.108
CON_FRU	-4.863	1.15e-06
CON_BOR	-7.763	8.25e-15
FRU_ENG	-3.906	9.35e-05
FRU_CON	-2.075	0.037
FRU_BOR	-0.007	0.99
BOR_ENG	-1.344	0.178
BOR_CON	-8.885	6.37e-19
BOR_FRU	-3.861	1.12e-04

The transitions significantly more likely than chance are highlighted in bold. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

2.6.3 Comparison of Datasets from the US and the Philippines

In the next analysis, we investigated if the model is more correct if we restrict the scope of its applicability. In the 12 datasets analyzed in this paper, datasets #1 to #5 were collected in the United States (US), the country where D'Mello's work was conducted, and datasets #6 to #10 were collected in the Philippines. As noted in section 2.2.1, the manifestation of affect may differ in different cultures. Thus, we analyze whether there is a difference in the significance pattern for the two countries, looking in particular at whether one of the

countries conforms better to the theoretical model. Table 2.10 and Table 2.11 present the results of these tests for the US and the Philippines group, respectively.

Table 2.10 Stouffer's Z and Combined p-values for the Data Collected in the United States.

Transition	Stouffer's Z	Combined p
ENG_CON	**18.337**	**4.14e-75**
ENG_FRU	**8.114**	**4.90e-16**
ENG_BOR	-0.511	0.609
CON_ENG	**2.028**	**0.042**
CON_FRU	-2.581	0.009
CON_BOR	-2.183	0.029
FRU_ENG	-1.768	0.077
FRU_CON	0.752	0.452
FRU_BOR	-0.905	0.365
BOR_ENG	**2.110**	**0.034**
BOR_CON	-5.150	2.60e-07
BOR_FRU	0.264	0.791

The transitions significantly more likely than chance are highlighted in bold. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

On combining the p-values from the datasets collected only in the US (Table 2.10), we see that a greater number of transitions are significantly more likely than chance as compared to the results in Table 2.9 (the analysis in both countries). Note that all 4 of these are either from *engaged concentration*, the most frequent state in all datasets (*engaged concentration -> confusion; engaged concentration -> frustration*) or into it (*confusion -> engaged concentration; boredom -> engaged concentration*). Yet only two of these transitions (*engaged concentration -> confusion; confusion -> engaged concentration*) belong to the theoretical model [5].

Table 2.11 Stouffer's Z and Combined p-values for the Data Collected in the Philippines.

Transition	Stouffer's Z	Combined p
ENG_CON	-6.634	3.26e-11
ENG_FRU	-21.100	7.88e-99
ENG_BOR	-15.668	2.46e-55
CON_ENG	-3.816	1.35e-04
CON_FRU	-4.144	3.40e-05
CON_BOR	-8.319	8.80e-17
FRU_ENG	-3.561	3.69e-04
FRU_CON	-2.973	0.0029
FRU_BOR	0.674	0.499
BOR_ENG	-3.543	3.94e-04
BOR_CON	-7.281	3.32e-13
BOR_FRU	-5.279	1.29e-07

The transitions significantly more likely than chance are highlighted in bold. A transition from affect1 to affect2 is denoted as "affect1_affect2." For instance, CON_BOR is a transition from confusion to boredom. Transitions hypothesized in the D'Mello & Graesser's model are highlighted in grey.

In contrast, none of the transitions are significantly more likely than chance when the p-values in the datasets from the Philippines is combined (Table 2.11).

Hence, affect transitions appear to be more stable in the United States than in the Philippines, but neither country shows patterns that conform particularly well to the theoretical model.

2.7 Discussions

D'Mello and Graesser's model [5] has been one of the most influential theoretical frameworks in affect dynamics research. It theorizes how affect develops over time during learning and describes how the transitions in affect that are hypothesized may contribute to processes of learning and disengagement.

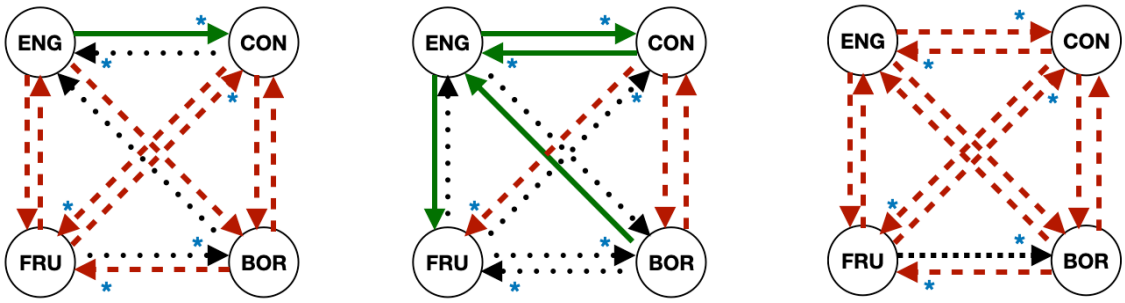


Figure 2.3 Visualization of the significantly likely (green solid arrows), significantly unlikely (red dashed arrows), and null (black dotted arrows) transitions using combine p-values from – (a) all the datastes combined, (b) data collected in the US, and (c) data collected in the Philippines. Transitions hypothesized in the D'Mello and Graesser's model is marked with a blue * next to the arrowhead.

Despite this model's influence on the research community, our survey of the published literature in this area indicates that most of the empirical studies on affect dynamics do not conform to the theoretical model. Even the two empirical studies presented in support of the model in the original paper [5] do not fully support all the hypothesized transitions. Further investigation of the literature reveals that at least some of the differences in the literature may be culturally driven. The studies that do show some evidence for the model were all conducted in the United States with undergraduate populations, but other student populations seem to show more variance in their transition patterns.

In this paper, we reanalyzed and synthesized the data collected in ten publications (twelve datasets) from diverse learning contexts. Our goal was to better understand the pattern of results across these datasets and see what empirical evidence we find for D'Mello and Graesser's theoretical model.

Two methodological concerns drove our decision to reanalyze the data from past publications rather than just synthesizing their results. First, it was not clear from any of

the past articles on affect dynamics how certain edge cases were handled, possibly impacting their results. Second, upon investigating methodological differences between the past studies, we observed that the studies that were showing some evidence for the model (including D'Mello and Graesser's studies) incorporated a pre-processing technique (removing self-transitions) that can sometimes produce spurious false positives.

To address these concerns, we first presented a detailed description of the steps involved in affect dynamics analysis, clarifying the edge cases. We also proposed a new correction to the interpretation of the transition metric for reanalysis. By further investigating which factors are associated with studies matching the predictions in D'Mello and Graesser's model (i.e., studies in different countries), we seek to better understand not just its validity but its scope of applicability.

2.7.1 Non-Conformance to the Theoretical Model

By reanalyzing 12 datasets using D'Mello and Graesser's approach (but with the correction to the transition metric), we show that the data generally does not seem to back up the D'Mello and Graesser [5] model (Figure 2.3). When all data is analyzed using Stouffer's Z, only one (*engaged concentration -> confusion*) of the six hypothesized transitions is more likely than chance, and no other transitions are more likely than chance (Table 2.9). This finding indicates that the differences between D'Mello and Graesser's hypothesized model and past published results are not simply due to differences in the analytical method. At best, we can conclude that this widely-accepted model of affect dynamics has a more limited scope than what it is currently being used for. The future use of this model needs a

thorough exploration of other aspects of design or contexts to understand where it is an accurate depiction of affective processes.

It is worth noting that most of D'Mello and colleagues' data were collected in lab settings, while the other datasets are from real-world classrooms. Is it possible that real-world events cause an affect to shift more rapidly than in the lab? Is it possible that the coarser grain size of BROMP as compared to retroactive affect judgements is part of what gives us a different result? One way to investigate this question in real-world learning is to use affect detectors to get a finer-grained look at affective processes (i.e. [4]), but given that only one study has used this method so far, and that paper appears to have made the same error around the transition metric as seen in several other studies, more work awaits the synthesis of findings using this method.

There has been considerable interest over the last few years in better understanding the dynamics and trends of affect. The primary assumption here is that learners do not randomly shift between emotions and that there are systematic, recurrent shifts between certain states during learning. Our results suggest that affect may instead be generally irregular, raising the question of whether modeling affect dynamics, *in general,* is still fruitful or useful. Our results suggest that it is highly unlikely that there is a general multi-step pattern in affect dynamics like the *engaged concentration -> confusion -> frustration -> boredom* trend suggested by the theoretical model. However, there may still be some contextually relevant patterns useful to understand the student experience. For instance, students in US classrooms (Table 2.10) may oscillate back and forth between *engaged concentration* and other states. But more broadly, perhaps we should be looking at the

affective changes associated with specific events during student learning more than overall trends and patterns.

2.7.2 Methodological Implications for Future Affect Dynamics Research

For around a decade, affect dynamics researchers have used the metric ***L*** to evaluate the probability of transitions in affect. ***L*** is largely believed to have a value of 0 when a transition is at chance, and this is true for the original use of the metric. However, this study provides mathematical evidence that the exclusion of self-transitions leads to a violation of the assumption of independence in the equations used to calculate ***L***. Therefore, this metric does not have a value of 0 at chance if self-transitions are removed.

The primary implication of this finding is in how the ***L*** value is interpreted to understand the direction of a transition. If an affective dynamics study excludes self-transitions, we find that when self-transitions are excluded, the value for ***L*** that represents chance shifts from 0 to $1/(n-1)^2$, where n is the number of affective states studied. Accordingly, the test for the significance of these transitions must be adjusted so that the null hypothesis is set at the appropriate chance levels and not zero.

This finding, thus, has important implications for the interpretation of past publications. For instance, for a study with four affective states, transitions with an ***L*** value less than 0.11 should be interpreted as being less likely than chance. In past studies that excluded self-transitions [4, 5, 21, 23, 37, 38, 50] like the original paper from [5], results must be reinterpreted in terms of the corrected chance value.

In cases where we were unable to analyze the raw data, results need to be reinterpreted on the basis of appropriate chance values for ***L***, given in Table 2.5. For instance, in first study

presented in [5], the transition *confusion -> frustration* is reported to have an $\boldsymbol{L} = 0.060$ and is significant with $p < 0.05$. This is interpreted as a transition that is more likely than chance, but since self-transitions were removed, researchers should apply a corrected chance value to $\boldsymbol{L}$ ($\boldsymbol{L} = 0.11$, as shown for n=4 in Table 2.5) before interpreting this result. This means that the *confusion -> frustration* transition is actually less likely than chance, and the same is true for six of the other ten statistically significant transitions in their two studies.

At the same time, it is important to remind the reader that many past publications using $\boldsymbol{L}$ are unaffected by this concern. Over half of the past studies using this metric included self-transitions [3, 19, 24, 25, 26, 27, 28, 29, 30, 44, 51] and are therefore unchanged by this finding. The choice of whether or not one ought to include self-transitions in an affect dynamics analysis depends on the research goals and questions of the study. Excluding self-transitions reveals a larger number of affective patterns that might otherwise be suppressed by the presence of persistent affective states (although, as our findings indicate, relatively few of these patterns appear to be consistent across studies). Including self-transitions in analysis helps us to better understand each state's persistence, but dilutes the transitions between different affective states. Better understanding transitions is likely important in theoretical models, but understanding persistence might be particularly useful for algorithms being used to trigger interventions, for example.

Recent research has also focused on finding alternative approaches to conducting affect dynamics research, including an update to the L statistic formula [59, 60], and use of epistemic network analysis [61], and marginal models [62].

2.7.3 Need to Focus on Cultural Factors in Affect Dynamics Research

One other important finding in this study is that affective patterns seem to differ based on the country in which the research was conducted (US versus Philippines). Across studies, no affective transitions were more likely than chance in the Philippines, while there were 4 significant transitions in the US (*engaged concentration -> confusion; engaged concentration -> frustration*; *confusion -> engaged concentration; boredom -> engaged concentration*). A similar pattern can be seen in past affect dynamics studies from the Philippines (Table 2.3), which included self-transitions (our analysis excluded them).

Currently, it is not clear why affect dynamic results are so different in the Philippines and the United States. It is not that the most common affect differs – this seems to be relatively consistent across studies, many of which started with additional affective states. It is not that the relationship between affect and learning differs – the negative correlation between boredom and learning and the positive correlation between *engaged concentration* and learning are seen in both countries (along with the instability of correlation within each country for confusion and frustration) (i.e., 43, 52, 63, 64). Despite these commonalities, the affect dynamics seem to differ.

Given the many differences between schools in the United States and the Philippines – national culture, school culture, use of educational technology, prevalent forms of disengagement [65] – it is difficult at this point to understand *why* we see these differences. It may even be the case that the affect being recognized in different countries is fundamentally different in kind, in a way that the researchers conducting these studies cannot fully recognize. The BROMP field observation protocol was co-developed by

researchers in the USA and Phillipines and has now been applied in several other countries, but that does not guarantee that the same constructs are captured when a researcher in each country identifies “engaged concentration” or “frustration”. Indeed, many individuals have found it difficult to achieve acceptable inter-rater reliability out of their native culture [44], and there are systematic biases in cross-cultural attempts to recognize affect [64]. A better understanding of the role that culture plays in the manifestation and recognition of affect is important for any future attempts to study affect dynamics as a generalizable phenomenon. Ultimately, it may make the best sense to re-consider that question after affect dynamics has been studied in a wider range of cultures, potentially looking at the kinds of traits that are known to vary at a national level.

2.7.4 Limitations and Future Work

There are some limitations to this paper’s findings. First, all but one dataset among the ten studies are collected through quantitative field observations, which may sample at slower rates than many other approaches to collect affect data, like video, sensors, emote-aloud methods, and self-reports. This choice was driven primarily by the previous papers that investigate the phenomena of interest, as well as the availability of datasets. However, other methods have different virtues and flaws in terms of cost, scale, accuracy, and time. It would be interesting for future work to explore how each of these methods captures student emotions differently and how these differences impact the validity or applicability of the affect dynamics analysis.

Second, this analysis did not consider the impact of other attributes of the study design like observation grain size and the length of observation session. As we note above, finer-

grained affect observations could potentially yield a different result than the coarser-grained data from BROMP observations. Moreover, it seems likely that students may be more likely to hit points of frustration and boredom later rather than earlier in a long observation system, particularly if they were working within a learning system that became increasingly challenging over time.

Third, all our significant transitions have *engaged concentration* in them. *Engaged concentration* also happens to be the most frequent affective state in all our datasets. In contrast, *frustration* is the rarest affective state and has the most null or negative results. Thus, future work needs to analyze if there exists a threshold for the minimum length of the affect sequence or minimum base rate of each affective state to be able to see significant positive transitions. Alternatively, it may be worth explicitly seeking out more difficult tasks and contexts where frustration may be more common, such as learning systems that are known to be less effective or students working alone at home within fully asynchronous virtual schooling. It is also possible that studies have seen less frustration because most are conducted in short-term studies rather than ongoing use; studies like [38] that involve an entire year of usage may be more able to detect affective sequences associated with frustration.

Fourth, this work synthesizes across multiple affect datasets. Like the studies it builds on, it does not consider individual differences in affect incidence and dynamics that may appear in the data. It is possible that different patterns – both related and unrelated to the theoretical model – may be characteristic of sub-groups. Differences in affective dynamics

may be associated with a variety of individual differences, such as differences in personality.

Overall, this paper provides a comprehensive look at affect dynamics across published work. Broadly, work so far does not seem to accord with the most popular theoretical model. Further work is needed to understand what is general in the dynamics of affect, both for specific contexts and across contexts.

BIBLIOGRAPHY

[1] McQuiggan, S. W., & Lester, J. (2009). Modeling affect expression and recognition in an interactive learning environment. *International Journal of Learning Technology, 4*(3-4), 216-233.

[2] D'Mello, S., Person, N., Lehman, B. (2009). Antecedent-Consequent Relationships and Cyclical Patterns between Affective States and Problem Solving Outcomes. In *AIED*, 57-64.

[3] Rodrigo, M.M.T., Anglo, E., Sugay, J., Baker, R. (2008). Use of un-supervised clustering to characterize learner behaviors and affective states while using an intelligent tutoring system. In *International Con-ference on Computers in Education*, 57-64.

[4] Bosch, N., & D'Mello, S. (2017). The Affective Experience of Novice Computer Programmers. *International Journal of Artificial Intelligence in Education*, 1-26.

[5] D'Mello, S. Graesser, A. (2012). Dynamics of Affective States during Complex Learning. *Learning and Instruction,* 22, 145-157.

[6] Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009, July). Emotion sensors go to school. In *AIED* (Vol. 200, pp. 17-24).

[7] Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014, June). Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International conference on intelligent tutoring systems* (pp. 29-38). Springer, Cham.

[8] D'Mello, S., & Kory, J. (2012, October). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 31-38).

[9] Nye, B. D., Karumbaiah, S., Tokel, S. T., Core, M. G., Stratou, G., Auerbach, D., & Georgila, K. (2018, June). Engaging with the scenario: Affect and facial patterns from a scenario-based intelligent tutoring system. In *International Conference on Artificial Intelligence in Education* (pp. 352-366). Springer, Cham.

[10] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis* (pp. 1-10). Springer, Cham.

[11] Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020, October). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 105-114).

[12] DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C. (2018) Detecting and Addressing Frustration in a Serious Game for Military Training. *International Journal of Artificial Intelligence and Education, 28* (2), 152-193.

[13] Karumbaiah, S., Lan, A., Nagpal, S., Baker, R. S., Botelho, A., & Heffernan, N. (2021, April). Using Past Data to Warm Start Active Machine Learning: Does Context Matter?. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 151-160).

[14] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011, March). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *In Proceedings of the 6th International Conference on Human-robot Interaction* (pp. 305-312). ACM.

[15] Clavel, C., & Callejas, Z. (2015). Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing, 7*(1), 74-93.

[16] DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). De-tecting and addressing frustration in a serious game for military train-ing. *International Journal of Artificial Intelligence in Education, 28* (2), 152-193.

[17] Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B. P., Arroyo, I., & Wixon, N. (2017). Addressing Student Behavior and Affect with Empathy and Growth Mindset. *Proceedings of the 10th International Conference on Educational Data Mining*.

[18] Kuppens, P. (2015). It's about time: A special section on affect dynamics. *Emotion Rev., 7*(4), 297-300.

[19] Andres, J.M.L., & Rodrigo, M.M.T. (2014). The Incidence and persis-tence of affective states while playing Newton's playground. *7th IEEE International Conference on Humanoid, Nanotechnology, Information Tech., Communication and Control, Environment, and Management.*

[20] Baker, R.S., Rodrigo, M.M.T., Xolocotzin, U. (2007). The dynamics of affective transitions in simulation problem-solving environments. *International Conf. on Affective Computing and Intelligent Interaction. Springer Berlin Heidelberg,* 666-67.

[21] Bosch, N., & D'Mello, S. (2013). Sequential patterns of affective states of novice programmers. In *The 1st Workshop on AI-supported Education for Computer Science (AIEDCS 2013),* 1-10.

[22] D'Mello, S., Graesser, A. (2015). Feeling, thinking, & computing with affect-aware learning technologies. Calvo, D'Mello, Gratch, Kappas (eds.) *Handbook of Affective Computing*. Oxford UP.

[23] D'Mello, S., Taylor, R., & Graesser, A. (2007). Monitoring Affective Trajectories during Complex Learning. D. McNamara & J. Trafton (Eds.), *Proc. 29th Annual Cognitive Science Soc.*, 203-8.

[24] Guia, T.F.G., Rodrigo, M.M.T., Dagami, M., Sugay, J., Macam, F., Mitrovic, A. (2013) An exploratory study of factors indicative of af-fective states of students using SQL-Tutor. *Research & Practice in Technology Enhanced Learning, 8*(3), 411-430.

[25] Guia, T.F.G., Sugay, J., Rodrigo, M.M.T., Macam, F., Dagami, M., Mitrovic, A. (2011). Transitions of Affective States in an Intelligent Tutoring System. *Proc. Philippine Computing Soc.* 31-5.

[26] McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective transitions in narrativecentered learning environments. *Educational Technology & Society, 13*(1), 40-53.

[27] Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education, 13(*2-4), 173-197.

[28] Ocumpaugh, J., Andres, J.M., Baker, R., DeFalco, J., Paquette, L., Rowe, J., et al. (2017). Affect Dynamics in Military Trainees using vMedic: From Engaged Concentration to Boredom to Confusion. In *International Conf. on Artificial Intelligence in Ed*., 238-249. Springer, Cham.

[29] Rodrigo, M.M.T., Baker, R., Agapito, J., Nabos, J., Repalam, M., Reyes Jr, S., San Pedro, M.O.C. (2011). The effects of an embodied conversational agent on student affective dynamics while using an in-telligent tutoring system. *IEEE Transactions on Affective Computing, 2*(4), 18-37.

[30] Rodrigo, M.M.T., Baker, R., Agapito, J., Nabos, J., Repalam, M., Reyes, S., San Pedro, M.O.C. (2012). The effects of an interactive software agent on student affective dynamics while using; an intelli-gent tutoring system. *IEEE Transactions on Affective Computing, 3*(2), 224-36.

[31] Schwarz, N. (2012). Feelings-as-Information Theory. In P. Van Lange, A. Kruglanski & T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (pp. 289-308). Thousand Oaks, CA: Sage.

[32] Izard, C. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review, 2*(4), 363-370.

[33] Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion, 24*(7), 1259-1268.

[34] Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought- action repertoires. *Cognition & Emotion, 19*(3), 313-332.

[35] Csikszentmihalyi, M. (1990). *Flow and the psychology of discovery and invention*. Harper Collins, New York.

[36] Baker, R.S., D'Mello, S., Rodrigo, M.M.T., Graesser, A. (2010). Better to be frustrated than bored: The incidence, persistence, & impact of learners' cognitive-affective states during interactions with 3 different computer-based learning environments. *Int'l J. Hum-Comp. Stu., 68* (4),223-41.

[37] D'Mello, S., Graesser, A. (2010). Modeling cognitive-affective dy-namics with Hidden Markov Models. *Proceedings of the 32nd Annual Cognitive Science Society*, 2721-2726.

[38] Botelho, A.F., Baker, R., Ocumpaugh, J., Heffernan, N. (2018) Study-ing Affect Dynamics and Chronometry Using Sensor-Free Detectors. *Proceedings of the 11th International Conference on Educational Data Mining*, 157-166.

[39] Tsai, J., Levenson, R. (1997). Cultural influences on emotional re-sponding: Chinese Am. & European Am. dating couples during inter-personal conflict. *J. Cross-Cultural Psych., 28*(5), 600-25.

[40] Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, emo-tion, and well-being: Good feelings in Japan and the United States. *Cognition & Emotion, 14*(1), 93-124.

[41] Dunn, J., & Brown, J. (1994). Affect expression in the family, chil-dren's understanding of emotions, and their interactions with others. *Merrill-Palmer Quarterly* (1982-), 120-137.

[42] Gross, J. J., Carstensen, L. L., Pasupathi, M., Tsai, J., Götestam Skorpen, C., & Hsu, A. Y. (1997). Emotion and aging: Experience, expression, and control. *Psychology and Aging, 12*(4), 590.

[43] Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29*(3), 241-250.

[44] Baker, R.S., Ocumpaugh, J.L., Andres, J.M.A.L. (in press) BROMP Quantitative Field Observations: A Review. In R. Feldman (Ed.) Learning Science*: Theory, Research, and Practice*. New York, NY: McGraw-Hill.

[45] Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical & Training Manual.* Technical Report. NY, NY: Teachers College, Columbia U. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

[46] Gonzalez, C., Best, B., Healy, A., Kole, J. A., Bourne Jr, L. (2011). A cognitive modeling account of simultaneous learning and fatigue ef-fects. *Cognitive Systems Research, 12*(1), 19-32.

[47] Healy, A., Kole, J., Buck-Gengler, C., Bourne Jr, L., (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Ex-perimental Psychology: Applied,* 10, 188–199.

[48] Karumbaiah, S., Andres, J. M. A. L., Botelho, A. F., Baker, R. S., & Ocumpaugh, J. S. (2018). The Implications of a Subtle Difference in the Calculation of Affect Dynamics*. In 26th International Conference for Computers in Education.*

[49] Andres, J.M.L., Rodrigo, M.M T., Sugay, J., Banawan, M., Paredes, Y., Cruz, J., Palaoag, T. (2015). More Fun in the Philippines? Factors Affecting Transfer of Western Field Methods to One Developing World Context. In *AIED Workshops*.

[50] Karumbaiah, S., Baker, R. S., & Ocumpaugh, J. (2019, June). The Case of Self-transitions in Affective Dynamics. *In International Con-ference on Artificial Intelligence in Education* (pp. 172-181). Springer, Cham.

[51] Rodrigo, M.M.T., Rebolledo-Mendez, G., Baker, R., du Boulay, B., Sugay, J.., Lim, S., Luckin, R. (2008). The effects of motivational modeling on affect in an intelligent tutoring system. *Proc. of Interna-tional Conference on Computers in Education*, 57, 64.

[52] Rodrigo, M. M. T., & Baker, R. S. (2009, August). Coarse-grained detection of student frustration in an introductory programming course. In *Proceedings of the 5th International Workshop on Computing Education Research Workshop* (pp. 75-80). ACM.

[53] Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction*, 44, 128-143.

[54] DiStefano, D. (2018*). How Pre-Service Teachers' Engagement and Affect Informs Instructional Format of an Introductory Methods Course* (Doctoral dissertation, Fordham University).

[55] Nicaud, J. F., Bouhineau, D., Mezerette, S., & Andre, N. (2007). *Aplusix II* [Computer software].

[56] Shute, V., Ventura, M. (2013). Stealth assessment: Measuring & supporting learning in video games. *MIT Press*

[57] Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. Jr. (1949). *The American Soldier, Vol. 1: Adjustment During Army Life.* Princeton University Press, Princeton.

[58] Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral re-search: Methods and data analysis* (Vol. 2). New York: McGraw-Hill.split

[59] Bosch, N & Paquette, L. (2020). What's Next? Edge Cases in Measuring Transitions Between Sequential States. Submitted

[60] Matayoshi, J., & Karumbaiah, S. (2020). Adjusting the L Statistic when Self-Transitions are Excluded in Affect Dynamics. *Journal of Educational Data Mining*, *12*(4), 1-23.

[61] Karumbaiah, S., & Baker, R. S. (2021, February). Studying Affect Dynamics using Epistemic Networks. In *International Conference on Quantitative Ethnography* (pp. 362-374). Springer, Cham.

[62] Matayoshi, J., & Karumbaiah, S. (2021, April). Using Marginal Models to Adjust for Statistical Bias in the Analysis of State Transitions. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 449-455).

[63] Lagud, M. C. V., & Rodrigo, M. M. T. (2010, June). The affective and learning profiles of stu-dents using an intelligent tutoring system for algebra. In *International Conference on Intelligent Tutoring Systems* (pp. 255-263). Springer, Berlin, Heidelberg.

[64] Okur, E., Aslan, S., Alyuz, N., Esme, A.A., Baker, R.S. (2018) Role of Socio-Cultural Differences in Labeling Students' Affective States. Proceedings of the *19th International Conference on Artificial Intelligence in Education*, 367-380.

[65] Rodrigo, M. M. T., Baker, R. S. J. D., & Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: comparison to prior research in the USA. *Teachers College Record, 115*(10), 1-27.

CHAPTER 3

DIFFERING IMPLICATIONS OF TECHNOLOGY DESIGN ON STUDENT OUTCOMES

Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2021). Context Matters: Differing Implications of Motivation and Help-Seeking in Educational Technology. *International Journal of Artificial Intelligence in Education*.

Abstract. Educational Technology (EdTech) designers need to ensure population validity as they attempt to meet the individual needs of all students. EdTech researchers often have access to larger and more diverse samples of student data to test replication across broad demographic contexts as compared to either the small-scale experiments or the larger convenience samples often seen in experimental psychology studies of learning. However, the source of typical EdTech data (i.e., online learning systems) and concerns related to student privacy often limit the opportunity to collect demographic variables from individual students – the sample is diverse, but the researcher does not know how that diversity is realized in individual learners. In order to ensure equitable student outcomes, the EdTech community should make greater efforts to develop new methods for addressing this shortcoming. Recent work has sought to address this issue by investigating publicly-available, school-level differences in demographics, which can be useful when individual-level variation may be difficult or impossible to acquire data for. In this study, we use this approach to better understand the role of social factors in students' self-regulated learning and motivation-related behaviors, behaviors whose effectiveness appears to be highly

variable between groups. We demonstrate that school-level demographics can be significantly associated with the relationships between students' help-seeking behavior, motivation, and outcomes (math performance and math self-concept). We do so in the context of Reasoning Mind, an intelligent tutoring system for elementary mathematics. By studying the conditions under which these relationships vary across different demographic contexts, we challenge implicit assumptions of generalizability and provide an evidence-based commentary on future research practices in the EdTech community surrounding how we consider diversity in our field's investigations.

3.1. Introduction

As Educational Technology (EdTech) researchers and designers seek to support productive learning behaviors, they are faced with a challenge. Complex constructs like motivation, interest, and engagement are known to influence a variety of learning behaviors (Ryan & Deci, 2000; Renninger et al., 2018). However, due to practical constraints of research projects (e.g., budget, recruitment, accessibility, and time), many of these studies involve either small-scale experiments or larger convenience samples of middle-class, undergraduate students (see discussion in Kimble, 1987), which can make it difficult to determine the extent to which these findings will generalize to new and diverse populations of students. Consequently, decisions inspired by such studies could lead to inequitable outcomes for students, if findings are inapplicable for key groups of learners who are not studied. Even when EdTech researchers obtain larger sample sizes, the typical source of EdTech data (i.e., intelligent tutoring systems, adaptive learning platforms, educational

game websites) often limits the practicality of obtaining demographic variables from individual students. Beyond practicality (e.g., the ease of acquiring log data on student interactions compared to student demographic data), concerns such as student privacy can reduce the collection of demographic data. For example, even when a partner school or university has documented the demographics of individual students, their release to a researcher increases the risk of potentially re-identifying students, particularly in rural parts of the country where the analysis of (for instance) the seven children of a minority ethnic group in a small school narrows the potential matches for sensitive information considerably. Yet considerable research shows that demographic factors are often related to differences in educational outcomes more generally (see Childs, 2017) and to constructs related to motivation more specifically (Usher & Pajares, 2006; Zeldin & Pajares, 2000; Zeldin et al., 2008).

Adaptive EdTech with automated decision-making attempts to meet individual student needs, but past research points out that technologies that aim to benefit all students might disproportionately benefit the more advantaged groups. Despite some examples of the success of adaptive EdTech technologies for historically underrepresented groups (i.e., Koedinger et al., 1997; Finkelstein et al., 2013; Huang et al., 2016; Roschelle et al., 2016), learning technologies have not had overall success in closing society's achievement and opportunity gaps (Hansen & Reich, 2015). Despite removing technical and economic barriers (Attewell, 2001), the social and cultural barriers contributing to inequity remain challenging. Institutionalized and unconscious bias and social and cultural distance

between EdTech designers and those they seek to serve (especially low-income and minority groups) are the two common sources of failure for the equitable deployment of new technologies (Reich & Ito, 2017). Technology developers' lack of awareness of sociocultural contexts and the needs of different student subgroups can lead to unfortunate consequences. Reich and Ito (2017) emphasize that measuring differences in how various subgroups experience and benefit from EdTech will be a crucial component of our deepening understanding of EdTech and in addressing the inequalities that emerge. At the same time, they acknowledge that there are substantial barriers to collecting relevant demographic data for individual students. Without studies to ensure population validity (Ocumpaugh et al., 2014), such systems may have to compromise on how effectively they can individualize when they are used in unexpected ways by diverse learners (Doroudi, 2019) and in diverse settings. Thus, it seems important that EdTech designers and researchers make greater efforts to overcome the challenges involved in collecting demographic data in order to ensure population validity, equity-oriented EdTech design, and fairness-aware educational data mining (e.g., Ocumpaugh et al., 2014). As such, some researchers have sought to extend student learning models to include information from the broader context, building models at the class, school, or school-cluster level instead of just the student-level (Wang & Beck, 2013; Pardos & Heffernan, 2010, Yudelson et al., 2014).

Our broader research goal is to investigate if the current designs of adaptive EdTech lead to inequitable student outcomes across different demographics. Within this paper, we incorporate broader demographic contexts into an investigation of help-seeking (where a

student deliberately asks for assistance in trying to complete or understand a problem) and motivational constructs within EdTech. Help-seeking is chosen as a phenomenon for investigation based on a recent review of the help-seeking literature, which found that the effectiveness of help-seeking behaviors was highly variable across studies (Aleven et al., 2016). Most of the previous research on help-seeking in EdTech has typically used a cognitive lens. Thus, we attempt to shift the focus to also consider social factors. We conduct our investigations in the context of Reasoning Mind, an Intelligent Tutoring System for elementary mathematics. Specifically, we demonstrate how readily-available, school-level demographics might reveal how help-seeking and other motivational behaviors of students correlate with two student outcome measures: (1) mathematics performance and (2) mathematics self-concept (an affective measure of students' perception of their own cognitive ability, which is known to predict performance; Lee, 2009). This work has a direct implication to EdTech designers and researchers, who often rely on features such as the universal design of hints and gamification informed by small-scale experiments or larger convenience samples, ignoring group differences.

3.2. Prior Work

This section will review the relevant prior work on help-seeking and motivation and the role they play in EdTech design. We also review the past research on the influence of student demographics on help-seeking, motivation, and the two outcome variables – math performance and self-concept.

Previous research on student help-seeking in ITSs has often taken a cognitive approach, focusing on understanding the cognition behind students' choices around help-seeking and the relationship between different forms of help-seeking and student outcomes (Aleven et al., 2016). However, this research has often obtained conflicting findings, including contradictory (positive and negative) correlations between hint usage and learning (Koedinger & Aleven, 2007). Though accounts of these findings have often focused on *how* or *when* students seek help (e.g., Koedinger & Aleven, 2007), we believe that the conflicting findings may also relate to *who* chooses to seek help. In this section, we review prior work on help-seeking, comparing the findings in the ITS research that has typically taken a cognitive approach to research that has explored the social and demographic factors related to help-seeking. We then look at previous research on intrinsic and extrinsic motivation, especially as it relates to social and demographic differences in education, based on evidence that help-seeking is associated with the student's motivation (Nelson-Le Gall & Resnick, 1998; Butler, 2006). Finally, given the relationships between mathematics self-concept and help-seeking (Skaalvik & Skaalvik, 2013), we discuss prior research related to the role that demographics play in math self-concept and math performance, which allow us to better understand how students' behaviors are related both to their actual skill level and their perceptions about that ability.

Given that previous research shows social differences in help-seeking outside of ITS systems and in other constructs related to motivation, we hypothesize that we should expect that help-seeking and motivational behaviors may demonstrate demographic differences.

These differences may account for the sometimes contradictory findings that cognitive research has shown when comparing help-seeking practices to student performance.

3.2.1 Help-Seeking

Help-seeking functions - mostly in the form of on-demand, contextual, real-time hints - are common features in most Intelligent Tutoring Systems (ITSs; VanLehn, 2006), and they have long been believed to foster emerging concepts and principles in a student's learning (Anderson, 1993) and to support struggling students during problem-solving (Aleven & Koedinger, 2000). Yet help-seeking behaviors are not always beneficial (Aleven & Koedinger (2000, 2001); Aleven et al., 2016). While much of the prior work on help-seeking in ITSs has focused strictly on its cognitive effects, other research suggests that we should be exploring how motivational social factors may influence these findings, as these patterns may help us to better understand the social issues that govern when and how students choose to engage with help-seeking opportunities.

3.2.1.1 Help-Seeking: A Cognitive Lens

The literature on help-seeking behaviors in ITSs now stretches back over two decades (see extensive review in Aleven et al., 2016). As it quickly became apparent that the availability of hints did not ensure their effective use, work began to identify the factors that led to a positive relationship between help-seeking behaviors and student learning. In one of the earliest studies, Anderson et al. (1989) compared the use of explanatory hints and so-called bottom-out hints (which simply provided the student with the correct answer) and found

that neither hint type was correlated with learning. In part, this may have been due to selection bias. That is, hint usage is typically a sign of struggling students, who often do not make substantial learning gains (see discussion in Aleven et al., 2016).

After early findings showed a negative correlation between hint usage and student learning in one context (Aleven & Koedinger, 2000), researchers began to develop a taxonomy of maladaptive help-seeking behaviors—including categories like *help abuse* (the overuse of help) and *help avoidance* (the underuse of help) (Aleven et al., 2006). Most studies analyzed the effectiveness of hints by focusing on the relationship between help-seeking behavior(s) and student outcome(s), with some researchers emphasizing that the intentionality of help-seeking behavior makes it a good candidate for understanding students' self-regulated learning (SRL) strategies (Aleven et al., 2016; Goldin et al., 2012).

A number of studies have attempted to identify the degree of help needed at any given moment (e.g., Koedinger & Aleven's (2007) assistance dilemma). These studies have shown several interesting findings. For example, (1) on-demand hints lead to greater learning gains than automatic hints in middle-school mathematics (Razzaq & Heffernan, 2010); (2) hint content (goal feedback versus other kinds of feedback) is related to student learning in Geometry (McKendree, 1990); (3) hints about which step to try next to improve student learning of logic proofs (Stamper et al., 2011).

In general, the literature suggests that increasing hint usage does not always lead to better domain-level learning (Aleven et al., 2016). However, the literature on help-seeking in ITSs has produced research that aggregates into a complicated and contradictory narrative,

including: (1) a negative association between hint usage and learning (Aleven & Koedinger, 2001); (2) a positive association between hint usage and learning (Beck et al., 2008; Wood & Wood, 1999); (3) a positive association between hint usage and learning only when time per hint level is considered (Long & Aleven, 2013); (4) a positive association between time spent in bottom-out hints and learning (Shih et al., 2008); (5) a negative association between the number of bottom-out hints used and learning (Mathews et al., 2008); (6) positive benefits for students but only when they have a medium level of skill (Roll et al., 2014); (7) a negative association between help avoidance and learning early within practice (Almeda et al., 2017) and on a transfer post-test (Baker et al., 2011). In addition, individual differences in self-regulation were observed in how students process hints and how that impacts their performance (Goldin et al., 2012). Vaessen et al. (2014) found that students' achievement goals (mastery and performance goals) are closely related to their help-seeking and could be used to predict their strategies for help-seeking. Overall, despite a considerable volume of research, the effectiveness of help-seeking remains an open question—and the clearest thing that we can say is that the relationship between hint usage and learning is complicated.

3.2.1.2 Help-Seeking: A Social Lens

While the role of social factors on help-seeking behaviors has not been the primary focus of the EdTech community (see Aleven et al., 2016), the social evaluation of help-seeking behaviors is well established in the literature. For instance, some learners may feel that asking for help is either a sign of incompetence or a challenge to their autonomy (Tessler

& Schwartz, 1972). Relatedly, Howley et al. (2014) suggest that asking for help (within in-person learning) may trigger experiences of evaluation anxiety – the fear of being judged. In fact, early work on student help-seeking sometimes focused on its maladaptive uses (Baltes, 1997), a categorization that suggests that some students might avoid help after accurately assessing classroom expectations of independence in their learning processes. Meanwhile, Butler (1998) identified three factors related to help-seeking behaviors, including the desire to work autonomously, the desire to demonstrate high ability, and the desire to finish the assignment quickly.

These kinds of concerns seem ripe for sociocultural variation, and a few studies have begun to explore how these differences may emerge. For example, Tai et al. (2013) increased students' help-seeking behaviors by changing the way they labeled those actions within the system. That is, they started by referring to the ITS as the students' teammate, and they designed the system so that students who needed help could choose to "work together" with the system. This adaption apparently reduced the ego-threat related to admitting a lack of knowledge (e.g., Tessler & Schwartz, 1972) and improved student learning.

3.2.2.3 Student Demographics and Help-Seeking Behaviors

When social expectations guide behaviors, researchers should expect to find demographic differences, and some studies have specifically investigated this with respect to help-seeking behaviors. For example, Ogan et al. (2015) found that the models on effective help-seeking did not transfer well between countries (namely Costa Rica, the Philippines, and the USA). Likewise, Arroyo et al. (2000) found that the effectiveness of different hint

designs varied by gender. Specifically, girls benefited more from highly interactive hints, while boys did better with less interactive hints. This work matches findings in other learning contexts, which has shown both that there may be racial and gendered interactions influencing differences in help-seeking behaviors and that these different behaviors may explain subsequent achievement patterns (Ryan et al., 2009). Combined, these findings suggest that researchers in the ITS community should be paying attention to cultural differences that may influence how students perceive help-seeking opportunities to affect their sense of *competence* and *autonomy*. That is, if we are going to design ways to improve appropriate help-seeking behaviors, we have to understand which students are currently reluctant to use these behaviors.

3.2.2 Intrinsic and Extrinsic Motivation

While help-seeking has been studied in terms of either cognitive or social factors, student motivation tends to be classified into either intrinsic or extrinsic motivational factors. As described in Deci and Ryan's Self Determination Theory (SDT; 1985), "intrinsic motivation refers to doing something because it is inherently interesting or enjoyable and extrinsic motivation refers to doing something because it leads to a separable outcome." There is a general consensus on the role of intrinsic motivation in high-quality learning and creativity, reflecting natural human propensities to learn (Ryan & Deci, 2000). However, the importance of extrinsic motivation is argued to be dependent on autonomy as experienced by the student. Thus, a student's extrinsic motivation could reflect either true self-regulation or external control. Since it is difficult to expect students to be intrinsically

motivated by all subject matter or to inherently enjoy all learning activities, educators and EdTech designers often rely on extrinsic motivators. However, passive and controlling forms of extrinsic motivation can leave students only externally propelled into action (e.g., with the expectation of being tested on it). In contrast, more active and volitional forms of extrinsic motivation can win students' acceptance (e.g., with the expectation of teaching it to a peer; see review in Ryan & Deci, 2000).

3.2.2.1 Social Factors Influencing Intrinsic Motivation

Researches examining the role of autonomy in intrinsic motivation suggest that immediate contextual conditions (e.g., those found in students' schools and homes) can systemically catalyze or undermine the needs of competence and autonomy (Ryan & Stiller, 1991). Cognitive Evaluation Theory (CET) - a sub-theory of SDT – specifies social factors that lead to differences in intrinsic motivation. It argues that interpersonal events and structures that are conducive to feelings of competence and autonomy can elicit, sustain, or enhance intrinsic motivation for the action performed. Examples of such structures include optimal challenge, constructive feedback (Harackiewicz, 1979), and the absence of shaming evaluations (Deci & Cascio, 1972). Along with the increases in perceived competence (Vallerand & Reid, 1984), students must experience their behavior to be self-determined for intrinsic motivation to increase (Ryan, 1982).

In fact, the issue of autonomy versus control has been a popular field of motivation research with considerable controversy. Lepper et al. (1973) first reported that extrinsic rewards could undermine intrinsic motivation. A later meta-analysis argued that any type of

expected tangible reward made contingent on task performance undermines intrinsic motivation by shifting the perceived locus of causality from internal to external (Deci et al., 1999). On the other hand, a parallel school of thought has argued against prematurely dismissing the value of tangible extrinsic rewards for students who are not intrinsically motivated (Hidi & Harackiewicz, 2000). Other structures that have been reported to have a negative outcome on intrinsic motivation include deadlines (Amabile et al., 1976), directives (Koestner et al., 1984), and competition pressure (Reeve & Deci, 1996). Students who were overly controlled also learned less and lost their initiative to learn (Grolnick & Ryan, 1987). In contrast, choice and opportunity to engage in self-direction (Zuckerman et al., 1978) were reported to enhance intrinsic motivation. A similar positive effect on intrinsic motivation, curiosity, and desire for challenge was reported with autonomy-supportive teacher practices (Ryan & Grolnick, 1986).

3.2.2.2 Social Factors Influencing Extrinsic Motivation

An important challenge for educators and EdTech designers is to design activities that, when not intrinsically interesting, could still motivate students to value and self-regulate on their own without external pressure (Zimmerman, 1985). Organismic Integration Theory (OIT) - another sub-theory of SDT – emphasizes the role of student autonomy in designing activities and experiences that improve extrinsic motivation (Ryan & Connell, 1989). Specifically, more autonomous extrinsic motivation is associated with greater engagement (Connell & Wellborn, 1990), better performance (Miserandino, 1996), higher

quality learning (Grolnick & Ryan, 1987), and greater psychological well-being (Sheldon & Kasser, 1995).

As with intrinsic motivation, social-contextual conditions that foster a students' feeling of competence and autonomy support self-regulation with extrinsic motivation. In addition, since extrinsically motivated behaviors do not reflect inherent interest, their value to the people, group, or culture whom the student identifies with becomes important (Ryan & Deci, 2000). For example, students' relatedness to teachers in their classrooms, along with their sense of being valued by their teacher, is strongly linked to their adoption of classroom values (Ryan, Stiller, & Lynch, 1994). Similar findings are reported for the importance of autonomy, relatedness, competence supportive practices in extrinsically valued activities (Grolnick & Ryan, 1987; Williams & Deci, 1996).

3.2.2.3 Student Demographics Influencing Motivation

Demographic differences in the development of students' motivational profiles and a corresponding need for different supports are noted in several studies (Renninger et al., 2018). One notable finding includes a significant linear decrease in intrinsic motivation from 3rd grade through 8th, while extrinsic motivation showed few differences across grade levels (Lepper et al., 2005), a finding which has been replicated in other studies (Gottfried, Fleming, & Gottfried, 2001). Similarly, gender has been shown to influence the relationship between students' motivation and the topic or context of the learning task (Hoffmann and Häussler, 1998). Other demographic categories, including those that are more clearly sociocultural (as opposed to biological) in nature, have also proved important

to motivation research and interventions. For example, both underrepresented students and first-generation students were positively influenced by interventions involving reflections on utility value or relevance, resulting in increased interest in the subject matter (Hulleman, Kosovich, Barron, & Daniel, 2016).

Thus, there is a need for more research to look at social factors while studying student motivation and help-seeking in EdTech. Such studies should consider student demographics to understand how to foster positive student outcomes. In this paper, we study students' help-seeking behavior and their intrinsically and extrinsically motivated behaviors in an online math tutor used in traditional classrooms during regular instruction. We aim to shift the focus of EdTech research for constructs like help-seeking from purely cognitive factors to the contextual factors that might play a more prominent role than is assumed. We focus on school as the social context and analyze the influence of school demographics on the relationship between student outcomes (math performance and math self-concept) and their help-seeking and motivational behaviors.

3.2.3 The Role of Demographics in Predicting Student Outcomes

Two outcome measures are used in this study: math performance and math self-concept. The first helps us to better understand how much help a student might need, while the second helps us to better understand whether how confident they are in their own ability (which may be more linked to help-seeking choices than actual proficiency). This section summarizes prior work on the role of demographics in the student outcomes of interest in this study - math performance and math self-concept.

3.2.3.1 Demographics and Math Performance

The literature addressing demographic differences in learning outcomes is now so vast that it would be difficult to review even if it were limited to a single domain (e.g., mathematics). Once referred to as the achievement gap, many scholars are now instead discussing an opportunity gap, as findings generally show that achievement patterns favor groups for whom the educational system was initially designed (see discussion in Chambers, 2009; Flores, 2007). Scholars point out that reframing this discussion in terms of opportunities to learn emphasizes the need to address the environmental inadequacies that are driving inequitable outcomes (Flores, 2007; Ladson-Billings, 2013). Childs' (2017) analysis shows, for example, that minority students are just as likely to value mathematics as other students but are less likely to attend schools where advanced mathematics classes are offered.

However, less tangible cultural and linguistic differences may also play a role. We know, for example, that the strategies for speech act like asking questions can vary substantially even in the same language. (See, for example, Greenbaum & Greenbaum's (1982) review of classroom practices among different Native American groups or Chavajay and Rogoff's (2002) review of the literature on classroom practices among cultures that do not use known-answer questions.) If students' patterns of communication are different from those expected by educators, their attempts at communication—including help-seeking—may not receive appropriate responses (Hudley & Mallinson, 2015). Such experiences could discourage students from future help-seeking behaviors, although one could imagine that

the ability to get help from an ITS could also mitigate this reluctance if the help-seeking system were appropriately designed.

3.2.3.2 Demographics and Self-Concept

Demographic variables have also been shown to correlate with constructs like math self-concept (self-beliefs related to a specific task; Bandura, 1982). Math self-concept (sometimes used interchangeably with self-efficacy, although see Bong & Skaalvik (2003) for discussion) has been found to be a predictor of various measures of achievement and career choice (see Brown & Lent, 2006), although the relationship between self-concept and mathematics achievement does vary in magnitude in different countries (Wilkins, 2004). It has also been linked to motivational constructs, including achievement goal orientation, anxiety, and self-concept (Schunk & Pajares, 2005).

Early work proposed that self-efficacy was a product of a person's own accomplishments and the feedback they receive on their work (Bandura, 1982; Urdan & Pajares, 2006); however, more recent studies have indicated that the source of self-efficacy may vary along demographic lines like gender and ethnicity (Zeldin & Pajares, 2000; Zeldin et al., 2008; Usher & Pajares, 2006). For example, Klassen's (2004) investigation of self-efficacy among seventh-grade students found that ethnic majority students followed Bandura's (1982) predictions, citing personal achievements as a source of self-efficacy, but ethnic minority students were more likely to cite group capabilities for collective efficacy. By contrast, Else-Quest et al. (2013) studied the intersection of gender, ethnicity, and achievement in tenth-grade students from a large northeastern city and found that males

reported higher math self-concept and expectation of success as compared to females, but no gender differences across ethnic groups were found.

Other research on self-efficacy suggests that it is malleable and can be influenced by social interactions (Zeldin et al., 2008), and there are significant efforts to understand how to support underrepresented groups, who may struggle against implicit stereotypes on top of normal learning struggles as their domain knowledge matures (Steele, 1997). Previous research shows that assimilation to social identity (e.g., gender and cultural identity) increases when people are experiencing uncertainty (Hogg, 2000). This could suggest that students could become more susceptible to negative cultural stereotypes (e.g., Steele, 1997), particularly those related to STEM performance, during periods of confusion associated with learning, making help-seeking an important behavior to study.

Given these findings, it seems likely that self-concept could vary not just by the demographics of individual students but also based on how those demographics influence the cultural interactions at a school level.

3.3. Data Collection

3.3.1 Reasoning Mind

This study analyzes data from students using Imagine Learning's Reasoning Mind (RM) Foundations (Figure 3.1), an intelligent tutoring system for elementary mathematics. The majority of Reasoning Mind's students are in Texas, but they represent a range of traditionally underrepresented populations across rural, urban, and suburban schools. RM

includes features that are designed to mimic other social experiences in the classroom, including both virtual peers and the system's signature pedagogical agent, known as the Genie, that guides students in their learning.

In this blended environment, students learn through self-paced problem solving, interactive explanations, and skill-based games. Problem sets are classified into three groups based on increasing levels of difficulty: (1) A-level problems on fundamental skills; (2) B-level (optional) problems on a combination of skills; and (3) C-level (optional) problems on higher-order thinking skills. Reasoning Mind Foundations is generally used in a classroom environment. Teachers assign/unlock problem sets for students based on the topic of instruction. Past studies of Reasoning Mind Foundations have shown high student and teacher acceptance, increases in test scores, high time on task, and a positive affective profile (see review in Khachatryan et al., 2014).

3.3.2 Hints in Reasoning Mind

Hints are an integral part of RM Foundations. They are delivered only on student request and contain conceptual feedback intended to help students solve the problem. Figure 3.2 demonstrates a hint in the system for one of the basic A-level problems in RM Foundations. The system's hints are multi-level and do not always contain a bottom-out hint.

Figure 3.1 Left - Reasoning Mind Foundations home screen; Right – An example problem displaying the Genie

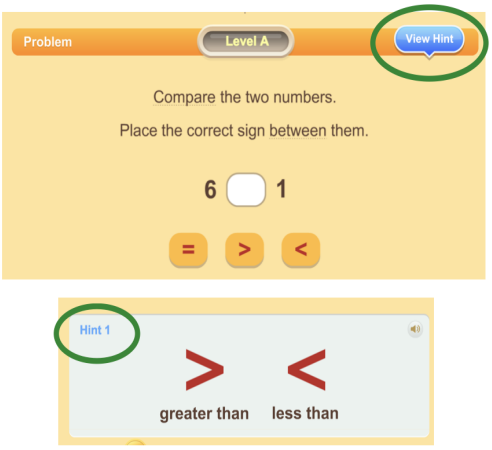


Figure 3.2 *Top* - Problem screen with a button to view hint (highlighted in green); *Bottom* - Hint displayed to the student when they request to view

3.3.3 Intrinsic and Extrinsic Motivations in Reasoning Mind

Most studies on motivation rely on student self-reports, which are dependent on how self-aware and reflective participants are (Renninger et al., 2018). Instead, we use student choices in the online math tutor as proxies for their intrinsic or extrinsic reason for it (cf. Barron, Gomez, Pinkard, & Martin, 2014). When students practice their skills in RM

Foundations, they are awarded points for solving problems correctly (and more points if they are consistent). They can use these points to buy virtual prizes or items like e-books, animations, and decorations for a virtual room called "My Place" (Figure 3.3), a feature seen in other learning systems as well. These extrinsic motivators are analogous to the game-like features in other EdTech systems such as iSTART-ME, the Motivational Enhanced (ME) version of iSTART, an intelligent tutoring system for reading (i.e., Jackson et al., 2009).

Another automatic outcome of earning more points and streaks in RM Foundations is that doing so opens up more challenging and optional problems (B and C level), which are otherwise locked until students demonstrate mastery in simpler problems (A level). Since B/C level problems are almost always optional, a student can choose to continue working on A level problems, allowing them to more easily earn points to purchase more items. The desire for challenge within the task rather than external rewards is a strong indicator of intrinsic motivation (Ryan & Deci, 2000), as is the will to pursue the learning task when it is a "free choice" (Deci, 1971). Thus, we use the number of items purchased and the number of B and C level problems attempted as proxies for students' extrinsic and intrinsic motivation, respectively. Our proxy for differentiating extrinsic motivation from intrinsic assumes that students understand that it is easier to earn points by solving the simpler problems than it is to do so by taking on more challenging material.

Figure 3.3 Left – "My place" lobby with entrances to library and great hall. At the bottom are the points awarded; Middle – Library with books and movies purchased using points; Right – The great hall decorated with furnishing items purchased using points

3.3.4 Participant Schools

We analyze data from 110 Texas schools across 25 school districts who used Reasoning Mind during the academic year 2017-2018 as part of their regular mathematics instruction, in schools where at least 25 students were using the software. There is a total of 9,122 2nd through 5th-grade students in this data (4,749 2nd graders, 1,964 3rd graders, 1,582 4th graders, and 827 5th graders) – i.e., Reasoning Mind was more widely used in 2nd-grade classes than older students. However, there was considerable variation in the use of Reasoning Mind across grades in different schools -- the standard deviations of the proportion of grades across schools are 33.06%, 16.87%, 15.03%, and 19.29% for grade 2, grade 3, grade 4, and grade 5 respectively. On average, there were 75 students using Reasoning Mind Foundations per school (min = 25; SD = 70) and 364 per school district (SD = 730), with one large urban district in Texas constituting the majority of our data, with 3,039 students across 62 schools.

Comprehensive log data captured student interactions with the system for the entire period, resulting in data for all 9,122 students. Surveys were administered once at the beginning and once at the end of the year to collect data on student math identity, resulting in complete surveys for 2,238 students in 22 schools.

3.4. Data Exploration

Considerable variation exists in the measures being analyzed in this study: help-seeking behaviors (i.e., hint usage), math performance, and pre- and post-year measures of math self-concept.

3.4.1 Exploring Help-Seeking

From the interaction log data, we operationalize help-seeking behavior as the number of hints used by a student in Reasoning Mind Foundations. As shown in Figure 3.4 (left), students in this study averaged less than 30 hint requests annually (mean = 27.01, SD = 55.72). The overall low hint usage could be attributed to RM being highly scaffolded – anticipating many student questions beforehand.

3.4.2 Exploring Intrinsic Motivation

As described in section 3.3.3, we use the student choice of solving advanced and optional B and C-level problems as our proxy for intrinsic motivation. Accordingly, we operationalize intrinsically motivated behavior as the number of B and C-level problems attempted by a student in Reasoning Mind Foundations. As shown in Figure 3.4 (left),

students in this study averaged less than 30 B and C-level problems annually (mean = 26.81, SD = 61.60).

3.4.3 Exploring Extrinsic Motivation

As described in section 3.3.3, we use the student choice of buying items (virtual prizes) to decorate a virtual room called "My Place" as our proxy for extrinsic motivation. Accordingly, we operationalize extrinsically motivated behavior as the number of items purchased by a student in Reasoning Mind Foundations. As shown in Figure 3.4 (left), students in this study averaged less than 15 item purchases annually (mean = 12.51, SD = 10.87).

3.4.4 Exploring Math Performance

For the purposes of this paper, math performance is defined as the accuracy of student responses to A-level problems in Reasoning Mind Foundations, i.e., the ratio of the number of correct answers to the number of problems attempted. We choose only A-level problems because they represent the core curriculum within the software. We obtain the problems attempted and the correctness of student answers from the interaction log data. As presented in Figure 3.4 (right), student-level calculations show a mean of 0.77 (SD = 0.14) – i.e., students obtained correct answers 77% of the time.

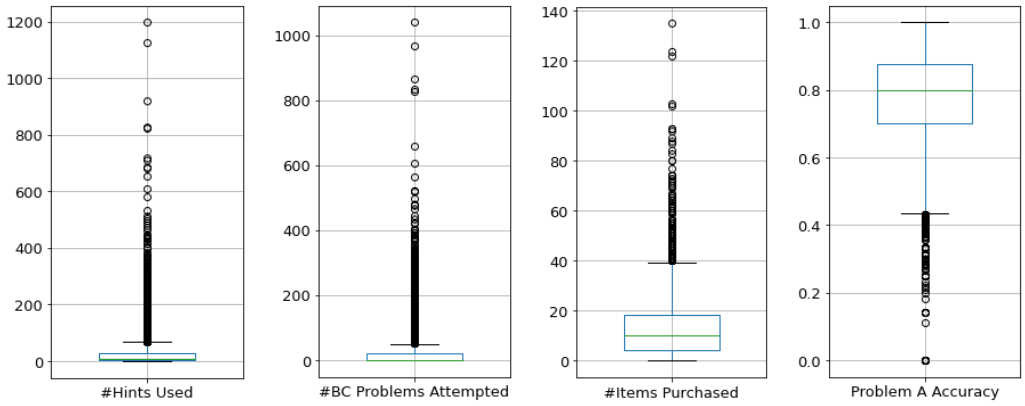


Figure 3.4 From left to right: Distribution of the number of hints (*leftmost*), number of B and C-level problems attempted, number of items purchased, and math performance (accuracy in A-level problems; *rightmost*). The middle line in the box indicates the median value.

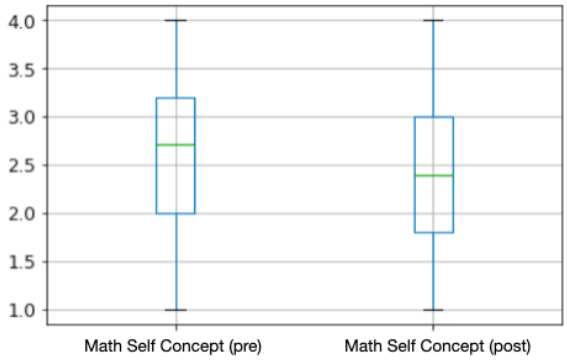


Figure 3.5 Distribution of the pre and post measures of math self-concept

3.4.5 Exploring Math Self-Concept

Students' self-concept in mathematics was measured using a five-item survey adapted from Marsh et al. (2005). This survey was administered twice--once at the beginning of the academic year (pre) and once at the end of the academic year (post). The survey included questions like "Math just isn't my thing" and "Some topics in math are just so hard that I know from the start I'll never understand them." Students took the survey voluntarily, and

each item in the survey was answered with a four-point Likert scale. Our previous work used this data to build a predictive model of math identity-related constructs like self-concept using language and behavior patterns (Crossley et al., 2020).

The distribution of students' responses is given in Figure 3.5 (self-concept pre: mean = 2.64 standard deviation = 0.77; self-concept post: mean = 2.44, standard deviation = 0.80). As summarized in Marsh et al. (2005), domain-specific self-concept (e.g., mathematics self-concept) shows developmental patterns of decline from early childhood to adolescence and then increases during early adulthood. We see a similar pattern in our student population, with the self-concept post-test score statistically significantly lower than the pre-test ($t = 5.2$, $p < 0.001$). The internal consistency of these items was found to be satisfactory, with a Cronbach's $\propto$ of 0.74.

3.4.6 Exploring School-Level Differences

Next, we explored the school-level differences in student outcomes (math performance and self-concept) and hint usage. As we can see in Figure 3.6, Figure 3.7, and Table 3.1, there is considerable variance in the variable aggregates (mean) across the schools, especially in hint usage and math performance.

Table 3.1 Mean and standard deviation (SD) of the school-level aggregates of the variables and outcomes

	Mean	SD
Hint Usage	24.52	21.30
Number of B & C level problems attempted (Proxy for intrinsic motivation)	23.79	24.04
Number of Items Purchased (Proxy for extrinsic motivation)	12.82	4.91
Math Performance	0.78	0.04
Math Self-Concept (Pre)	2.69	0.35
Math Self-Concept (Post)	2.43	0.15

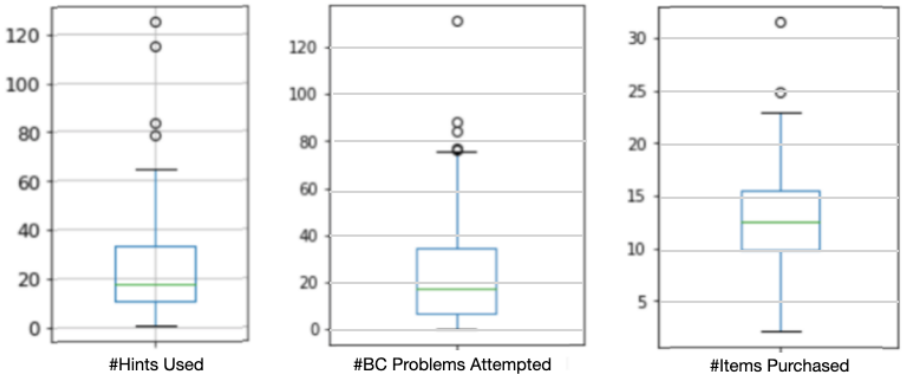


Figure 3.6 Distribution of school-level aggregates of the variables

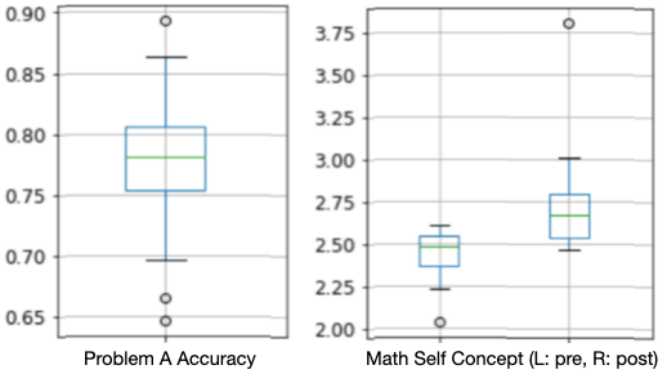


Figure 3.7 Distribution of school-level aggregates of the outcomes

3.4.7 Summarizing School-Level Demographics

We characterize the schools in our sample using demographics from the Texas Education Agency's (TEA) public data repository. These data capture some contextual factors that are likely to affect the school culture or climate and thereby may affect student use of RM Foundations.

Table 3.2 Mean and standard deviation (SD) of the school-level demographics

	Mean	SD
% Economic Disadvantage	78.3	16.6
% Limited English Proficiency	41.4	20.6
% Special Education	6.9	3.1
Urbanicity (binary)	60.4%	-
Charter (binary)	27.1%	-

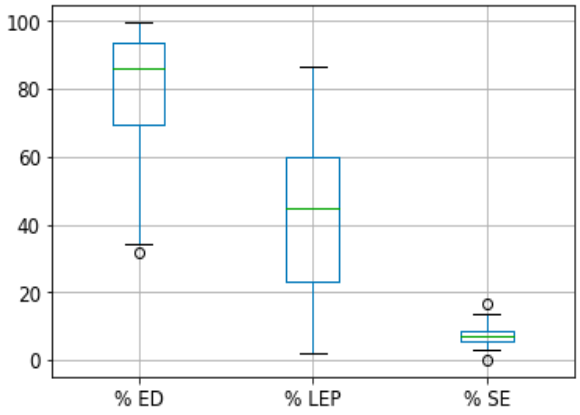


Figure 3.8 Distribution of non-binary school-level demographics for the 110 schools selected in this study. ED - Economically Disadvantaged; LEP - Limited English Proficiency; SE - Special Education

Table 3.2 summarizes the first set of school-level demographics obtained from TEA sources, including the percentage of students at the school who are classified as (1) Economically Disadvantaged (EcD), as (2) Limited English Proficiency (LEP), or as (3)

Special Ed (SpEd), as well as (4) the urbanicity of the school and whether or not it is a (5) charter school. These terms are defined by the State of Texas as follows (TEA, 2018). Students are classified as EcD if they qualify for free or reduced-price meals under the National School Lunch and Child Nutrition Program; it is worth noting that a large proportion (avg = 40%) of Texas public school students qualify for this status (TEA, 2018a). SpEd classifications are given to students who qualify for services for cognitive, emotional, or physical disabilities. LEP status is conferred for students whose primary home language is not English and who also fail to meet proficiency standards as established by either an approved testing measure or by a Language Proficiency Assessment Committee (LPAC). Finally, the TEA (2018b) classifies a school district as urban (or not) based on whether its school district (a) is located in a county with a population of at least 960,000; and (b) has the largest enrollment in the county or its enrollment is greater or equal to 70% of county's largest district. As seen in Table 3.2 and Figure 3.8, we have a diverse set of schools along these dimensions.

We also considered school-level data on the percentage of students representing major ethnic/racial groups, using the categories provided by the TEA. As Table 3.3 shows, Hispanic students (the TEA's term) are by far the largest group in these schools (mean = 63.5%), followed by African American students (mean = 17.5%), White students (mean = 13.5%), and then Asian students (4.5%), but as Figure 3.9 the schools show considerable variance in terms of this composition. To avoid noisy results, this analysis considers only

groups that constitute at least 5% of the student population: Hispanic, African American, White, and Asian.

Table 3.3 Mean and standard deviation (SD) of the school-level percentages of ethnicities

	Mean	SD
% Hispanic	63.5	24.5
% African American	17.5	17.8
% White	13.1	16.2
% Asian	4.5	7.8
% American Indian*	0.36	0.4
% Pacific Islander*	0.04	0.1
% Two or More Races*	1	1

*Categories constituting less than 5% of the data were excluded from further analysis.

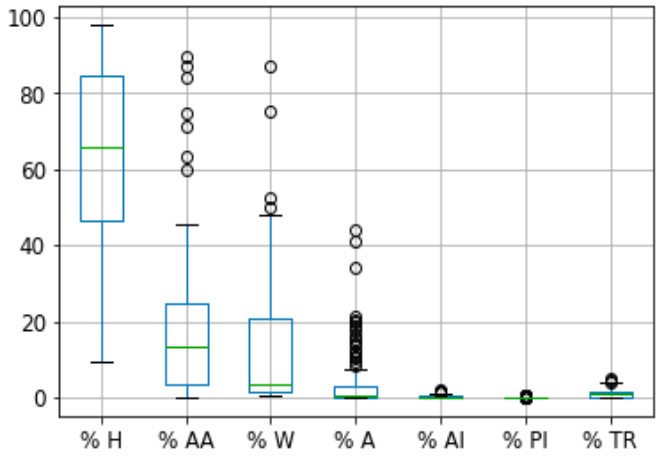


Figure 3.9 Distribution of percentages of school-level ethnicities for the 110 schools selected in this study. H - Hispanic; AA - African American; W - White; A - Asian; AI - American Indian; PI - Pacific Islander; TR - Two or more races

3.5. Analysis

Our data exploration (Section 3.4) suggests that help-seeking behavior, intrinsic and extrinsic motivation, math performance, math self-concept, and demographics each vary by school. Our goal with the analysis is to investigate whether the relationship between the behavior (hint usage and motivational behaviors) and the outcomes vary for different student populations. Thus, we conduct a two-step data analysis to explore how help-seeking and the two motivations might differ based on student demographics, while controlling for performance and math self-concept.

In the first step, we determine how closely students' math performance and self-concept measures correlate to their hint usage and motivational behaviors within each school, using Spearman ρ correlations due to non-normality in the data. That is, we produce three types of measures for the three behaviors for each student, the correlation between a behavior and performance on A-level problems, the correlation between a behavior and the pre-year survey of self-concept, and the correlation between a behavior and the post-year survey of self-concept.

In the next step, we determine whether the differences in these correlations are themselves correlated to school-level demographics. Note that in the first step, the unit of analysis for the correlations is the student, but in the second step, the unit of analysis is the school. We conduct two-tailed tests to report the significance levels.

3.6. Results

3.6.1 Relationship Between Variables and Student Outcomes

3.6.1.1 Help-Seeking and Student Outcomes

Figure 3.10 shows the distribution of correlations across schools between students' hint usage and their math performance and math self-concept (taken once at the beginning (pre) and again at the end of the year (post)).

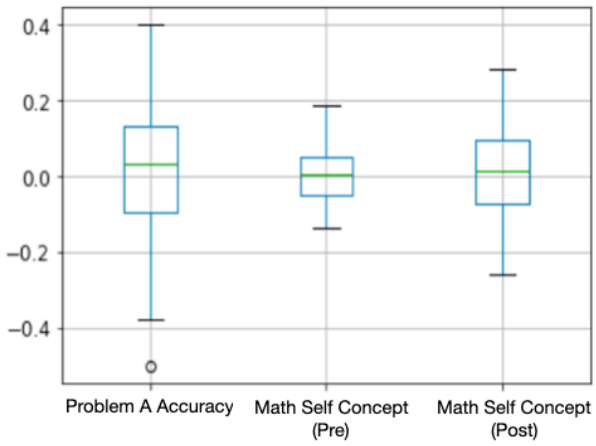


Figure 3.10 School-level correlations between hint usage and math performance vs. the correlations between hint usage and self-concept measures

Help-Seeking and Math Performance

Grouping students by schools allows us to see that the relationship between hint usage and math performance differs in important ways, even before we look at demographic variables more directly. Specifically, when student measures are aggregated at the school level, as they are in Figure 3.10, the correlation between hint usage and math performance ranges from -0.39 to 0.40 (SD = .18). In contrast, when we do not aggregate students into school-level populations (instead, treat them all as a single population), there is not a significant relationship between hint usage and math performance (ρ = -0.008, p = 0.44). In other

words, while there is an appearance of varied effects within individual schools, it appears to cancel out when considering all schools together.

Help-Seeking and Math Self-Concept

Like math performance, math self-concept also shows signs of sub-population differences. When students are aggregated into school-level populations, as shown in Figure 3.10, the correlations between hint usage and math self-concept show a relatively wide range. For pre-year surveys, the correlation ranges from -0.14 (students with lower self-concept are most likely to use hints) to 0.19 (students with higher self-concept are most likely to use hints), and an even wide range is found for post-year survey correlations (-0.27 to 0.30). In contrast, when the students in this data were treated as a single population, the correlations were essentially zero ($\rho = -0.008$, $p = 0.442$ for pre and $\rho = -0.007$, $p = 0.77$ for post).

Summary of Help-Seeking Variance

There is considerable variance in the school-level correlations between hint usage and student outcome measures (SD = 0.18 for math performance, SD = .084 for pre-year self-concept, SD = 0.118 for post-year self-concept). This variance indicates that students likely have different motivations for using hints, and that hints are associated with positive outcomes in some student populations but not in others.

As seen in Figure 3.10, the median of the correlations is centered close to zero. For these schools, there is no association between hint usage on student outcomes. Figure 3.10 also

shows that the distribution of these correlations is not skewed, meaning that hint usage is not universally positively or negatively associated with student outcomes across schools.

3.6.1.2 Motivational Behaviors and Student Outcomes

Figure 3.11 shows the distribution of correlations across schools between the students' motivational behaviors (number of B and C level problems attempted and number of items purchased) and their outcomes (math performance and the pre- and post- measures of math self-concept).

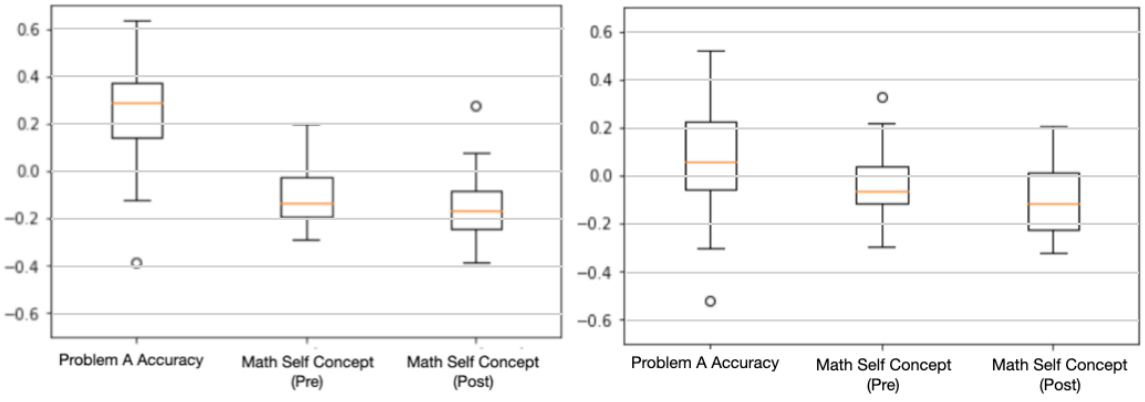


Figure 3.11 Distribution of correlations across schools between student outcomes and the proxies of their intrinsic (*left*) and extrinsic (*right*) motivation

Intrinsic Motivation, Extrinsic Motivation and Math Performance

When student measures are aggregated at the school level, we see that the correlation between the intrinsic motivation behaviors and math performance ranges from -0.12 to 0.62 (SD = .18). The majority of the schools show a significantly positive correlation between intrinsic motivation and performance, in line with the vast body of empirical

research confirming this relationship (Ryan & Deci, 2000). When we do not aggregate students into school-level populations (instead treat them all as a single population), there is still a significant relationship between intrinsic motivation and math performance ($\rho = 0.17$, $p < 0.001$) with a correlation value closer to the average across schools but doing so fails to capture the range and variance in this relationship.

We also find that the correlation between extrinsic motivation behaviors and math performance ranges from -0.31 to 0.52 (SD = .19). When we do not aggregate students into school-level populations, there is still a significant relationship between extrinsic motivation and math performance ($\rho = 0.04$, < 0.001), but the correlation value is very close to zero.

The school-level correlations between the motivations and math performance range from -0.12 to 0.62 for intrinsic motivation and – 0.31 to 0.52 for extrinsic motivation (SD = 0.18 for intrinsic motivation and SD = 0.19 for extrinsic motivation). This variance indicates that the relationship between the motivations and student performance differs by student populations. The average correlation, when grouping students by schools and taking an average across the schools, is 0.26 for the intrinsic motivation, while extrinsic motivation has a relatively lower average correlation of 0.08. The range of the correlations is also shifted upwards for intrinsic motivation when compared to the extrinsic motivation – with a tail of -0.12 vs. -0.31, head of 0.62 vs. 0.52, and a similar standard deviation. Our results show that there are school-level differences in the correlation between the different

motivations and student performance, with extrinsic motivation showing more negative associations in some schools than others.

Intrinsic Motivation, Extrinsic Motivation and Math Self-Concept

Like math performance, math self-concept also shows signs of sub-population differences. When students are aggregated into school-level populations, the correlations between the motivations and math self-concept show a relatively wide range, although not as prominent as with math performance.

For pre-year surveys, the correlations for intrinsic motivation ranges from -0.27 (students with lower self-concept are most likely intrinsically motivated) to 0.21 (students with higher self-concept are most likely intrinsically motivated), shifted lower for post-year survey correlations (-0.38 to 0.15). When the students in this data were treated as a single population, the correlations are close to zero, though statistically significantly for pre (ρ = -0.067, p = 0.002) and statistically significantly negative for post (ρ =-0.163, < 0.001).

For pre-year surveys, the correlations for extrinsic motivation ranges from -0.29 to 0.23, which is similar to the post-year survey correlations (-0.33 to 0.21). When the students in this data were treated as a single population, the correlations are very small or close to zero for both pre (ρ = -0.037, p = 0.12) and post (ρ =-0.099, p < 0.001).

The schools at the tail ends of these distributions are interesting case studies. They represent the cases where hint usage and motivations have either a notably high positive correlation or a notably high negative correlation with our outcome measures. As such, it

becomes important to understand what demographics are involved in order to address any potential disparate impacts of the hint function in the system.

3.6.2 The Influence of School Demographics

3.6.2.1 Demographic Differences in the Relationship between Help-Seeking and Student Outcomes

School-level demographic variables help to capture some of the variance in the relationship between hint usage and the student outcomes measured in this study (math performance and math self-concept). These findings are summarized in Tables 3.4 and 3.5.

Table 3.4 Correlations between school-level demographics and the correlations resulted between student outcomes (math performance, self-concept) and help-seeking. p-value in parenthesis. Significant correlations after Benjamini and Hochberg post hoc corrections in bold.

	Correlation between number of hints and		
	Math performance	*Self-concept Pre*	*Self-concept Post*
Urbanicity (binary)	**0.292** (0.002)	0.130 (0.564)	0.080 (0.729)
% Economic Disadvantage	**0.256** (0.007)	0.182 (0.417)	-0.288 (0.205)
% Limited English Proficiency	**0.314** (0.001)	**-0.452** (0.035)	**-0.565** (0.008)
% Special Education	-0.002 (0.982)	**0.463** (0.030)	**0.444** (0.044)
Charter Status (binary)	-0.083 (0.389)	-0.058 (0.799)	-0.269 (0.225)

Table 3.5 Correlations between school-level ethnicity and the correlations resulted between student outcomes (math performance, self-concept) and help-seeking. p-value in parenthesis. No significant correlations were obtained after Benjamini and Hochberg post hoc correction was conducted.

	Correlation between number of hints and		
	Math performance	*Self-concept Pre*	*Self-concept Post*
% Hispanic	0.094 (0.329)	0.123 (0.587)	-0.153 (0.507)
% African American	0.054 (0.579)	-0.260 (0.243)	-0.174 (0.451)
% White	-0.194 (0.042)	0.103 (0.647)	0.095 (0.683)
% Asian	-0.037 (0.703)	-0.071 (0.753)	-0.107 (0.644)

School-level Demographics, Help-Seeking, and Math Performance

As Table 3.4 (above) shows, the relationships between hint usage and math performance differ significantly in terms of the school's urbanicity ($\rho = .292$, $p = .002$) as well as differences in the percentage of students who are economically disadvantaged (EcD; $\rho = .256$, $p = .007$) and limited English proficiency (LEP; $\rho = .314$, $p = .001$). Specifically, more hints are associated with higher math performance among urban students, but more hints are associated with lower math performance among suburban/rural students. In schools with a higher percentage of students who are economically disadvantaged (EcD) or limited English proficiency (LEP), the use of hints is associated with higher math performance. However, as Table 3.5 shows, other demographic categories that are often considered in educational research, namely ethnicity/race, are not predictive in this context.

School-level Demographics, Help-Seeking, and Math Self-Concept

The relationships between hint usage and math self-concept differ significantly in terms of the percentage of students with limited English proficiency (LEP; $\rho = -.452$, $p = .035$ for pre; $\rho = -.565$, $p = .008$ for post), and the percentage of students in special education (SpEd; $\rho = .463$, $p = .030$ for pre; $\rho = .444$, $p = .044$ for post). Specifically, in schools that serve a higher percentage of LEP students, students with low self-concept are more likely to use hints, while in schools with fewer LEP students, students with high self-concept are more likely to use hints. This finding is somewhat stronger for the end of year surveys than the start of year surveys.

The opposite pattern is shown among schools that serve a higher percentage of SpEd students. In these schools, their students with high self-concept use more hints, whereas that relationship is negative in schools that serve fewer SpEd students. This finding is consistent across the start of the year and end of the year surveys.

Other demographic factors from Table 3.4, namely urbanicity and EcD, were not significant for the relationship between help-seeking and math self-concept, despite being predictive of the relationship between help-seeking and math performance. School-level descriptions of ethnicity (Table 3.5) again did not help to explain the variance between math self-concept and hint usage.

3.6.2.2 Group Differences in the Role of Motivations in Student Outcomes

School-level demographic variables help to capture some of the variance in the relationship between hint usage and the student outcomes measured in this study (math performance and math self-concept). These findings are summarized in Table 3.6 and Table 3.7.

Table 3.6 Correlations between school-level demographics and the correlations resulted between student outcomes (math performance, self-concept) and behaviors related to intrinsic and extrinsic motivation. p-value in parenthesis. Significant correlations after Benjamini and Hochberg post hoc corrections in bold.

	Correlation between behaviors related to motivation and					
	Math performance		*Self-concept Pre*		*Self-concept Post*	
	Intrinsic Motivation Behavior	*Extrinsic Motivation Behavior*	*Intrinsic Motivation Behavior*	*Extrinsic Motivation Behavior*	*Intrinsic Motivation Behavior*	*Extrinsic Motivation Behavior*
Urbanicity (Binary)	**0.281** (0.003)	**-0.238** (0.010)	0.186 (0.408)	0.037 (0.869)	0.132 (0.510)	-0.049 (0.826)
% Economic Disadvantage	**0.230** (0.018)	-0.151 (0.116)	0.021 (0.926)	0.280 (0.205)	0.275 (0.227)	**0.499** (0.017)
% Limited English Proficiency	0.161 (0.100)	**-0.226** (0.017)	0.038 (0.863)	0.024 (0.914)	0.018 (0.937)	0.369 (0.090)
% Special Education	**0.247** (0.011)	0.0548 (0.569)	0.167 (0.459)	0.117 (0.604)	0.149 (0.519)	0.109 (0.627)
Charter (Binary)	**-0.244** (0.012)	**0.217** (0.022)	0.086 (0.702)	0.072 (0.750)	0.167 (0.469)	0.115 (0.610)

Table 3.7 Correlations between school-level ethnicity and the correlations resulted between student outcomes (math performance, self-concept) and behaviors related to intrinsic and extrinsic motivation. p-value in parenthesis. No significant correlations were obtained after Benjamini and Hochberg post hoc correction was conducted.

	Correlation between behaviors related to motivation and					
	Math performance		*Self-concept Pre*		*Self-concept Post*	
	Intrinsic Motivation Behavior	*Extrinsic Motivation Behavior*	*Intrinsic Motivation Behavior*	*Extrinsic Motivation Behavior*	*Intrinsic Motivation Behavior*	*Extrinsic Motivation Behavior*
% Hispanic	0.012 (0.903)	-0.203 (0.034)	0.065 (0.774)	0.196 (0.382)	0.116 (0.616)	0.427 (0.047)
% African American	0.072 (0.463)	-0.038 (0.696)	0.045 (0.844)	-0.290 (0.190)	0.173 (0.454)	-0.161 (0.474)
% White	-0.091 (0.356)	0.207 (0.030)	-0.408 (0.056)	-0.191 (0.395)	-0.192 (0.404)	-0.123 (0.585)
% Asian	-0.203 (0.038)	0.081 (0.398)	-0.045 (0.842)	-0.330 (0.134)	-0.008 (0.971)	-0.312 (0.158)

School-level Demographics, Intrinsic & Extrinsic Motivations, and Math Performance

As Table 3.6 (above; column 2) shows, the relationships between the intrinsic motivation and math performance differ significantly in terms of the school's urbanicity ($\rho = .281$, $p = .003$), whether or not it is charter ($\rho = -.244$, $p = .012$) as well as differences in the percentage of students who are economically disadvantaged (EcD; $\rho = .230$, $p = .018$) and in special education (SpEd; $\rho = .247$, $p = .011$). Specifically, there is a positive association between higher usage of more advanced (B- and C-Level) problems and math performance among students from urban schools, and a negative association between higher usage of more advanced (B- and C-Level) problems and math performance among students from suburban/rural schools. More attempts of advanced problems are associated with higher math performance among non-charter students, but more attempts of advanced problems are associated with lower math performance among charter students. Schools with a higher percentage of students who are economically disadvantaged (EcD) have a positive

association between higher usage of more advanced problems and math performance. Finally, schools with a higher percentage of students in special education (SpEd) have a positive association between higher usage of more advanced problems and math performance.

In contrast, Table 3.6 (above; column 3) shows that the relationships between extrinsic motivation and math performance differ significantly in terms of the school's urbanicity (ρ = -.238, p = .010), whether or not it is charter (ρ = .217, p = .022) as well as differences in the percentage of students who have limited English proficiency (LEP; ρ = -.226, p = .017). Specifically, the association between higher usage of the gamified extrinsic motivations and math performance is positive among students from suburban and rural schools and negative among students from urban schools. Furthermore, the association between higher usage of the gamified, extrinsic motivation behaviors, and math performance is positive among students from charter schools and negative among students from non-charter schools. Schools with a higher percentage of students who have limited English proficiency (LEP) show a negative relationship between extrinsic motivation and math performance. Also, like in the hint usage results, ethnicity/race are not significantly correlated in this context as well (Table 3.7). As such, the correlations between extrinsic motivation and math performance across schools range relatively more negative than intrinsic motivation (-0.12 (intrinsic) vs. -0.31 (extrinsic); Section 3.6.1.2), suggesting the behaviors associated with extrinsic motivation may be more harmful to certain schools.

An interesting pattern in these results is the inverse relationship of the same demographic variable with intrinsic and extrinsic motivation. For example, suburban and rural schools have a negative association between math performance and intrinsic motivation and a positive association between math performance and extrinsic motivation. This result suggests that more support may be needed to transition students in rural and suburban settings from gamified extrinsic motivators to develop competence in more advanced and challenging math problems. Similar support may also be needed for students in charter schools. Another concerning trend among schools with a higher percentage of students who have limited English proficiency is the negative association between extrinsic motivation and math performance and the lack of a strong relationship between intrinsic motivation and math performance. This could imply that the gamified motivation in the system may not be enough to improve math competence in these schools, while their students also fail to find motivation in the increased challenge in advanced problems. On the other hand, schools with a higher percentage of economically disadvantaged students could be inherently motivated to solve challenging problems, based on the correlation between the percentage of students economically disadvantaged and intrinsic motivation and the lack of correlation between economic disadvantage and extrinsic motivation. The same is true for schools with a higher percentage of special education students, though intrinsically motivated behaviors may be promoted in this case by the teachers' aides and specialists helping these students.

School-level Demographics, Intrinsic & Extrinsic Motivations, and Math Self-Concept

The relationships between the usage of extrinsic motivation and math self-concept (post) differ significantly in terms of the percentage of students who are economically disadvantaged ($\rho = 0.499$, $p = .017$). This relationship is not consistent across the start of the year and end of the year surveys, unlike what we observed in hint usage (Table 3.4). Also, other demographic factors from Table 3.6 that were predictive of the relationship between the motivations and math performance were not significant for the relationship between the motivations and math self-concept, neither were the ethnicity/race variables (Table 3.7).

3.7. Discussion

Talking about the social and ethical impacts of computer bias, Garcia (2016) said, "the side effects of unintentionally discriminatory algorithms can be dramatic and harmful." In recent years, data-driven systems have come under scrutiny for amplifying existing social inequities or, in some cases creating new ones (Garcia, 2016). As such, there has been increasing amounts of research on fairness in data and machine learning systems. However, much of this work has focused on optimizing systems based on abstract universal notions of fairness or de-contextualized quantitative metrics and ignoring social, political, and cultural deliberation (Green & Hu, 2018). However, education is a context where achieving fairness with sociotechnical solutions poses unique challenges tied closely to the sociocultural aspects of the domain (Ocumpaugh et al., 2015; Ito, 2017; Karumbaiah et al., 2019, 2021). Despite the increase in the use of data analytics and AI-based systems in education, relatively little work has focused on establishing what fairness means in this

context and exploring approaches to achieving it (Holstein & Doroudi, 2019; Subotzky and Prinsloo, 2011; Slade & Prinsloo, 2013). The current literature on ethics in the field has mostly been interested in issues of data ownership and privacy and institutional and policy level considerations (Draschler et al., 2015; Tsai & Gasevic, 2017). Through this paper, we hope to contribute to the emerging conversations on fairness and equity in EdTech systems in two ways – 1) by investigating inequitable outcomes across student subpopulations on two of the fundamental psychological constructs in EdTech design, 2) by demonstrating the use of publicly-available, school-level demographics for fairness research where individual student demographics may be difficult or impossible to acquire.

3.7.1 Difference in Outcomes Across Student Subpopulations

3.7.1.1 Help-Seeking

Hint-seeking behaviors have been a source of interest among EdTech researchers since the early days of the adaptive EdTech field (Anderson, 1993; Aleven & Koedinger, 2000), yet understanding which hints are effective, for whom, and under what conditions, remains a somewhat elusive goal. A large part of answering these questions likely lies in understanding what motivates a student to seek help. Ideally, we would like students to use hints to improve their understanding of the material, but as these results show, students who are struggling do not always make use of available resources effectively. Within this data (with a relatively low hint usage overall)—which involves students in the same state using the same mathematics learning system—there are also schools where hint usage is associated with low-performance. If these students are benefiting from this hint usage, it is

not measurable with the variables considered in this study. This finding suggests that the hints could be ineffective at helping these particular students to learn the material.

At least part of the school-level differences in the correlation between hint usage and math performance seems to be related to school-level demographics, but interestingly, the schools where hint usage appears to be most advantageous are those that enroll larger numbers of students who would typically be thought of as disadvantaged by the school system. That is, schools with fewer LEP students are more likely to have low performers who do not appear to be benefitting from hint use. Schools with fewer students receiving free or reduced-price lunch are more likely to have low performers who do not appear to be benefitting from hint use. Schools in large urban centers are less likely to have low-performing students who do not appear to be benefitting from hint use.

The relationship between hint usage and self-concept is also complicated. Students in schools that serve more LEP students tend to show a negative relationship between self-concept and hint usage. That is, those students with low math self-concept appear to use more hints (in those schools). However, in schools that serve more SpED students, the relationship between self-concept and hint usage is positive (i.e., students who are relatively more sure of themselves ask for more hints). It is also possible that the smaller number of schools sampled for self-concept (compared to math performance) made it more difficult for these relationships to emerge.

Ethnic population differences were not particularly revealing in this study, and it is not entirely clear why. It is possible that, say, the LEP findings are strong enough to warrant

further divisions to the subpopulations included in this study, a possibility that has not yet been explored in this data, such as dividing LEP students based on different languages. However, it is also possible that some of the linguistic differences that influence classroom practices of different ethnic groups (e.g., Hudley & Mallinson, 2015)—practices that may include figuring out how to ask for help—are less relevant in an online context like Reasoning Mind where the student is simply pressing a button to request a hint.

3.7.1.2 Intrinsic and Extrinsic Motivations

Research on fostering student motivation has a long history of research in psychology. There have been mixed opinions on the efficacy of extrinsic motivations like tangible rewards, while there is more consensus on the important role of intrinsic motivation in high-quality learning (Deci & Ryan, 2000). Given that many students are not intrinsically motivated in any given subject, EdTech designers often turn to features that can increase extrinsic motivation, such as rewards and gamified activities. Research suggests that the design of extrinsic motivators should value student autonomy and foster self-regulation instead of exerting external control (Ryan & Connell, 1989). Our literature review establishes the role of social factors and the need for different supports for different learners to catalyze student competence and autonomy. However, very little research in EdTech has explored the role of social context in the efficacy of extrinsic motivation.

In this data, we observe that the relationship between students' intrinsic and extrinsic motivation and their outcomes (math performance, math self-concept) is associated with student demographics. In the case of our binary variables, urbanicity and charters, we find

the inverse correlations between math performance and intrinsic versus extrinsic motivations. Suburban and rural schools (like charter schools) have a negative association between math performance and intrinsic motivation and a positive association between math performance and extrinsic motivation. Our results suggest that students from rural and suburban schools may need better supports to transition from gamified extrinsic motivations to developing interest in more advanced and challenging math problems. Similarly, our results suggest that such supports may also be needed for students in charter schools. A rather concerning trend in schools with a higher percentage of students who have limited English proficiency is the negative relationship between extrinsic motivation and students' math competence. These results suggest that EdTech designers need to pay special attention to their students' demographic context when designing extrinsic motivations in EdTech systems if our goal is to deliver equitable student outcomes. Overall, if future studies establish a causal link, then we recommend that the system behavior should change based on the student needs. For instance, the design of extrinsic rewards and hints should vary based on student self-concept. At a minimum, teachers should have the ability to override system decisions or make changes to certain design elements based on their assessment of varying student needs, to address possible mismatches between design and the needs of specific learner populations.

3.7.2 Implications for EdTech Designers and Researchers

One of the main implications of this paper for EdTech designers is that a universal design that focuses on improving student outcomes while ignoring individual or group differences

might not produce the desired results. One key finding of this paper is that the group differences that matter most for design might not be the groups that are the most immediately obvious. The work presented here is a step towards understanding which group differences may matter, but as recommended by Baker and colleagues (2019), research in this area should explore a broader range of conceptualizations of context and identity than are currently considered. Ultimately, the vision of culturally aware tutoring systems (Blanchard & Mizoguchi, 2008) can only be fully achieved if we know which groups to adapt them to, and how. Thus far, however, personalization in help design has not taken these types of issues into consideration, primarily focusing on understanding student cognition to provide hints based on the pedagogical content (Aleven et al., 2016). Our findings suggest that developers and researchers on adaptive EdTech should explore broader contextual factors to analyze the effectiveness of hint usage across student subpopulations, and adapt help to a broader vision of student need. Similarly, features to increase extrinsic motivation, such as gamification, may not always be beneficial to academically unmotivated students, depending on group differences.

To illustrate this, let's take the example of students with limited English proficiency. As shown in Section 3.6.2, there is an inverse relationship between help-seeking and the two student outcomes (performance vs. self-concept). In schools with a higher percentage of limited English proficient students, higher hint usage is associated with high math performance but low math self-concept. On the other hand, in schools with more native English speakers, higher hint usage is associated with low math performance but high math

self-concept. This is an interesting case for ITS designers to investigate further. Is the text-heavy nature of the hints contributing to this finding? Are limited English proficient students using hints to improve on their math skills, but the cognitive load in processing more verbal content is causing a negative impact on their self-efficacy? Such investigations could open up opportunities for design innovations to better support students. Would it help to use multiple representations (visual, auditory, symbols) and give autonomy to the students to choose which representation to use? In summary, including school-level demographics to the analysis of complex constructs like help-seeking is an important step in developing designs that are appropriate for all learners.

Our results suggest that research on complex yet widely-used constructs related to student learning and engagement may not generalize well across diverse student populations, especially when the studies are conducted in a small-scale or with convenience samples. Although this finding is not novel in itself, this paper demonstrates an approach to assess the generalizability of the EdTech research findings by using publicly-available, school-level demographics when we have access to larger data (e.g., interaction logs), which may be particularly important when access to individual student demographics is restricted. A broader implication of our work for EdTech researchers is to consider the student demographic factors when explaining contradictory findings on the relationship between student behaviors and outcomes in virtual learning environments. As such, we would echo Paquette et al.'s (2020) recommendation that the research community pay more attention to student demographics, including both commonly reported categories of gender and

race/ethnicity, and factors like LEP, EcD, SpEd, urbanicity, and school type (e.g., public/private/charter), as there is substantial evidence that these factors often influence student behaviors. In addition to providing a more holistic picture of research to the readers, this practice of reporting diverse contextual factors could also help with situating the research in prior literature and aid replication or application in a similar context.

3.7.3 Implications for the Fair-ML Community

Despite recent advancements, the field of fair Machine Learning (ML) has been criticized for focusing on algorithmic concerns and mathematical definitions of fairness rather than engaging with the broad set of ethical, social, and political concerns within the contexts in which applications are deployed (Green & Hu, 2018). The interaction between ML systems and social worlds can sometimes lead to effects unanticipated from a purely technical perspective. It is time for application domains like EdTech to actively contribute by bringing our nuanced challenges to the multidisciplinary conversation around fairness (Holstein & Doroudi, 2019). As Selbst and colleagues (2019) explain, the popular ML approach of abstraction - abstracting out domain-specific aspects of a problem - risks rendering ML ineffective when used to define fairness and develop fair-ML algorithms in a social context. Instead, they argue, fairness requires a nuanced understanding of the social context, its politics, and all the actors involved. In this study, we investigate how the implications of student behavior and motivation within a learning system may be dependent on social and group factors. Specifically, we identify which subpopulations need particular attention by identifying potential blind spots in the generalizability of past

results. This work is necessary in order to understand how to make EdTech systems fairer and foster more equitable student outcomes.

A popular strategy to mitigate bias in the industry is to collect more training data (Holstein et al., 2019). However, collecting more rich and complex data in domains like education may not always be feasible. Specifically, collecting fine-grained data on social context and individual demographics can be difficult in education due to student privacy concerns. This study demonstrates a potential workaround to this challenge by collecting coarser-grained school-level demographics data.

Lastly, ML-based sociotechnical systems need to recognize and adapt to the changing social circumstances in the context in which they are deployed - challenging the current practice of treating historical data as the ground truth. For instance, Schofield (1995) reported that students in urban schools skipped lunch and stayed after school to use an intelligent tutoring system – not a common pattern 25 years later. Patterns of interaction and properties of students may change even over a shorter period of time, and may change as a result of using the systems we develop. Moreover, continuing to make predictions using the learning from past data could also impede students' progress in developing interest. To illustrate this, take the example of extrinsic and intrinsic motivation discussed in this paper. The goal of interventions in education is to build student interest in the subject matter, hopefully to the extent that activities started out with extrinsic motivations lead to the student developing an inherent interest in the content. Even when an adaptive EdTech system is trained on data from the same population where it is applied, it may fail to adapt

to the improving conditions as a result of the technological intervention if the algorithm stays blind to these changes. For example, continuing to motivate students with extrinsic tangible rewards over increasing student autonomy to attempt challenging problems.

3.7.4 Limitations and Future Work

This paper only considered a small number of sociocultural variables (albeit more than are commonly seen in research on help-seeking or motivation within EdTech). We acknowledge that there are many other sociocultural aspects that influence a student's engagement and learning with an ITS. In the case of students' help-seeking behavior, the perceptions of help-seeking within their classroom (peers, teachers) and outside (family, friends) can influence student choices. Similarly, social conditions in school and at home can systemically influence intrinsic and extrinsic motivation. While this paper focuses on broadly-defined school-level demographics, we believe that it would be beneficial to look at other influencers from the student's social context. For instance, the pedagogical practices of the teacher in the math classroom could influence what students perceive as appropriate help-seeking and the value of intrinsic vs. extrinsic motivation.

More broadly, the priorities of the school district and state might also impact the pedagogical choices made in schools. Teachers' choices are influenced by public policy. Shortly after the completion of our data collection, Texas issued letter grades (A-F) (The Texas Tribune, 2018) to its school districts based on a complex formula involving overall student performance on standardized exams, overall year-to-year improvement, and improvement for specific sub-groups. These ratings were generally lower in districts with

higher rates of economically disadvantaged students, creating different degrees of pressure where demographics differ. The pressure of performing well (as measured by standardized tests), in many cases with limited resources, could influence what is being prioritized as the goal of math learning in these schools. While quantifying these factors to include in an analysis is not straightforward, these factors no doubt drive the type of differences that are seen between schools with different demographics.

Another potential limitation to our findings is seen in this paper's lack of explicit consideration of gender. Gender was not investigated in the current paper, as public schools generally have balanced gender distributions (as was the case in this dataset), leading to limited power to observe any difference that might exist. This leads to a more general point. It would be beneficial to analyze the impact of demographics at the student level, both to replicate the relationships seen here and to study whether students who are outliers in their own schools have different patterns. However, collecting student-level data is not always feasible, and this study has demonstrated that school-level aggregates can still help us understand the role of demographic factors in understanding motivation and help-seeking behaviors.

In addition, it is important to note that the findings were obtained in a single EdTech system, geographical region, and point in time. When using findings of this nature, it is important not just to consider whether the findings are substantial in effect size, but whether they continue to apply. For instance, the impacts of the variables studied here may change between regions (where variables such as limited English proficiency may correspond to

different population groups), or over time, as society changes. In general, work investigating the impact of demographics on students' responses to EdTech must be sensitive to the contextual applicability of the phenomena being studied. This indicates that research of the nature presented here must continue to occur over time as well as across learner groups, if we intend our EdTech systems to be effective for all of the learners using them.

The psychological constructs studied in this work - like others that are important to EdTech design - are complex and nuanced. Even though many motivational constructs are studied individually, in practice, they are most likely to co-occur and interact with each other (Renninger et al., 2018). This paper does not represent a complete or comprehensive study on the role of motivation and help-seeking in student performance and identity. However, its finding indicates the importance of future work to consider broader social factors around these constructs when incorporating them in design. In this work, we assume that an outcome is inequitable if particular student groups are observed to be advantaged or disadvantaged by the system usage. We acknowledge that fairness can be conceptualized in other ways too. We hope that our work can contribute to the emerging discussion on fairer EdTech research, design, and development.

3.8. Conclusion

In order to close the opportunity gap, we must improve the learning experiences for *all* students who use EdTech systems. This study attempts to answer calls to be socially responsible and accountable in Ed Tech (e.g., Porayska-Pomsta & Rajendra, 2019), which

has historically shown considerable social distance between its developers and the students that they want to serve. In order to identify the potential blind spots that lead to inequitable student outcomes, we suggest a method for explicitly identifying the varied needs of student subgroups even when data is unavailable at the student level. We make these recommendations within the context of help-seeking behaviors, which are an important part of self-regulation, and student autonomy more broadly, but which is also a behavior that may be particularly susceptible to cultural differences (i.e., Ogan et al., 2015). Intelligent Tutoring Systems like Reasoning Mind Foundations provide an opportunity for students to practice self-regulation by taking control over their choices in the learning environment. Help-seeking is a particularly relevant SRL process within this type of learning system, given the prominence of hints in EdTech systems. Similarly, student autonomy plays an important role in the development of their interest and motivation in learning. In this paper, we demonstrate that school-level demographics can have a significant influence on the relationships between students' help-seeking behavior, their motivations, and student outcomes. In doing so, we question the implicit assumption that complex constructs like these can be considered without also considering student context. This calls for greater consideration within our field of social, cultural, and economic influences on student choices and outcomes (cf. Baker et al., 2019).

Amidst the mixed results from empirical studies on the effectiveness of hints, Aleven and colleagues (Aleven et al., 2016) continue to recommend the use of hints in EdTech systems and suggest making four key methodological distinctions when studying interventions

designed to promote help-seeking - (1) effects on learning in the same learning environment versus a new environment; (2) effects on current learning versus future learning; (4) effects on learning in the same domain versus another; (3) effects on SRL processes versus domain-level learning. We propose to extend upon the list of these methodological considerations, suggesting that researchers also (5) explore the effects of help-seeking designs in one demographic context versus another. Similarly, we recommend EdTech designers and researchers consider the role of students' demographic contexts while making design choices to motivationally enhance their systems.

This is not to say that there are not both practical and definitional issues in doing so. However, as we can see that such demographic effects are present even within a single U.S. state (albeit one of the larger and more diverse U.S. states), it is worth considering the ways in which different groups of people may attach different meanings to the behavior of help-seeking and of intrinsic and extrinsic motivations. For instance, research should consider the ways in which help-seeking might be interpreted as an imposition or as an admission of failure or how value is attached to autonomy over control for extrinsic motivation, since, as we discussed in Section 3.2, these interpretations likely vary from one culture to another. By considering demographics in our research on these constructs - and on SRL in general - we increase the likelihood that our findings will apply to the full diversity of learners using EdTech and related systems today.

BIBLIOGRAPHY

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), Proceedings of the 5th *International Conference on Intelligent Tutoring Systems*, ITS 2000 (pp. 292–303). Berlin: Springer.

Aleven, V., & Koedinger, K. R. (2001). Investigations into help-seeking and learning with a Cognitive Tutor. In R. Luckin (Ed.), Papers of the AIED-2001 *Workshop on Help Provision and Help-Seeking in Interactive Learning Environments* (pp. 47–58).

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101-128.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(1), 205- 223.

Almeda, M. V. Q., Baker, R. S., & Corbett, A. (2017). Help Avoidance: When students should seek help, and the consequences of failing to do so. In Meeting of the *Cognitive Science Society* (Vol. 2428, p. 2433).

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science*, 13(4), 467–505. doi 10.1016/0364-0213(89)90021-9.

Amabile, T. M., DeJong, W., & Lepper, M. R. (1976). Effects of externally imposed deadlines on subsequent intrinsic motivation. *Journal of Personality and Social Psychology,* 34, 92–98.

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science*, 13(4), 467–505. doi:10.1016/0364-0213(89)90021-9.

Anderson, J. R. (1993). Rules of the Mind. Hillsdale: Lawrence Erlbaum Associates.

Arroyo, I., Beck, J., Woolf, B. P., Beal, C. R., & Schultz, K. (2000). Macro adapting animal watch to gender and cognitive differences with respect to hint interactivity and symbolism. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), Proceedings of the 5th *International Conference on Intelligent Tutoring Systems*, ITS 2000 (pp. 574–583). Berlin: Springer Verlag. doi 10.1007/3-540-45108-0_61.

Attewell, P. (2001). "Comment: The First and Second Digital Divides." *Sociology of Education* 74(3): 252–259.

Baker, R. S. J. d., Gowda, S. M., & Corbett, A. T. (2011). Towards predicting future transfer of learning. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), Lecture Notes in Computer Science: *Artificial intelligence in Education: 15th International Conference*, AIED 2011 (Vol. 6738, pp. 23– 30). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-21869-9_6.

Baker, R.S., Ogan, A.E., Madaio, M., Walker, E. (2019) Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context, 1*(1), 1-13.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122.

Baltes, M. M. (1997). The many faces of dependency. New York: Cambridge University Press.

Beck, J. E., Chang, K., Mostow, J., & Corbett, A. T. (2008). Does help help? Introducing the Bayesian evaluation and assessment methodology. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), Proceedings of the 9th *International Conference on Intelligent Tutoring Systems*, ITS 2008 (pp. 383–394). Berlin: Springer.

Blanchard, E. G., & Mizoguchi, R. (2008). Designing culturally-aware tutoring systems: towards an upper ontology of culture. *Culturally aware tutoring systems (CATS)*, 23-34.

Bong, M., Skaalvik, E. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psych Review*. 15(1), pp. 1-40.

Brown, S. D., & Lent, R. W. (2006). Preparing adolescents to make career decisions: A social cognitive perspective. *Adolescence and Education*, 5, 201-223.

Butler, R. (1998). Determinants of help seeking: Relations between perceived reasons for classroom help-avoidance and help-seeking behaviors in an experimental context. *Journal of Educational Psychology*, *90*(4), 630.

Butler, R. (2006). An achievement goal perspective on student help seeking and teacher help giving in the classroom: Theory, research, and educational implications. *Help seeking in academic settings: Goals, groups, and contexts*, 15-44.

Chambers, T. V. (2009). The" Receivement Gap": School Tracking Policies and the Fallacy of the" Achievement Gap". *The Journal of Negro Education*, 417-431.

Chavajay, P., & Rogoff, B. (2002). Schooling and traditional collaborative social organization of problem solving by Mayan mothers and children. *Developmental psychology*, *38*(1), 55.

Childs, D. S. (2017). Effects of Math Identity and Learning Opportunities on Racial Differences in Math Engagement, Advanced Course-Taking, and STEM Aspiration. PhD Dissertation. Temple University.

Connell, J. P., & Wellborn, J. G. (1990). Competence, autonomy and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *The Minnesota symposium on child psychology* (Vol. 22, (pp. 43–77). Hillsdale, NJ: Erlbaum.

Crossley, S. A., Karumbaiah, S., Ocumpaugh, J., Labrum, M. J., & Baker, R. S. (2020). Predicting Math Identity Through Language and Click-Stream Patterns in a Blended Learning Mathematics Program for Elementary Students. *Journal of Learning Analytics*, *7*(1), 19-37.

Deci, E. L., & Cascio, W. F. (1972, April). Changes in intrinsic motivation as a function of negative feedback and threats. Presented at the meeting of the *Eastern Psychological Association*, Boston.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125, 627-688.

Doroudi, S., & Brunskill, E. (2019, March). Fairer but not fair enough on the equitability of knowledge tracing. In Proceedings of the 9th *International Conference on Learning Analytics & Knowledge* (pp. 335-339).

Drachsler, H., Hoel, T., Scheffel, M., Kismihók, G., Berg, A., Ferguson, R., Chen, W., Cooper, A., & Manderveld, J. (2015). Ethical and privacy issues in the application of learning analytics. In *LAK'15* (pp. 390-391). ACM.

Else-Quest, N. M., Mineo, C. C., & Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gender and ethnicity. *Psychology of Women Quarterly*, 37(3), 293-309.

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013). The effects of culturally congruent educational technologies on student achievement. In *International Conference on Artificial Intelligence in Education* (pp. 493-502). Springer, Berlin, Heidelberg.

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap?. *The High School Journal*, *91*(1), 29-42.

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. World Policy Journal, 33(4), 111-117

Goldin, I. M., Koedinger, K. R., & Aleven, V. (2012). Learner differences in hint processing. In K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), Proceedings of the 5th International Conference on *Educational Data Mining* (EDM 2012) (pp. 73–80). Worcester: International Educational Data Mining Society.

Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology*, 93, 3–13.

Green, B. & Hu, L. (2018). The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *Proceedings of the International Conference on Machine Learning: The Debates Workshop.*

Greenbaum, P. E., & Greenbaum, S. D. (1983). Cultural differences, nonverbal regulation, and classroom interaction: Sociolinguistic interference in American Indian education. *Peabody Journal of Education*, *61*(1), 16-33.

Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology,* 52, 890–898.

Hansen, J. D., & Reich, J. (2015). Democratizing education? Examining access and usage patterns in massive open online courses. *Science* 350, 6265 (2015), 1245–1248.

Harackiewicz, J. (1979). The effects of reward contingency and performance feedback on intrinsic motivation. *Journal of Personality and Social Psychology,* **37**, 1352–1363.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of educational research*, *70*(2), 151-179.

Hoffmann, L. & Häussler, P. (1998). An intervention project promoting girls' and boys' interest in physics. In L.

Hoffmann, A. Krapp, K.A. Renninger, & J. Baumert (Eds.), *Interest and learning* (pp. 301–316). Kiel, Germany: IPN.

Hogg, M. A. (2000). Subjective uncertainty reduction through self-categorization: A motivational theory of social identity processes. *European review of social psychology*, 11(1), 223-255.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need? In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM.

Holstein, K., & Doroudi, S. (2019). Fairness and equity in learning analytics systems (FairLAK). In *Companion Proceedings of the Ninth International Learning Analytics & Knowledge Conference (LAK 2019)*.

Howley, I., Kanda, T., Hayashi, K., & Rosé, C. (2014). Effects of social presence and social role on help-seeking and learning. In G. Sagerer, M. Imai, T. Belpaeme, & A. Thomaz (Eds.), HRI '14: Proceedings of the 2014 *ACM/IEEE International Conference on Human-robot Interaction* (pp. 415–422). New York: ACM. doi:10.1145/2559636.2559667.

Huang, X., Craig, S. D., Xie, J., Graesser, A., & Hu, X. (2016). Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. *Learning and Individual Differences*, *47*, 258-265.

Hudley, A. H. C., & Mallinson, C. (2015). Understanding English language variation in US schools. Teachers College Press.

Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2016). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*, *109*(3), 387–404. doi:10.1037/edu0000146

Jackson, G. T., Boonthum, C., & McNamara, D. S. (2009). iSTART-ME: Situating extended learning within a game-based environment. In In Proceedings of the Workshop on Intelligent Educational Games at the 14th Annual Conference on *Artificial Intelligence in Education* (pp. 59-68).

Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2019). The influence of school demographics on the relationship between students' help-seeking behavior and performance and motivational measures. *Educational Data Mining (EDM)*, 4, 16.

Karumbaiah, S., Lan, A., Nagpal, S., Baker, R. S., Botelho, A., & Heffernan, N. (2021, April). Using Past Data to Warm Start Active Machine Learning: Does Context Matter? *International Learning Analytics and Knowledge Conference*, 151-160.

Khachatryan, G. A., Romashov, A. V., Khachatryan, A. R., Gaudino, S. J., Khachatryan, J. M., Guarian, K. R., & Yufa, N. V. (2014). Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education*, 24(3), 333-382.

Kimble, G. A. (1987). The scientific value of undergraduate research participation. *American Psychologist,* 42(3), 267- 268.

Klassen, R. M. (2004). Optimism and realism: A review of self-efficacy from a cross-cultural perspective. *International Journal of Psychology*, 39(3), 205-230.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3), 239-264.

Koestner, R., Ryan, R. M., Bernieri, F., & Holt, K. (1984). Setting limits on children's behavior: The differential effects of controlling versus informational styles on intrinsic motivation and creativity. *Journal of Personality,* 52, 233–248.

Ladson-Billings, G. (2013). Lack of achievement or loss of opportunity. *Closing the opportunity gap: What America must do to give every child an even chance*, *11*.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and individual differences*, *19*(3), 355-365.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic rewards: A test of the ''over justification'' hypothesis. *Journal of Personality and Social Psychology,* 28, 129–137.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of educational psychology*, *97*(2), 184.

Long, Y., & Aleven, V. (2013). Skill diaries: Improve student learning in an intelligent tutoring system with periodic self-assessment. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), Proceedings of the 16th *International Conference on Artificial Intelligence in Education,* AIED 2013 (pp. 249–258). Berlin Heidelberg: Springer. doi 10.1007/978-3-642-39112-5_26.

Mathews, M., Mitrović, T., & Thomson, D. (2008). Analysing high-level help-seeking behaviour in ITSs. In W. Nejdl, J. Kay, P. Pu, & E. Herder (Eds.), *Adaptive Hypermedia and Adaptive Web-based Systems*: 5th International Conference, AH 2008 (pp. 312–315). Berlin, Heidelberg: Springer. doi:10.1007/978-3-540- 70987-9_42.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., and Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2):397–416.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-Computer Interaction*, 5(4), 381–413. doi:10.1207/s15327051hci0504_2.

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology,* 88, 203–214.

Nelson-Le Gall, S., & Resnick, L. (1998). Help seeking, achievement motivation, and the social practice of intelligence in school. *Strategic help seeking: Implications for learning and teaching*, 39-60.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487-501.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. International Journal of *Artificial Intelligence in Education*, 25(2), 229- 248.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, J.M.A.L., Baker, R.S. (2020) Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining, 12* (3), 1-30.

Pardos, Z. A., & Heffernan, N. T. (2010, June). Modeling individualization in a bayesian networks implementation of knowledge tracing. In International Conference on *User Modeling, Adaptation, and Personalization* (pp. 255-266). Springer, Berlin, Heidelberg.

Porayska-Pomsta, K., & Rajendran, G. (2019). Accountability in Human and Artificial Intelligence Decision-Making as the Basis for Diversity and Educational Inclusion. In *Artificial Intelligence and Inclusive Education* (pp. 39-59). Springer, Singapore.

Razzaq, L., & Heffernan, N. T. (2010). Hints: Is it better to give or wait to be asked? In V. Aleven, J. Kay, & J. Mostow (Eds.), Lecture Notes in Computer Science: Proceedings of the 10th International Conference on *Intelligent Tutoring Systems*, ITS 2010 (Vol. 1, pp. 115–124). Berlin: Springer.

Reeve, J., & Deci, E. L. (1996). Elements of the competitive situation that affect intrinsic motivation. *Personality and Social Psychology Bulletin,* 22, 24–33.

Reich, J., & Ito, M. (2017). From good intentions to real outcomes: Equity by design in learning technologies. *Digital Media and Learning Research Hub*.

Renninger, K. A., Ren, Y., & Kern, H. M. (2018). Motivation, engagement, and interest:"In the end, it came down to you and how you think of the problem". In *International handbook of the learning sciences* (pp. 116-126). Routledge.

Roll, I., Baker, R. S. J. D., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, 23(4), 537–560. doi:10.1080/10508406.2014.883977.

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, *2*(4), 2332858416673968.

Ryan, A. M., Shim, S. S., Lampkins-uThando, S. A., Kiefer, S. M., & Thompson, G. N. (2009). Do gender differences in help avoidance vary by ethnicity? An examination of African American and European American students during early adolescence. *Developmental Psychology, 45*(4), 1152–1163.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology,* 43, 450–461.

Ryan, R. M., & Grolnick, W. S. (1986). Origins and pawns in the classroom: Self-report and projective assessments of individual differences in children's perceptions. *Journal of Personality and Social Psychology,* 50, 550–558.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology,* 57, 749–761.

Ryan, R. M., & Stiller, J. (1991). The social contexts of internalization: Parent and teacher influences on autonomy, motivation and learning. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement* (Vol. 7, pp. 115–149). Greenwich, CT: JAI Press.

Ryan, R. M., Stiller, J., & Lynch, J. H. (1994). Representations of relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem. *Journal of Early Adolescence,*, 14, 226–249.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54-67.

Schofield, J. W. (1995). *Computers and classroom culture*. Cambridge University Press.

Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. *Handbook of Competence and Motivation*, 85, 104.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). ACM.

Sheldon, K. M., & Kasser, T. (1995). Coherence and congruence: Two aspects of personality integration. *Journal of Personality and Social Psychology,* 68, 531–543.

Shih, B., Koedinger, K. R., & Scheines, R. (2008). A response time model for bottom-out hints as worked examples. In R. S. J. d. Baker, T. Barnes, & J. Beck (Eds.), Proceedings of the 1st International Conference on *Educational Data Mining*, EDM 2008 (pp. 117–126). Montreal, Canada.

Skaalvik, E. M., & Skaalvik, S. (2013). School goal structure: Associations with students' perceptions of their teachers as emotionally supportive, academic self-concept, intrinsic motivation, effort, and help seeking behavior. *International Journal of Educational Research*, *61*, 5-14.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510-1529.

Stamper, J., Barnes, T., & Croy, M. (2011). Enhancing the automatic generation of hints with expert seeding. *International Journal of Artificial Intelligence in Education*, 21(1–2), 153–167. doi:10.3233/JAI-2011-021.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52(6), 613.

Subotzky, S., & Prinsloo, P. (2011). Turning the tide: A socio-critical model and framework for improving student success in open distance learning at the University of South Africa. *Distance Education*, *32*(2), 177-193.

Tsai, Y. S., & Gasevic, D. (2017). Learning analytics in higher education---challenges and policies: a review of eight learning analytics policies. In *LAK'17.* ACM.

Tai, M., Arroyo, I., & Woolf, B. (2013). Teammate relationships improve help-seeking behavior in an intelligent tutoring system. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), Lecture Notes in Computer Science: *Artificial Intelligence in Education* (Vol. 7926, pp. 239–248). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-39112-5_25.

Tessler, R.C., and Schwartz, S.H. (1972). Help seeking, self esteem, and achievement motivation: an attributional analysis. *Journal of Personality and Social Psychology* 21, 3 (1972), 318-326.

Texas Education Agency. (2018a, October 4). State and School District Summary. Retrieved from http://www.texaseducationinfo.org/infopage/Summary_Report_Glossary.pdf

Texas Education Agency. (2018b, July). District Type Glossary of Terms. Retrieved from https://tea.texas.gov/acctres/analyze/1617/gloss1617.html#M ajor20Urban

The Texas Tribune. (2018, August 24). State and School District Summary. Retrieved from https://www.texastribune.org/2018/08/24/texas-schooldistricts-a-f-grades-takeaways/

Urdan, T., & Pajares, F. (Eds.). (2006). Self-efficacy beliefs of adolescents. IAP.

Usher, E. L., & Pajares, F. (2006). Sources of academic and self-regulatory efficacy beliefs of entering middle school students. *Contemporary Educational Psychology*, 31(2), 125-141.

Vaessen, B. E., Prins, F. J., & Jeuring, J. (2014). University students' achievement goals and help-seeking strategies in an intelligent tutoring system. *Computers & Education*, 72, 196- 208.

Vallerand, R. J., & Reid, G. (1984). On the causal effects of perceived competence on intrinsic motivation: A test of cognitive evaluation theory. *Journal of Sport Psychology,* 6, 94– 102.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3), 227–265.

Wang, Y., & Beck, J. (2013, July). Class vs. student in a bayesian network student model. In International Conference on *Artificial Intelligence in Education* (pp. 151-160). Springer, Berlin, Heidelberg.

Wood, H., & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, 33(2/3), 153– 169.

Wilkins, J. L. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education*, *72*(4), 331-346.

Williams, G. C., & Deci, E. L. (1996). Internalization of biopsychosocial values by medical students: A test of self-determination theory. *Journal of Personality and Social Psychology,* 70, 767–779.

Yudelson, M., Fancsali, S., Ritter, S., Berman, S., Nixon, T., & Joshi, A. (2014, July). Better data beats big data. In *Educational Data Mining* 2014.

Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, 37(1), 215-246.

Zeldin, A. L., Britner, S. L., & Pajares, F. (2008). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 45(9), 1036-1058.

Zimmerman, B. J. (1985). The development of "intrinsic" motivation: A social learning analysis. *Annals of Child Development*, 117-160. Greenwich, Conn. JAI.

CHAPTER 4

VARYING EFFECTIVENESS OF METHODOLOGICAL IMPROVEMENTS IN ANNOTATED DATA COLLECTION

Karumbaiah, S., Lan, A., Nagpal, S., Baker, R. S., Botelho, A., & Heffernan, N. (2021). Using Past Data to Warm Start Active Machine Learning: Does Context Matter? *International Learning Analytics and Knowledge Conference*.

Abstract. Despite the abundance of data generated from students' activities in virtual learning environments, the use of supervised machine learning in learning analytics is limited by the availability of labeled data, which can be difficult to collect for complex educational constructs. In a previous study, a subfield of machine learning called Active Learning (AL) was explored to improve the data labeling efficiency. AL trains a model and uses it, in parallel, to choose the next data sample to get labeled from a human expert. Due to the complexity of educational constructs and data, AL has suffered from the cold-start problem where the model does not have access to sufficient data yet to choose the best next sample to learn from. In this paper, we explore the use of past data to warm start the AL training process. We also critically examine the implications of differing contexts (urbanicity) in which the past data was collected. To this end, we use authentic affect labels collected through human observations in middle school mathematics classrooms to simulate the development of AL-based detectors of engaged concentration. We experiment with two AL methods (uncertainty sampling, L-MMSE) and random sampling for data selection. Our results suggest that using past data to warm start AL training could be

effective for some methods based on the target population's urbanicity. We provide recommendations on the data selection method and the quantity of past data to use when warm starting AL training in the urban and suburban schools.

4.1 Introduction

New forms of digital data captured in various learning settings have made it possible to build meaningful models to understand and optimize learning [18]. Interaction logs, and sensors, among others, make it easy to generate abundant data from students' learning activity [2]. Yet, in many cases, the modeling effort is limited by the availability of ground truth labels of complex educational constructs (which are used as target variables in supervised ML) related to student affect, cognition, behavior, and other sociocultural factors [18]. In some cases, it is possible to attain labels at little to no cost (e.g., college enrollment, in-system behavior, and test performance). But for other constructs, the success in using ML depends on human annotation that is time-consuming, expensive, and sometimes difficult depending on the pedagogical context and the complexity of the construct being modeled. For instance, if the data labels are collected in authentic settings like physical classrooms, then fieldwork opportunities are limited in time, resource-intensive, and involve several tedious tasks such as background verification of the observers, approvals from the school administration and institutional review boards, and obtaining written consent from students and parents. Even video replay coding still takes a substantial amount of time per label [25]. Thus, despite having an abundance of student activity data overall, the limitations in collecting human labeled data are pushing us to find

new ways to develop better performing ML models with a smaller amount of annotated data [31].

A potential solution lies in a subfield of ML called Active Learning (AL) that tries to learn a good model from fewer data samples by letting the ML model choose the data it trains from – thus, focusing the labeling efforts on a smaller subset of carefully selected data samples [28]. This subfield is not to be confused with the instructional approach of active learning in education [3]. In this paper, AL refers to the ML-based data selection algorithms aimed at improving the data labeling efficiency (discussed further in Section 4.2.1). AL works by training a model and choosing data iteratively: in each iteration, it first uses the current model to choose which data point to use next (i.e., on which data point to query for a human-generated label), and then uses the label to update the model. In contrast to simple classification problems (e.g., classifying apples from oranges in an image), complex settings like education pose some challenges to the adoption of AL. First, the labels could be highly subjective (e.g., self-reports of student emotions). Second, the data can be highly noisy (e.g., video data from a physical classroom). Third, the input feature set could be large (e.g., hundreds of features summarizing student activity in a virtual learning environment). Thus, the complexity involved in the ML tasks in application fields like education may require AL to seek significantly more samples to reach reasonable quality. As such, AL has found limited use in LA thus far, especially in the cold start situation, where the model doesn't have access to sufficient data yet [28].

In this paper, we investigate the use of past data on the same construct to overcome the cold-start problem when using AL. Using past data to warm start the AL training process, even for the same construct, may not be straightforward given the diversity in the student population [12]. Considerable research shows that demographic factors are often related to differences in educational outcomes [6]. Thus, we also examine whether the differing context of the past data used to warm start AL has an implication in building a model in the target population. This is important because LA models need to ensure population validity as they attempt to meet the needs of all students [22]. Given the feasibility challenges around collecting learning data in schools, the first data that is collected may in many cases come from convenience samples of middle-class students (see discussion in [14]) or another highly accessible student population where it is easy to collect labeled data. Since AL follows a greedy approach to optimize data collection, using data from a dominant student population could drive the model training process for a different population of learners to a suboptimal solution - a biased model. Hence, it is necessary to critically investigate the role of differing contexts of the past data if we want to use them to overcome the cold-start problem. As we still don't know what a "population" is [1], we focus our experiments in this study on one contextual dimension of urbanicity [cf. 22] to examine the use of past data from urban and suburban schools to warm start AL training in a school from the other context. Thus, our primary research questions in this study are –

- Does using past data help warm start the AL training process effectively?

- How does the urbanicity of the past data impact the effectiveness of the warm start process?
- How much past data from a different urbanicity is appropriate to use while warm starting the AL training process?

To this end, we use authentic affect labels collected through human observations in middle school mathematics classrooms to simulate the development of AL-based detectors of engaged concentration (described further in section 4.2), a common affective state among students. We experiment with two AL methods and one non-AL method (random sampling) for data selection. Our results suggest that using past data to warm start AL training could be effective for some methods. We also see that the urbanicity of the past data matters. We provide recommendations on the data selection method and the quantity of past data to use when warm starting AL training in the urban and suburban schools.

4.1.1 Contributions

Our primary contribution to AL research with education data is the critical analysis of the use of past data to overcome the cold start problem of AL training with complex constructs. More importantly, we show that mismatches in the urbanicity of the past data (and possibly other demographic dimensions) could be detrimental to effective model training in some cases.

4.2. Background

The next subsections provide a brief introduction to AL methodology, its use in label data collection for affect detection, and the importance of studying differing contexts in this paradigm.

4.2.1 Active Learning Algorithms

Supervised learning is a machine learning task that involves learning a function that maps the input (a set of feature values for a data point) to a predefined output (a target label of the data point). A commonly used function is a classifier that maps the input to a set of categories (class labels). In the typical supervised learning setup, all labeled data is collected before model development starts and available at training time for the model to learn from. On the other hand, in an AL setting, data collection and model training occur concurrently. The label collection process is iterative since all or a relevant subset of the training data collected thus far is available in real-time to make a choice on which point will get labeled next. AL methods are used in a scenario where there is limited opportunity to obtain labeled data – typically, when one can only selectively label a small subset of an otherwise abundant unlabeled data. The goal of AL algorithms is to enable training a high-quality classifier with fewer data samples by selecting those that are the most informative to the classifier. Thus, as the training of a model progresses, the AL algorithm aims to select the next data point to obtain a label for, such that it will be the most informative for the current model and hopefully lead to the largest improvement in its predictive power [28]. Several metrics of informativeness have been explored in AL research such as entropy (or observation uncertainty) [19], expected error reduction [26], expected variance

reduction [32], and model change [5]. A suite of algorithms has also shown promising results with a range of classifiers from logistic regression [30] to deep convolutional neural networks [27]. In this paper, we compare the following three approaches (two AL methods and one non-AL method) that have previously been applied to affect detection [31].

4.2.1.1 Uncertainty Sampling (UncS)

Uncertainty sampling is the simplest and most commonly used AL method and has been shown to achieve comparable or even better performance than other more sophisticated AL methods on real-world data [30]. It uses the prediction entropy of the model's predictive distribution over each possible class label to quantize the informativeness of each data point. Therefore, in each iteration, it takes the current model and predict the label distribution of each unlabeled data point and selects the one that the model is the least certain of, i.e., the data point that has the highest predictive entropy under the current model.

4.2.1.2 The (L-MMSE)-based method

One limitation of the UncS method is that the accuracy of its notion of data informativeness, i.e., model uncertainty, is highly associated with the quality of the current model. Therefore, when the model only has access to limited data, this estimate of uncertainty may not be accurate. The Linear Minimum Mean Square Error (L-MMSE) Estimator, first proposed in [16, 17], provides a set of closed-form approximations of the estimation error (a proxy of model uncertainty), which is shown to be highly accurate when

the number of data points is small. Therefore, the L-MMSE-based AL method [31] selects the next data point as the one that leads to the maximum reduction of the MSE. Roughly speaking, it looks at how similar each unlabeled data point (behavior) is to previous labeled data points the model has seen, which means we want to label the next data point that looks the most like an outlier. It is shown to mostly outperform UncS for student affect detection, especially when the number of data points is small [31].

4.2.1.3 Random Sampling (Random)

We also use a third, non-AL method for data selection, which is simply to randomly select an unlabeled data point from all possible points.

4.2.2 Student Affect Detection

Affective computing is an important area of interest in LA due to the close connection between a student's affect and their learning and experience. Affect has been shown to correlate with important educational constructs like self-efficacy [20], motivation [24], and learning [7]. Accordingly, affect-sensitive interventions have been designed in virtual learning environments to improve students' learning [8], and overall experience [10]. Thus, several research studies in the past decade have focused on building good quality affect detectors using physical and physiological sensors [21], and interaction log data [10] - the latter being the more affordable, less intrusive, and scalable option. Sensor-free affect detectors are classifiers categorizing a set of student interaction features into a predefined set of student affective states such as confusion, frustration, boredom, and engaged

concentration. The features are distilled from the interaction log data which is easily available in most virtual learning environments. However, the affect labels required for supervised training involves a labor-intensive data collection process.

One commonly used approach to collect labels for affect is through field observations in a real classroom by certified expert coders. A frequently-used technique for classroom observations is the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; [23]) - an affect coding protocol wherein students are observed by certified coders in a round-robin fashion. Each observation lasts up to 20 seconds. The affective state labels are boredom, confusion, frustration, and engaged concentration. During coding, some observations are labeled "NA", corresponding to the cases where i) the student could not be observed, ii) their affective state was unclear to the observer, or iii) they were in an affective state other than the states being coded. Following common practice in other detector development work [e.g. 22], we do not use these "NA" cases in our analysis.

4.2.3 Active Learning for Affect Detection

In a typical BROMP-based affect label collection, the students are observed in a pre-defined order. Thus, it is likely that the observers will miss opportunities to observe more informative cases. This could be an inefficient use of the limited time of the expert coders in an already short time window of fieldwork. Finding efficient ways to collect affect data is necessary as there are several constraints to conducting fieldwork in real classrooms (discussed in Section 4.1) along with the limited availability of the certified BROMP coders. AL provides an adaptive method to collect affect labels by directing the observers'

attention to the more informative cases. For the first time in 2019, Yang and colleagues [31] investigated the use of AL to collect higher quality data for training affect detectors with fewer data samples. We will be adopting a similar experimental protocol as [31], which is the standard in most AL studies (elaborated in Section 4.4.1).

In addition to experimenting with the existing AL methods, Yang and colleagues [31] also proposed the new method of L-MMSE, which appears to be particularly suited for the affect data collection setting where the data is small and noisy. Their results suggest that, when compared to other AL methods, L-MMSE leads to efficient modeling i.e., high-quality sensor-free affect detectors with fewer labeled data. By letting the model pick the next observation to learn from, the AL models were able to reach a desirable performance with as little as 70 observations which would translate to around 20 minutes of field observations with BROMP. This could tremendously reduce the burden on human labeling. However, before we adopt this methodology in our data collection practices, we also need to critically examine any possible biases that the model could have picked while greedily choosing the next best data sample to get labeled.

4.2.4 Role of Student Population in Affect Detection

In the previous study, the empirical analysis was conducted on combined data from multiple schools in different urbanicities (urban, suburban, and rural). In the current study, we split the data based on the urbanicity to assess any potential discrepancies in using models trained on one population to test in another population. This is important because student demographics are known to influence several aspects of affect [13]. Differences in

culture are known to influence variation in beliefs and personal dispositions towards emotional expression and moderation [29], and the frequency and emergence of certain affective states [15]. A recent study synthesizing results across multiple affect datasets showed that affective patterns seem to differ based on the country in which the data was collected (US versus Philippines; [13]). We chose to explore urbanicity as a contextual dimension in this study because past work suggests that affect detectors do not always transfer well between urbanicity categories [22]. In this study, we want to examine if this result holds true in the AL paradigm, especially when using data from different urbanicity to warm start the AL training.

4.3 Data

We use a previously collected dataset from ASSISTments [4]– a computer based learning platform which allows teachers to assign content and monitor student performance while supplying students with immediate correctness feedback and on-demand supports in the form of hint messages and scaffolding [9]. The affect data was collected in middle school mathematics classrooms using BROMP (see Section 4.2.2). The dataset consists of 2511 affect observations for 367 students. For each observation, a set of 92 features is extracted from the log of the student's interactions with practice problems within ASSISTments. These features summarize student within- and across-problem behaviors in the 20-second interval of the affect observation, such as the number of hints they seek, time spent on solving problems, the accuracy of responses, etc. In this paper, we will study affect detection for engaged concentration, a binary classification model (engaged concentration

vs other affective states). Consistent with past studies in other learning systems [11], engaged concentration has the highest incidence among all the affective states in this dataset. However, the rate differs significantly between the two urbanicities - 93.85% among suburban students and 56.06% among urban students. Affect data were collected in schools located in northeastern US - 1772 observations from 153 students in three suburban schools and 755 observations from 222 students in one urban school [22]. All the schools are non-charter and non-magnet public schools. The original dataset also had three rural schools that we do not include in the current analysis due to data quality issues; past research with ASSISTments data in a similar context reported that affect detection models (without AL) generalize better between suburban and urban students than rural students [22]. In this paper, we would like to investigate if this property holds true between urban and suburban data when data from students in one urbanicity is used to warm start the AL-driven affect detector training process for students in the other urbanicity.

4.4 Analysis and Results

In this section, we examine the effectiveness of using past data in the initial batch to warm start the AL training process. In addition, we investigate the impact of using a mismatching student population in the initial batch used for model development for differing initial batch sizes. We present the results for the experiments we conducted to answer the following research questions –

1. Does using past data help warm start the AL training process effectively?

2. How does the urbanicity of the past data impact the effectiveness of the warm start process?
3. How much of the past data from a different urbanicity is appropriate to use while warm starting the AL training process?

4.4.1 Active Learning Experimental Design

As detailed in Section 4.2.1, we use three different approaches: 1) the linear minimum mean square error (L-MMSE)-based method [31], 2) uncertainty sampling (UncS), and 3) random sampling (Random). The first two are AL methods. We perform a train-validation-test split of the full dataset (70%-10%-20% ratio) at the student level i.e., the instances corresponding to an individual student are all in a single split. We use a simple logistic regression-based affect detector in all the experiments since it performs well and makes it possible to use all AL methods [31]; other more advanced affect detectors are not compatible with many AL methods. We use the standard area under the receiver operating characteristic curve (AUC) as the performance metric. The first step of the AL training is to select an initial batch from the training set. The initial batch size is a variable of interest in this research and we vary it based on each individual experiment (details in subsections below). Using the observations and affect labels in the initial batch, a base classifier is trained. We train our affect detectors using gradient descent and stop training as soon as performance on the validation set stops improving. Next, we select a data point for each AL method from the remaining training set based on its feature values. The model is re-trained with the selected data point, and the AUC is calculated using the test set. This

process is repeated for 70 new observations. Each experiment is repeated 100 times by splitting the dataset randomly into train-validation-test sets and randomly selecting an initial batch from the training set each time. The plots presented in the results section contain the average AUC across the 100 random splits.

4.4.2 Baselines (Experiment Set #0)

We report baseline performances on two testing setups: i) test set drawn from only urban students, and ii) test set drawn from only suburban students, with three training setups each: i) training set drawn from only urban students, ii) training set drawn from only suburban students, and iii) training set drawn from both urban and suburban students) - leading to a total of six train-test setups. To ensure a fair comparison across the urbanicities, we match the randomly chosen test sets across the 100 runs for all three training setups. We ran the following two sets of baseline models -

1. *Full data model (without AL)* - For each of the six train-test setups mentioned above, a logistic regression model is trained using all the data in the training set. This baseline represents the typical scenario where we collect the full data without using AL to optimize the label data collection process. It is the best-case scenario in terms of having all the data that can practically be collected given the resource constraints.

2. *AL without warm start* - AL algorithms without a warm start. This baseline represents the scenario where we choose to disregard any past data we have collected for the same

construct. Instead, we collect new data using AL. We run this for all the three approaches - L-MMSE, US, and random.

Table 4.1 Baseline test performances (measured by mean AUC across 100 random splits) for the six train-test setups of the logistic regression model with full data (without AL) and AL training without warm start. We report the performance of the AL algorithms without a warm start for L-MMSE, US, and random at the last iteration of AL training for.

Exp#	Training Set	Test Set	Full data model (without AL)	AL Without Warm Start		
				L-MMSE	UncS	Random
Testing on Suburban Students						
1	Suburban	Suburban	0.628	0.617	0.607	0.578
2	Urban	Suburban	0.583	0.548	0.581	0.585
3	Urban + Suburban	Suburban	0.652	0.649	0.631	0.599
Testing on Urban Students						
4	Urban	Urban	0.657	0.617	0.645	0.652
5	Suburban	Urban	0.663	0.583	0.582	0.607
6	Urban + Suburban	Urban	0.638	0.626	0.639	0.638

In Table 4.1, Figure 4.1, and Figure 4.2 we present the baseline performances (measured by average AUC) on the held-out test sets of a logistic regression model on full data (without AL) and AL models without a warm start.

Full data model (without AL). When compared to the within-urbanicity performance, the between-urbanicity transfer is relatively better for suburban -> urban (Table 4.1, exp# 5 vs exp# 4) than urban -> suburban (Table 4.1, exp #2 vs exp #1). For testing with suburban data, the model trained on urban data (different urbanicity) has 0.045 AUC value less (0.583 vs 0.628) than the one trained on suburban data (same urbanicity). In contrast, for testing with urban data, the model trained on suburban data (different urbanicity) has 0.006 AUC value more (0.663 vs 0.657) than the one trained on urban data (same urbanicity).

The better transferability of the model trained with suburban data to urban data could be due to the higher diversity within suburban data (from three different schools) as compared to urban data (from a single school). The three suburban schools may also vary in terms of the teacher practices and use of the system. The best performing model for the suburban data is the one trained on the combined dataset (urban+suburban). By contrast, the best performing model on the urban data is the one trained on the suburban dataset. We see that the models trained on the full data transfer well between urbanicities for testing on urban students. The models tested on urban data have similar performances with a 0.025 difference in AUC value between the best (0.663) and worst-performing (0.638) models. The AUC value difference between the best (0.583) and worst-performing (0.652) models is higher for the suburban test data at 0.069.

AL Without Warm Start. The AL models have a relatively larger difference in performances across the six train-test setups. The urbanicity mismatch in training and testing sets hurts test performance in all the three approaches. A model trained using the data from a single urbanicity does not transfer well when tested on the other urbanicity. For testing with suburban data (Table 4.1, exp# 1-3), training with combined data leads to a better performance for all three approaches. This observation is consistent with the pattern in the models trained on the full data without AL. The best performing AL model (L-MMSE) has only 0.003 less AUC than the best model trained on full data without AL (0.649 vs 0.652). Note that random sampling does slightly better than the full data model when a model trained using urban data is used with a suburban test set (0.585 vs 0.583).

One possible explanation is that random sampling learns from a smaller subset (50 samples) of training data from a mismatched urbanicity as compared to the model trained on the full urban data without AL (744 samples) – potentially reducing its generalizability to suburban students.

For testing with urban data (Table 4.1, exp# 4-6), a similar pattern of better performance with combined data is seen only for L-MMSE. For UncS and random, the best performing model is the one trained on urban data (same urbanicity). In contrast to the full data model without AL, the model trained on suburban data leads to a worse performance in urban data for all the three approaches. Among the three approaches, random sampling has the best performance at an AUC of 0.652 which is only 0.011 less than the full data model without AL.

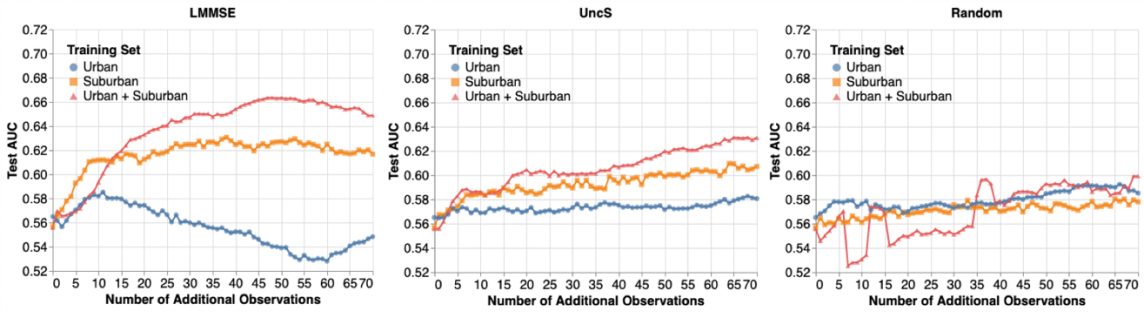


Figure 4.1 Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms trained on different training sets without warm start and tested on suburban data.

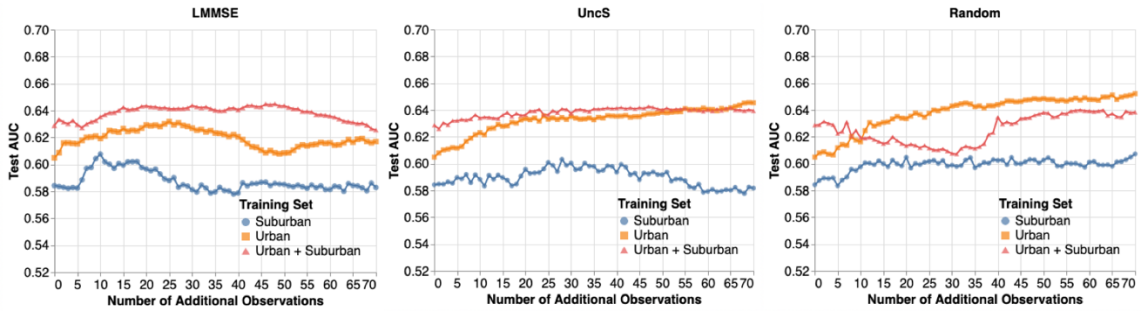


Figure 4.2 Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms trained on different training sets without warm start and tested on urban data.

4. 4.3 Within-Urbanicity warm start (Experiment Set #1)

In this set of experiments, we warm start the AL training process with data based on schools from the same urbanicity as the test school. We train and test AL algorithms on a single school but take the initial batch data from other schools in the same urbanicity. Since our data is from one urban school and three suburban schools, we will run these experiments only with the suburban data. Also, we have 1598 observations from one suburban school and only 103 and 68 observations from the other two suburban schools, which is not sufficient to have diverse enough random splits between training and test sets across the 100 runs of AL training – leading to unreliable evaluation. Hence, we will be running the AL models only on the school with 1598 observations with initial batches drawn from the other two schools. For comparison, we also report results for the experiments where the initial batch is drawn from the same school (no warm start) for the same test sets. Since the total number of students in the smallest suburban school is 68, we limit the initial batch size to 68 across all experiments for consistency. In cases where the school has more than

68 samples, we randomly sample 68 observations for the initial batch. These experiments represent the scenario where we choose to use the past data collected in a different school(s) from the same urbanicity to warm start the AL training. We run these experiments to answer our first research question on the effectiveness of using past data as an initial batch to warm start the AL training. The results should help us decide if we want to use past data from a similar student population (schools in the same urbanicity) to warm start the AL training process.

The three plots in Figure 4.3 present the results for the three approaches after a within-urbanicity warm start for a single suburban school (described in Section 4.4.3). Each plot has one line for the same school warm start (blue circles - School A), two lines for warm start with two other suburban schools (orange squares - School B and red triangles- School C), and one line for warm start with the combined data from the two other suburban schools (green pluses - Schools A&C). As one would expect, the same school warm start (blue circles) generally has a better performance for all the three approaches - close to the full data model without AL for suburban data (Table 4.1, exp# 1). The AL training starts at a high AUC value, potentially because the initial batch data (with 68 data samples) are all from the same school. The AUC doesn't improve with training and stays the same throughout (almost a straight line at AUC value around 0.62). This is not surprising as the AL algorithms are expected to do well with less data and we had enough data points from the same population in the initial batch to start with.

Despite the other two schools (B and C) being in the same urbanicity (suburban) as school A, using the data from these two schools to warm start AL training in school A leads to strikingly different results. The initial batch from one of the two schools (School B; orange squares) leads to a model performance that is consistently better than the other school (School C; red triangles) for all the three approaches. For the UncS approach to AL, warm start with school B quickly improves the AUC value and surpasses the same-school warm start with only 5 additional training samples from the target school. With random sampling, the AUC improvement is relatively slower as compared to the UncS approach. Nevertheless, the steady improvement eventually converges with the same-school warm start at the end of AL training with only 50 samples from school A as compared to 118 samples from school A for same-school warm start (68 in the initial batch + 50 during AL training). With L-MMSE, however, the improvement to AUC value saturates after 5 additional samples from school A, starts to dip slowly with more samples leading to a close to chance AUC value (~0.50) at the end of AL training. This raises concerns on using past data from a different school to warm start L-MMSE model training, even when the school is from the same urbanicity.

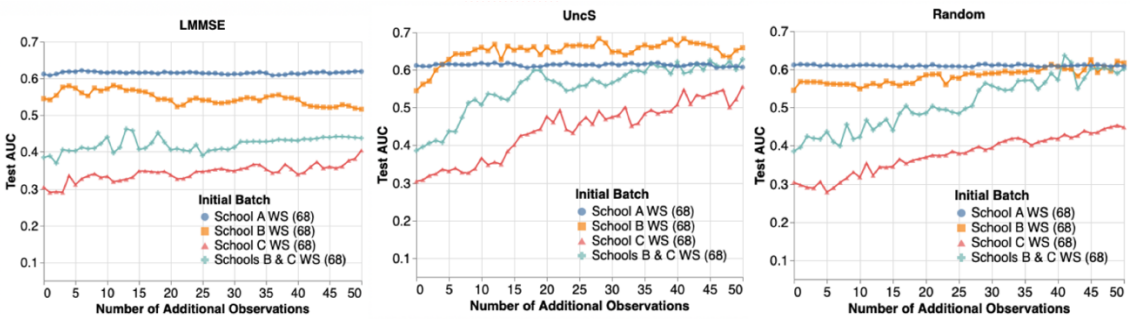


Figure 4. 3 Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms on a single suburban school (School A) with warm start data from the same and other suburban schools. In parenthesis in the legend are the initial batch sizes. WS = Warm Start.

The data from the other school (school C; red triangles) in the initial batch brings down the model performance severely (down by 0.30 AUC). In all the three approaches, the AUC value at the end of the training is below chance (<0.50), making the trained model inapplicable to the target population. The performance gradually improves as the AL training progresses for UncS and random sampling but fails to recover for L-MMSE. This observation raises questions on the robustness of the AL algorithms to out-of-context data during training - how quickly can L-MMSE recover when the new samples from the target population are introduced?

As one would expect, the initial batch with the combined data (green pluses) from the two other schools leads to a close-to-average performance compared to the two schools separately. With more observations, the combined data initial batch starts to improve steadily for the UncS and random (not for L-MMSE), and reaches a similar performance as the same-school warm start. The final performance for the UncS and random sampling is better than the AL without a warm start and is similar to the full data model without AL for suburban data (Table 4.1, exp# 1). Thus, there is some evidence supporting the use of combined data from multiple schools in the same urbanicity to warm start the AL training.

In summary, the within urbanicity warm start experiments suggest that not all schools in the same urbanicity have a similar effect when used to warm start the AL training process. Using a random sample from the combined data could be a better choice. More research

on the similarities and differences between the three schools on other demographic variables is needed to better understand the warm start process's differing implications. The UncS approach to AL and random sampling are seen to be more robust than L-MMSE in improving the model performance when new data from the target population is introduced. Overall, we recommend using past data from the same urbanicity, preferably from multiple schools, in warm starting some data collection approaches (UncS and random sampling, not L-MMSE).

4.4.4 Between-urbanicity warm start (Experiment Set #2)

Our next set of experiments are similar to the experiment set #1, except the initial batch comes from schools in a different urbanicity. Specifically, we train and test AL algorithms on student data from urban schools but draw the initial batch from suburban schools. Likewise, we run the experiment with training and test sets drawn from suburban data and initial batch from urban data. In these experiments, we take all the past data of the chosen urbanicity in the initial batch (the size is varied in the next subsection). These experiments correspond to the scenario where we choose to use the past data collected in schools from a different urbanicity to warm start the AL training. For comparison, we also report results for the within urbanicity warm start for the same test sets. We run these experiments to answer our second research question on the impact of urbanicity in using past data to warm start the AL training. The results should help us decide if we want to use past data from a different student population (in this case urbanicity) to warm start the AL training process.

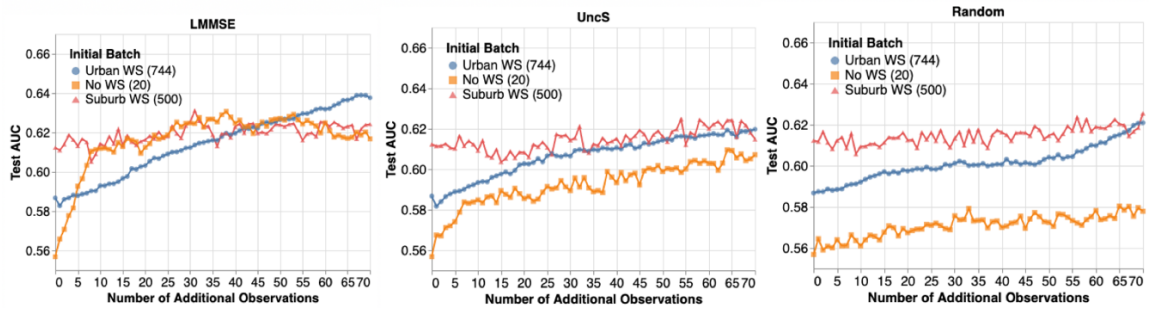


Figure 4.4 Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms on suburban student data with no warm start, warm start with past urban student data, and warm start with past suburban student data. In parenthesis in the legend are the initial batch sizes. WS = Warm Start.

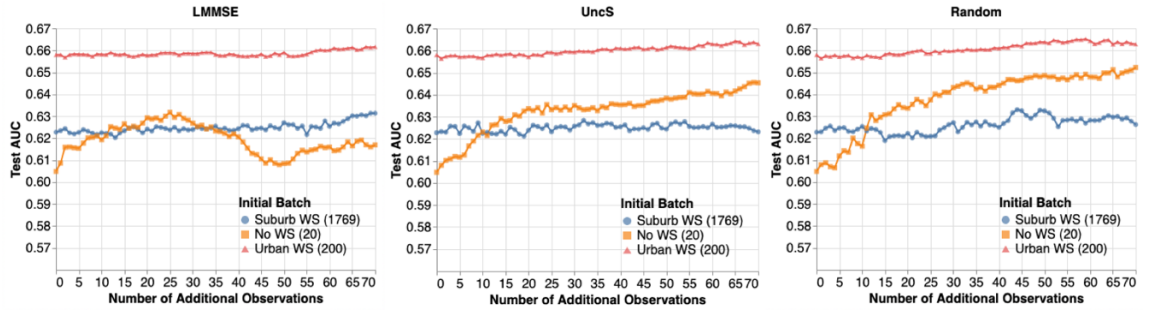


Figure 4.5 Comparing cross-validated performances of the L-MMSE, US, and random algorithms on urban student data with no warm start, warm start with past suburban student data, and warm start with past urban student data. In parenthesis in the legend are the initial batch sizes. WS = Warm Start.

The results for the between-urbanicity warm start is presented for two cases - a) test on suburban data (Figure 4.4), and b) test on urban data (Figure 4.5). In the case of the test on suburban data (Figure 4.4), using past data from urban schools to warm start the AL training (blue circles) leads to a better model performance when compared to the model trained with no warm start (orange squares). This observation is true for all three approaches. In fact, all the three approaches start at an AUC greater than 0.58 (above the chance value) with an initial batch drawn from the past urban data. This is in contrast with

like the last set of experiments where using past data from a different suburban school led to a performance below chance. The use of past data from urban schools gives a head-start of close to 0.02 AUC for the AL training with suburban data. Although the L-MMSE model without a warm start catches up after 20 new observations, the improvement in AUC saturates and stays at around 0.62, while the training with a warm start climbs up to 0.64. For UncS and random sampling, the progress in performance for both with and without a warm start is more gradual. Relative to the US, the gap widens further for random sampling as the training progresses and reaches 0.04 (0.58 without warm start vs. 0.62 with a warm start). As expected, a warm start using 500 samples from the same urbanicity (red triangles) leads to a peak performance right from the beginning of the AL training and shows little effect due to the new observations. This is not surprising because the AL algorithms are expected to do well with less data and we had enough data points (500) from the same population in the initial batch to start with. Unlike AL without a warm start, the between-urbanicity warm start catches up to this peak performance in all the three approaches - even exceeding it in the case of L-MMSE. L-MMSE with between-urbanicity warm start also manages to come close to the full data model without AL for suburban data, while random sampling only exceeds the AL without a warm start (Table 4.1, exp# 3). Overall, there is some evidence that using past data from urban schools effectively warm starts AL training in suburban schools.

The results are not as clear when testing on urban data (Figure 4.5). In comparison to AL training without a warm start (orange squares), the AL training with suburban data in the

initial batch (blue circles) performs better for L-MMSE and not for the UncS and random sampling. In all the three cases, the AL training starts at a higher AUC value (0.625 vs 0.605) with the suburban data in the initial batch but remains constant as new data samples are introduced. One possible explanation is that the initial batch data from a different urbanicity (suburban) outnumbers the additional samples from the new population (urban) and the model fails to improve its predictive power on the target population. After around 12 new samples, the model without a warm start exceeds the warm start model in its performance. However, as the training progresses, the L-MMSE performance for no warm start starts to decline, while the performance of UncS and random sampling rises steadily. Both UncS and random sampling models without a warm start exceed the full data model without AL (Table 4.1, exp# 6). In contrast, the performance of all the three models with between-urbanicity warm start doesn't meet both the baselines (AL without warm start and full data model without AL). Within-urbanicity warm start with urban data consistently leads to peak performance in all three models (red triangles). Overall, there is some evidence that using past data from suburban schools could be detrimental in warm starting AL training in urban schools.

4.4.5 Vary the initial batch size for the warm start (Experiment set #3)

In the experiment set #2, we took all the data from a chosen urbanicity as the initial batch. In this experiment set, we try to answer our third research question on how much past data from different urbanicity is appropriate to warm start the AL algorithms. We repeat the same experiments as before, varying the initial batch size stepwise. These experiments

represent the scenario where we may have a large amount of past data from a different urbanicity and need to find out how much data should be used to warm start the AL training process. The results should help us decide what amount of past data from the same or different student population (in this case urbanicity) is effective to warm start the AL training process.

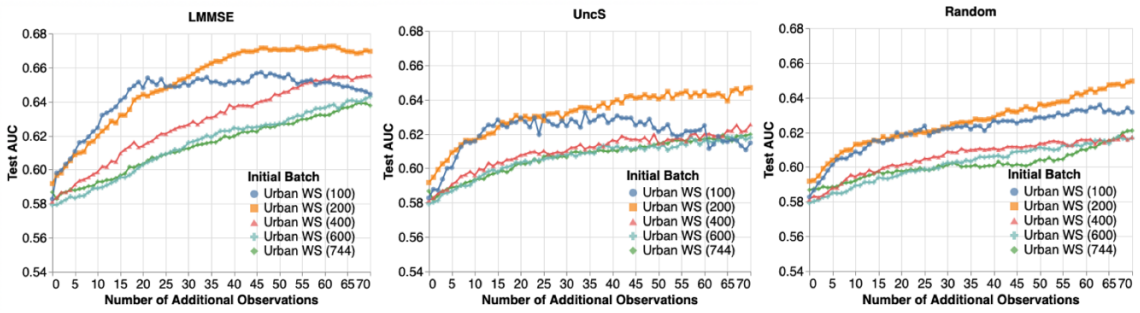


Figure 4.6 Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms on suburban student data with varying amounts of past urban student data for the warm start. In parenthesis in the legend are the varying initial batch sizes. WS = Warm Start.

4.5 Discussion

AL is a promising subfield of ML that can be used in LA to increase the efficiency of the label collection process that is time-consuming and requires extensive human effort in many cases. Model training and label data collection go hand-in-hand in AL, with the model iteratively choosing the most informative next data point to get labeled by a human. One of the challenges in the adoption of AL for education data is the cold-start problem when it is hard to accurately estimate the informativeness of a data point due to the lack of data at the early stages of the AL process [31]. In this paper, we have explored the use of

past data to overcome the cold-start problem seen for AL methods, along with the potential implications of differing student populations in the past data. Using an existing student affect dataset collected through human observations in middle school mathematics classrooms, we experimentally tested three training approaches (UncS, L-MMSE, Random) for the sensor-free detector of engaged concentration, studying performance within and across urbanicities (urban, suburban).

4.5.1 Summary of results

We conducted four sets of experiments to answer our research questions: how effective is AL i) without warm start, ii) using within-urbanicity warm start, iii) using between-urbanicity warm start, and iv) using varying batch sizes to warm start. Our results suggest the following. First, for all three approaches, training a model completely using data from a different urbanicity without warm start results in low detection accuracy in the target population [cf. 22]. Second, not all schools in the same urbanicity have a similar effect when used to warm start the AL training process. Using a random sample from the combined data (across suburban schools) could be a better choice when data from multiple schools are available. Third, using past data from urban schools effectively warm starts AL training in suburban schools. In contrast, using past data from suburban schools is detrimental to warm starting AL training in urban schools. One possible explanation is that the size of the suburban data is too large (1772 observations) compared to the urban data (755 observations); a trained model on suburban data may overfit to suburban students and loses generalizability to a different student population. Lastly, using the right (often small)

amount of past data can effectively warm start AL training if urbanicities do not match. When comparing the three approaches, we found that UncS and random sampling are more robust than L-MMSE in improving the model performance when new data from the target population is introduced during within-urbanicity warm start. With between-urbanicity warm start using a smaller batch size, it could help to switch from AL algorithms to random sampling after collecting some data points in the target population. In summary, with the experiments in this paper, we have shown that using past data to warm start some AL methods could be effective in training a good quality detector of engaged concentration (performance comparable to a detector trained using all data) with a fewer number of samples in some conditions and not so effective in others.

4.5.2 Implications for research

Our primary contribution to AL research with education data is the critical analysis of the use of past data to overcome the cold start problem of AL training with complex constructs. More importantly, we show that mismatches in the urbanicity of the past data (and possibly other demographic dimensions) could be detrimental to effective model training in some cases. As our AL modeling effort advances and finds innovative ways to improve model training with little data, it becomes essential to critically examine which data samples we are using. The need to consider human diversity in predictive modeling is becoming more apparent in LA research as the community moves to implement analytics solutions at a larger scale. If we aim to serve all students, we need to ensure the population validity of the models we build. This does not necessarily imply that all models must be within-

population (and, indeed, we do not entirely know what a population is [cf. 20]) – our findings suggest that there are better and worse ways to use data from other populations when building a model.

4.5.3 Implications for practice

Yang and colleagues [31] discuss the implications for data collection procedures when using AL in real classrooms. They present a brief design of a three-component system – i) an interface to record human observations, ii) a training paradigm to build a detector, and iii) an active learning method that connects the labeling and model training processes (see [31] for more details). In addition, there should be a provision for the expert coders to ignore the suggestion made by AL and use their intuition to pick the most informative cases when necessary. This agency could be important, especially if the AL-based model is picking up some unknown biases and leading to a suboptimal model training. Differences between AL recommendations and expert choices could also be valuable in conducting a post-hoc analysis of an AL approach's functioning. Also, even within a single class, there could be student subgroups that may end being under-observed by the AL recommendation. It is important that we need to set up conditions in AL recommendation to pick samples that are representative of these subgroups. Partnering with teachers will be a useful direction for identifying important subgroups in a specific class or school. Such research-practice partnership can help mitigate potential biases in data selection.

Although this work focuses on classroom observations, we could extend it to other forms of label data collection such as self-reports, video coding, and text replay coding. With

COVID-19 related school closures, we are currently exploring the use of AL in collecting labels through student self-reports. Since a student can be surveyed at any point in time, the observation window is not as strict as field observations. However, we must budget the surveys per student to not interfere with their learning or be too intrusive. Hence, our focus in using AL shifts from choosing which student to observe, to when and how often we survey each student. The data is likely to look different from classroom observations. It could have more missing data (e.g., student skips the survey) or be more noisy data (e.g., student responds incorrectly). In addition, the feature set for the AL algorithm will likely come from a longer time window when compared to being restricted to a single class period. Our next step is to collect some self-report data and conduct a similar analysis on warm starting AL training for the different student populations.

4.5.4 Limitations and future work

There are some limitations to our work presented in this paper. First, our experimental design does not consider the temporal nature of affect data collection in the real world. We choose the next most informative data point among all available data points in hindsight after they are already collected, while in the actual observation session, only a subset of these students and only the temporally close data samples will be available for human observation. Second, we have only experimented with the detector for engaged concentration, which is the most common affective state in our dataset. This work needs to be replicated with other important but relatively rarer affective states like boredom, frustration, and confusion. Third, due to data quality issues, we could not include rural

schools in this study. In general, further thought on categorizing urbanicity is warranted. Fourth, our mixed results on within-urbanicity warm start suggest that more research is need on the similarities and differences between the suburban schools on other demographic variables. Finally, we hope to explore more advanced AL methods to see if there are methods that respond better to the warm start condition than UncS and L-MMSE.

BIBLIOGRAPHY

[1] Baker, R.S., Ogan, A.E., Madaio, M., Walker, E. (2019). Culture in Computer-Based Learning Systems: Challenges and Opportunities. Computer-Based Learning in Context, 1 (1), 1-13.

[2] Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. Journal of Learning Analytics, 3 (2), 220-238.

[3] Bonwell. C. and Eison. J. (1991). Active Learning: Creating Excitement in the Classroom.Jossey-Bass.

[4] Botelho, A. F., Baker, R. S., and Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. In Proc. International Conference on Artificial Intelligence in Education, pages 40–51.

[5] Cai, W., Zhang, Y., Zhang, Y., Zhou, S., Wang, W., Chen, Z., and Ding., C. (2017). Active learning for classification with maximum model change. ACM Transactions on Information Systems, 36(2):15.

[6] Childs, D. S. (2017). Effects of Math Identity and Learning Opportunities on Racial Differences in Math Engagement, Advanced Course-Taking, and STEM Aspiration. PhD Dissertation. Temple University.

[7] D'Mello, S., Person, N., Lehman, B. (2009). Antecedent-Consequent Relationships and Cyclical Patterns between Affective States and Problem Sol- ving Outcomes. In International Conference on Artificial Intelligence in Education, 57-64.

[8] DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. International Journal of Artificial Intelligence in Education.

[9] Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. International Journal of Artificial Intelligence in Education.

[10] Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B. P., Arroyo, I., & Wixon, N. (2017). Addressing Student Behavior and Affect with Empathy and Growth Mindset. Proceedings of the 10th International Conference on Educational Data Mining.

[11] Karumbaiah, S., Andres, J. M. A. L., Botelho, A. F., Baker, R. S., & Ocumpaugh, J. S. (2018). The Implications of a Subtle Difference in the Calculation of Affect Dynamics. In 26th International Conference for Computers in Education.

[12] Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2019). The influence of school demographics on the relationship between students' help-seeking behavior and performance and motivational measures. Educational Data Mining (EDM), 4, 16.

[13] Karumbaiah, S., Baker, R. S., Ocumpaugh, J. & Andres, J. M. A. L. (2020). A Re-Analysis and Synthesis of Data on Affect Dynamics in Learning. Submitted.

[14] Kimble, G. A. (1987). The scientific value of undergraduate research participation. American Psychologist, 42(3), 267- 268.

[15] Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, emotion, and wellbeing: Good feelings in Japan and the United States. Cognition & Emotion, 14 (1), 93-124.

[16] Lan, A. S., Chiang, M., and Studer, C. (2018) An estimation and analysis framework for the Rasch model. In Proc. International Conference on Machine Learning, pages 2889–2897.

[17] Lan, A. S., Chiang, M., and Studer, C. (2018). Linearized binary regression. In Proc. Conference on Information Sciences and Systems, pages 1–6.

[18] Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017). Handbook of learning analytics. SOLAR, Society for Learning Analytics and Research.

[19] Lewis, D. D., and Gale, W. A. (1994) A sequential algorithm for training text classifiers. In Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12.

[20] McQuiggan, S. W., & Lester, J. (2009). Modeling affect expression and recognition in an interactive learning environment. International Journal of Learning Technology, 4 (3-4), 216-233.

[21] Nye, B. D., Karumbaiah, S., Tokel, S. T., Core, M. G., Stratou, G., Auerbach, D., & Georgila, K. (2018). Engaging with the scenario: Affect and facial patterns from a scenario- based intelligent tutoring system. In International Conference on Artificial Intelligence in Education (pp. 352-366). Springer, Cham.

[22] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), 487-501.

[23] Ocumpaugh, J. (2015). Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual.New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences, 60.

[24] Rodrigo, M. M. T., Anglo, E., Sugay, J., Baker, R. (2008). Use of unsupervised clustering to characterize learner behaviors and affective states while using an intelligent tutoring system. In International Conference on Computers in Education, 57-64.

[25] Rowe, E., Asbell-Clarke, J., Bardar, E., Almeda, M. V., Baker, R. S., Scruggs, R., & Gasca, S. (2020). Advancing Research in Game-Based Learning Assessment: Tools and Methods for Measuring Implicit Learning. In Advancing Educational Research With Emerging Technology (pp. 99-123). IGI Global.

[26] Roy, N., and McCallum, A. (2001). Toward optimal active learning through Monte Carlo estimation of error reduction. In Proc. International Conference on Machine Learning, pages 441–448.

[27] Sener, O., and Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In Proc. International Conference on Learning Representations, pages 1–13.

[28] Settles., B. (2012). Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning,6(1):1–114.

[29] Tsai, J., Levenson, R. (1997). Cultural influences on emotional responding: Chinese Am. & European Am. dating couples during inter-personal conflict. J. CrossCultural Psych., 28 (5), 600-25.

[30] Yang, Y., and Loog, M. (2016). A benchmark and comparison of active learning for logistic regression. arXiv preprint arXiv:1611.08618.

[31] Yang, T. Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active Learning for Student Affect Detection. International Educational Data Mining Society.

[32] Yu, K., Bi, J., and Tresp, V. (2006) Active learning via transductive experimental design. In Proc. International Conference on Machine Learning, pages 1081–1088.

CHAPTER 5

GENERAL DISCUSSION

This dissertation highlights the need for a nuanced approach in conducting bias research in adaptive and artificially intelligent learning systems. Although recent research has focused primarily on biases in downstream algorithmic aspects (see reviews in Baker & Hawn, 2021; Kizilcec & Lee, 2020), I hypothesized that the upstream components (e.g., theory, design, training data collection method) in the development of adaptive learning systems also contribute to the bias in these systems (Chapter 1). Towards this, I empirically demonstrate demographic disparities in three cases that are representative of the aspects that shape technological advancements in adaptive systems in education: non-conformance of data to a widely-accepted theoretical model (Chapter 2), differing implications of technology design on student outcomes (Chapter 3), and varying effectiveness of methodological improvements in annotated data collection (Chapter 4). In the following sections, I first detail the overall contribution of this dissertation by formally conceptualizing the upstream sources of bias (5.1). Then, I summarize the individual studies, highlight their contributions in identifying upstream biases, and discuss the implications to both bias research and to the design and development of adaptive systems (5.2). Finally, I present the potential future directions towards an equitable design and development of adaptive learning systems (5.3).

5.1 Conceptualization of Upstream Sources of Bias

The primary contribution of this dissertation work is the conceptualization of upstream sources of bias in adaptive learning systems. Adaptive and artificially intelligent learning systems have come under scrutiny for potential biases in their automated decision-making (see synthesis in Section 1.2). Several recent research projects have examined bias in algorithmic systems in education, focusing mostly on investigating bias in predictive modeling (Anderson et al., 2019; Bridgeman et al., 2009, 2012; Hu & Rangwala, 2020; Kai et al., 2017; Lee & Kizilcec, 2020; Ogan et al., 2015; Yu et al., 2020). This dissertation emphasizes the need to broaden the search for the sources of bias to those non-algorithmic aspects that may directly or indirectly influence the algorithm design. Inspired by Baker and Hawn's review (2021), I conceptualize these aspects as the "upstream sources of bias." I will first formally conceptualize upstream sources of bias within the standard workflow of machine learning algorithms and then present the broader impact of upstream biases in the design of adaptivity in learning systems.

In a standard workflow of machine learning algorithms, we start by collecting sample S from a probability distribution P (e.g., pairs of faces images and annotated affective states). Then, we choose a model class H (e.g., a set of neural networks). Using an algorithm or heuristic, we then find a model in this model class $h \in H$ which has the lowest prediction error $\hat{e}_S(h)$ on sample S. Lastly, we estimate the true error $e_P(h)$ of this model for any future data by testing it on new data that is also drawn from P. This methodology is justified by the fundamental theorem of machine learning on generalization (Kearns, 2021). In simple

terms, no matter how complicated the distribution P is, for a reasonable model class H, if we have enough data S, then for every model $h \in H$, $\hat{e}_S(h) \approx e_P(h)$, i.e., the error on current data is a close approximation of the error in the real-world. Simply put, if a model does well on the data in hand, it is expected to do well on any unseen data in the future.

While this approach may work well for predicting, say, the weather tomorrow, it falls short in serving diverse student populations, especially those who were underrepresented or unrepresented in sample S. As reported by Paquette and colleagues (2020) in their recent review of papers from educational data mining, student population information is often not reported with generalization estimates, making it hard to contextualize models. By design, most machine learning models are optimized for the majority populations because of their focus on reducing the overall error. This leads to a bias towards those who do not share similarities with the majority population. There are several reasons why this happens: 1) the data collected may be less accurate for some groups (e.g., earlier cameras were designed to bring out high contrast and better resolution for white skin color; Roth, 2009), 2) the construct of interest may look different in different subgroups (e.g., cultural differences in the expression of emotions; Tsai & Levenson, 1997), or 3) some groups may just be hard to predict (e.g., facial expression analysis for students wearing niqab). Identifying the serious issue of bias, several recent research projects have examined bias in algorithmic systems in education (see reviews in Baker & Hawn, 2021; Kizilcec & Lee, 2020). However, the focus has mostly been on investigating bias in downstream stages of predictive model development and evaluation. While this is an important step in the right

direction, I argue that the issue of bias arises well before the actual model development and that the current emphasis on downstream stages will limit the progress towards equity in adaptive learning systems if we don't broaden our search.

Towards this, I propose investigating what I call the "upstream sources of bias" that not only influence what sample (S) gets collected but also the conceptualization of true distribution (P). Figure 5.1 presents a visualization highlighting the three upstream sources of bias that were investigated in this dissertation, i.e., theory, design, and data collection method. As discussed in Section 1.3, data collection methods and system design directly impact what data gets collected and hence the variables used to design an algorithm. Moreover, theories impact system design, data collection, and algorithm design. Letting an algorithmic model perpetuate an upstream bias in its discriminatory decision-making may harm specific student subgroups, leading to outcomes that falsely confirm the upstream bias. Moreover, studying upstream bias is important not only because it could form a basis for bias in an algorithmic model but also because it could lead to direct discriminatory behaviors in an adaptive system where the design of adaptive interventions is based on a biased upstream source. By considering three upstream sources separately, this dissertation brought to attention those aspects that closely shape adaptive decision making but are often obscured in the evaluation of bias in an algorithmic model.

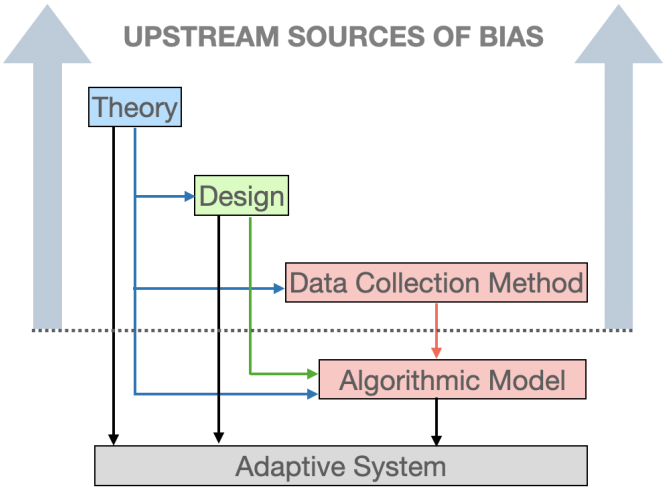


Figure 5.1 Upstream sources of bias studied in this dissertation that directly or indirectly shape the adaptivity of a learning system.

5.2 Summary of Studies, Contributions, and Implications

I study bias specifically in adaptive and artificially intelligent learning systems. These systems attempt to identify key moments in students' learning to adapt based on perceived student needs. Such automated decision-making often involves little to no human intervention. Studying bias in adaptive learning systems is important because they make real-time decisions that impact students' learning and experiences closely. In addition, these systems often serve a diverse student population. Hence, the system behavior needs to be non-discriminatory across all the student populations it serves. Below, I summarize the three studies conducted as part of this dissertation work and highlight the upstream biases discovered in them. I also discuss the implications of these findings to the research, design, and development of adaptive and artificially intelligent learning systems.

Study #1. Non-conformance of data to a widely-accepted theoretical model. In this study, I empirically tested the generalizability of a highly-cited theoretical model of affect dynamics (D'Mello & Graessar, 2012) that is often used in affect research and design of interventions in adaptive systems. First, a systematic literature review revealed that studies that do show some evidence for the theoretical model were all conducted in the United States with undergraduate populations. To better understand its scope of applicability, I analyzed ten past affect datasets collected in diverse contexts and found that the theoretical model has a more limited scope than what it is currently being used for. The results suggest that affective patterns seem to differ based on the country in which the research was conducted (the United States versus the Philippines), highlighting the need to focus on cultural differences in using this theory.

Theory as an upstream source of bias. Studying theory as an upstream source of bias becomes helpful in understanding differing effectiveness of theory-driven downstream decisions for different student subpopulations (e.g., students in the United States versus the Philippines). As illustrated in Section 1.3, the theoretical model investigated in this study has influenced several aspects of adaptive decisions, including the conceptualization of affective processes (e.g., Andres et al., 2019), the interpretation of student behaviors (e.g., Kort et al., 2001), the construction of variables in modeling (e.g., Hershkovitz et al., 2013), and the design of interventions in adaptive systems (e.g., Padron-Rivera et al., 2018). The findings of this study suggest that it is highly unlikely that there is a universal multi-step

pattern in affect dynamics in learning, while there may still be some contextually relevant patterns useful to understand student experience in specific student populations.

Impact of convenience sampling. One possible explanation for the theory bias revealed in this study is the convenience sampling in the studies that originally informed the theoretical model and the subsequent studies that showed some conformance. Most of these are small-scale experiments conducted in the United States with undergraduate students. Moreover, there is often little to no research on the generalizability of assumptions when the systems designed based on such theories are used in non-western contexts despite evidence on differences in student use (Ogan et al., 2012; Rodrigo et al., 2013). It is not an uncommon practice to sample from WEIRD (Western Educated Industrialized Rich and Democratic) contexts (Section 1.3). As reported by Blanchard (2015), there is "limited cultural diversity" (p. 225) in the student population investigated in the research involving adaptive learning systems despite the differing needs of students from non-WEIRD contexts. Careful attention must be paid both in recruiting participants from target populations and in testing generalizability assumptions of theories that may not be context-aware.

Study #2. Differing implications of technology design on student outcomes. Help-seeking and motivation have been widely studied since the early days of adaptive EdTech design for their important but complex role in learning (Aleven et al., 2016). However, past research has often taken a cognitive approach - for example, focusing on how or when students seek help but not as much on who chooses to seek help. In this study, I examined the conditions under which the relationships between students' behavior, motivation, and

outcomes vary across different demographic contexts. By doing so, I challenged the implicit assumptions of generalizability of design choices and provided an evidence-based commentary on future research practices in the community surrounding how we consider diversity in our field's investigations.

Design as an upstream source of bias. As discussed in Section 1.3, the design of adaptive learning systems influences the quality of student experience and learning directly and also shapes what data gets collected and how. Bias in design could directly impact student outcomes or indirectly lead to inequitable downstream decisions due to suboptimal data collection or modeling. Often design choices are informed by theory and/or experiments which may suffer from representation and measurement bias themselves. Due to the diverse student populations served by adaptive learning systems, it is all the more important that the design choices are vetted for the target populations as demographic factors are often related to differences in educational outcomes (Childs, 2017). This study contributes to this goal not only by empirically demonstrating demographic disparities in the effectiveness of help-seeking and motivation design, but also by focusing on those student populations that are currently under-studied in bias evaluation, such as students with poorer economic backgrounds, English as a second language, and students in special education.

Methodological implication for bias research. Practical constraints of research projects (e.g., budget, recruitment, accessibility, and time) make it difficult to collect data from diverse populations of students to test the generalization of design assumptions. Even when larger sample sizes are obtained (e.g., interaction log data from widely-used adaptive

learning systems), there are often limits in the student demographic data that can be obtained (e.g., due to concerns over student privacy). Thus, it seems important that greater efforts are made to overcome the challenges involved in investigating bias in the design of adaptive systems. This study contributes to this goal by demonstrating the use of publicly-available, school-level demographics where there is often access to larger and more diverse samples of student data, but individual student demographics may be difficult or impossible to acquire.

Study #3. Varying effectiveness of methodological improvements. New methods are emerging in the research of adaptive systems to make collecting data, building models, and visualizing results more efficient. In this study, I examined the demographic disparity in the effectiveness of one such methodological improvement in annotated data collection. Active learning - a subfield of machine learning - has been proposed to improve annotated data collection for complex constructs like affect (Yang et al., 2018). However, in practice, this method suffers from the cold-start problem where it does not have access to sufficient data yet to learn from. To overcome this limitation, I devised an approach that uses past affect data. More importantly, I conducted experiments to show that mismatches in the urbanicity (urban vs. suburban) of the past data and target student population could be detrimental to effective modeling and provided recommendations to mitigate this disparity.

Data collection method as an upstream source of bias. As discussed in Section 1.3, data collection methods closely shape the measurement of the educational constructs like affect. Accordingly, measurement bias can occur due to issues in data collection methods.

Measuring complex and sometimes subjective educational constructs needs caution to ensure that the measurements are reliable for different student subpopulations. This study suggests that applying a method across multiple populations brings challenges not seen in applying it within a single population, highlighting the need to conduct studies on demographic disparities in the benefits of methodological innovations.

Overall limitations. Although I present limitations of the individual studies separately in Chapters 2, 3, and 4, I believe it is important to explicitly discuss some of the limitations to the approach used in this dissertation to categorize students into specific subgroups. These limitations are also likely true for many other current efforts in studying subgroup bias and fairness in technology shaping education. First, this dissertation followed the standard race categorizations, which are politically influenced and likely oversimplified (Strmic-Pawl et al., 2018). Second, there was not enough data on indigenous students to include in the study (a challenge that has also faced other studies concerned with these students' success, such as Anderson et al., 2019). And because this was secondary data analysis, there was limited opportunity to collect more data. Third, there are other ways to categorize students for bias research that were not available in the current data but appear to be associated with algorithmic bias (e.g., first-generation students; Yu et al., 2020, military-connected status; Baker et al., 2020). Finally, bias may also emerge at the intersection of student demographics rather than a single dimension. This is yet to be studied with respect to biases in adaptive learning systems (both upstream and downstream).

5.3 Future directions

In my future research, I plan to develop theories and methods for understanding, identifying, and mitigating bias in adaptive learning systems. I plan to conduct empirical work focused on contextualizing the experiences of diverse student subgroups to serve the needs of all students, especially those who are likely to be underserved by biased systems. Towards this goal, I propose the following directions for future research.

Origins of upstream bias. In this line of inquiry, I want to focus on locating the origins of the upstream biases. So far, my dissertation identified upstream sources of biases in adaptive learning systems and discussed potential origins of it (Section 1.3). For example, representation bias can occur for reasons such as 1) most research studies tend to be conducted in western countries with adaptive systems developed by designers in the west, 2) small-scale experiments tend to recruit from a convenience sample due to the practical constraints of research projects, and 3) even when there is access to larger, more diverse datasets, it is often harder to collect student demographic data due to concerns over student privacy. Such representation bias can invalidate assumptions for student subpopulations not represented in the experiments informing these upstream components. Similarly, measurement bias can occur due to issues in data collection methods, such as the lack of reliability of measurements across different student subpopulations. And lastly, historical bias can occur due to forming theories and design choices by observing the world as it exists, including its biases. In my future work, I plan to formalize the understanding of the origins of upstream bias.

Theory Building on sources, origins, and harms of bias. In this line of inquiry, I want to further extend my dissertation work to develop a theoretical framework for bias in affective technologies shaping adaptive learning systems. Despite the recent spike in the research on bias in adaptive and artificially intelligent systems, there is no consensus on a single definition of bias. Furthermore, the often overlapping conceptualizations of bias and the harms they cause have themselves been considered a potential limitation that needs to be addressed as the work to mitigate it emerges (Blodgett et al., 2020). Lack of clarity in the description of bias has been argued to hinder the progress in conversations about "what kinds of system behaviors are harmful, in what ways, to whom, and why" (Blodgett et al., 2020, p. 2). Hence, there is a new urgency to provide definitions and theoretical grounding for the known biases in technology enabling adaptive systems.

To respond to this challenge specifically for the applications of affective computing in adaptive learning systems, I plan to extend my dissertation work and develop a theoretical framework of bias representing its sources (e.g., data, model, theory), origins (e.g., measurement, representation, historical), impacted populations (e.g., female students, rural school students, indigenous students), and harms (e.g., missed learning opportunity, reduced interest in subject learned). By narrowing the scope of this research away from generalities of data and algorithms, I hope to focus on developing sharper distinctions on the categories of bias for affective technologies in education while also grounding the categories in authentic educational contexts and theory. Although developed for affect, this work could serve as a basis for extending the taxonomy to other domains. By auditing real-

world affect datasets and adaptive learning systems, I plan to empirically test and further refine the framework. Through this work, I plan to provide a nuanced understanding of known biases to find ways to mitigate them and identify gaps for further investigation. I hope that it can serve as a shared language for researchers, designers, practitioners, and all stakeholders that can be used to audit for known biases in educational affect technologies.

Technical and human-centered approaches to addressing bias. In this line of inquiry, I want to develop context-aware methods that consider student diversity by design. There are new approaches being developed for fair machine learning. The general idea is to add fairness as a constraint while optimizing the model such that the model with the best fairness tradeoff is selected. More often than not, the issue of bias or unfairness in this discussion is conceptualized predominantly as a technical problem with little consideration to real-world contexts. The problem, however, is that there is no consensus on what "fair" means - Fair to whom? To answer, we need to take an explicit position on where we, as a society, stand on the social construct of fairness. We also need to define the groups of concern explicitly, which is also not straightforward. For example, is it a specific race or gender or sexual orientation, or a specific combination? As discussed earlier, one of my findings so far highlights how the group differences that matter most might not be the groups that are the most immediately obvious. There are also limitations to the completeness of data that can be collected from some student subgroups (e.g., categories that may require students to share sensitive information like LGBTQ identities).

By grounding my investigations in an authentic educational context, I want to test the limits of the advancements in fair machine learning by juxtaposing it with the constraints of the real world and highlighting the complexities involved in posing educational problems algorithmically. I also want to explore ways to bring the voices of teachers and learners in designing equitable human-AI adaptivity in learning systems. For this, I plan to conduct human-centered studies with students and teachers who use adaptive learning systems in their classrooms to develop approaches to elicit fairness understanding from them and potentially conduct collective audits to identify and detect bias in adaptive learning systems. I believe that it is both an imperative and an opportunity to engage and empower stakeholders to have an active voice in bias research.

Conclusion

Motivated by the inextricable link between learning and context, I investigated how ignoring learner context in the design and development of adaptive learning systems could introduce harmful biases in them. Specifically, I have tried to bring the attention of bias research to those aspects of the design and development of adaptive systems that are often obscured in the evaluation of bias in an algorithmic model. My research so far has identified some of the biases in upstream sources such as theories, design, and data collection methods for demographic categories such as urbanicity, race/ethnicity, economic status, English as a second language, special education, and charter school. Achieving equity in educational technology has a long way to go, but by investigating upstream bias we can reduce many disparities that are not being addressed by current approaches.

BIBLIOGRAPHY

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(1), 205- 223.

Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the Fairness of Graduation Predictions. *Proceedings of the 12th International Conference on Educational Data Mining*, 488–491.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2), 167-207.

Andres, J. M. A. L., Ocumpaugh, J., Baker, R. S., Slater, S., Paquette, L., Jiang, Y., ... & Biswas, G. (2019, March). Affect sequences and learning in Betty's Brain. *In Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 383-390).

Andres, J. M. L., Rodrigo, M. M. T., Sugay, J. O., Banawan, M. P., Paredes, Y. V. M., Cruz, J. S. D., & Palaoag, T. D. (2015). More Fun in the Philippines? Factors Affecting Transfer of Western Field Methods to One Developing World Context. *In AIED Workshops.*

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Arroyo, I., Beck, J., Woolf, B. P., Beal, C. R., & Schultz, K. (2000). Macro adapting animal watch to gender and cognitive differences with respect to hint interactivity and symbolism. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), Proceedings of the *5th International Conference on Intelligent Tutoring Systems*, ITS 2000 (pp. 574–583). Berlin: Springer Verlag.

Arroyo, I., Burleson, W., Tai, M., Muldner, K., & Woolf, B. P. (2013). Gender Differences in the Use and Benefit of Advanced Learning Technologies for Mathematics. *Journal of Educational Psychology*, 105(4), 957–969.

Baker, R., Ma, W., Zhao, Y., Wang, S., Ma, Z. (2020) The Results of Implementing Zone of Proximal Development on Learning Outcomes. *Proceedings of the 13th International Conference on Educational Data Mining*, 749-753.

Baker, R. S., Walker, E., Ogan, A., & Madaio, M. (2019). Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context*, 1(1), 1–13.

Baker, R. S., & Hawn, A. (2021). *Algorithmic Bias in Education*.

Baker, R. S., Berning, A., & Gowda, S. M. (2020). Differentiating Military-Connected and Non-Military-Connected Students: Predictors of Graduation and SAT Score. *EdArXiv*.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Binns, R. (2020, January). On the apparent conflict between individual and group fairness. *In Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 514-524).

Blanchard, E. G. (2012). On the WEIRD nature of ITS/AIED conferences. *In International Conference on Intelligent Tutoring Systems* (pp. 280-285). Springer, Berlin, Heidelberg.

Blanchard, E. G. (2015). Socio-cultural imbalances in AIED research: Investigations, implications and opportunities. *International Journal of Artificial Intelligence in Education*, 25(2), 204-228.

Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.

Borman, G. D., Grigg, J., & Hanselman, P. (2016). An effort to close achievement gaps at scale through self-affirmation. *Educational Evaluation and Policy Analysis*, 38(1), 21-42.

Campbell, N. J. (1989). Computer anxiety of rural middle and secondary school students. *Journal of Educational Computing Research*, 5, 2, 213-220.

Childs, D. S. (2017). Effects of Math Identity and Learning Opportunities on Racial Differences in Math Engagement, Advanced Course-Taking, and STEM Aspiration. *PhD Dissertation*. Temple University.

Clark, M., Rothstein, J., & Schanzenbach, D. W. (2009). Selection bias in college admissions test scores. *Economics of Education Review*, 28(3), 295-307.

Cole, P. (1986). Children's spontaneous control of facial expression. *Child development*, 1309-1321.

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. *In Handbook of human-computer interaction* (pp. 849-874). North-Holland.

Crawford, K. [The Artificial Intelligence Channel]. (2017, December 11). *The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford* [Video]. YouTube.

Csikszentmihalyi, M. (1990) Flow: *The Psychology of Optimal Experience*. New York, NY, USA: Harper & Row.

D'Mello, S. Graesser, A. (2012). Dynamics of Affective States during Complex Learning. *Learning and Instruction*, 22, 145-157.

Desmarais, M. C., & d Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1), 9-38.

Dorans, N. J. (2010). Misrepresentations in Unfair Treatment by Santelices and Wilson. *Harvard Educational Review*, 80(3), 404–413.

Doroudi, S., & Brunskill, E. (2019). Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 335–339.

Dunn, J., & Brown, J. (1994). Affect expression in the family, children's understanding of emotions, and their interactions with others. *Merrill-Palmer Quarterly* (1982-), 120-137.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. *In Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2), 124.

Elfenbein, H. A., Mandal, M. K., Ambady, N., Harizuka, S., & Kumar, S. (2002). Cross-cultural patterns in emotion recognition: highlighting design and analytical techniques. *Emotion*, 2(1), 75.

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013). The effects of culturally congruent educational technologies on student achievement. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 493–502). Springer Berlin Heidelberg.

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3), 330–347.

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, 33(4), 111–117.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234.

Graham, S. E. & Provost, L. E. (2012). Mathematics achievement gaps between suburban students and their rural and urban peers increase over time. *Issue Brief No. 52*. Carsey Institute.

Gross, J. J., Carstensen, L. L., Pasupathi, M., Tsai, J., Götestam Skorpen, C., & Hsu, A. Y. (1997). Emotion and aging: Experience, expression, and control. *Psychology and Aging*, 12(4), 590.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29-29.

Hershkovitz, A., de Baker, R. S. J., Gobert, J., Wixon, M., & Pedro, M. S. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*, 57(10), 1480-1499.

Holstein, K., Doroudi, S., Fikes, T., Jones, K., McCoy, C., Meaney, M., & Lang, D. (2019). Fairness and equity in learning analytics systems (FairLAK). *In Companion Proceedings of the Ninth International Learning Analytics & Knowledge Conference* (LAK 2019) (pp. 1-2).

Holstein, K., & Doroudi, S. (2021). Equity and Artificial Intelligence in Education: Will" AIEd" Amplify or Alleviate Inequities in Education?. *ArXiv E-Prints*.

Hu, S. (2003). Educational aspirations and postsecondary access and choice: Students in urban, suburban, and rural schools compared. *Education Policy Analysis Archives*, 11, 14, 14.

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. *Proceedings of The 13th International Conference on Educational Data Mining* (EDM 2020), 431–437.

Kai, S., Andres, J. M. L. ., Paquette, L., Baker, R. S. ., Molnar, K., Watkins, H., & Moore, M. (2017). Predicting Student Retention from Behavior in an Online Orientation Course. *Proceedings of the 10th International Conference on Educational Data Mining*, 250–255.

Karumbaiah, S., Andres, J. M. A. L., Botelho, A. F., Baker, R. S., & Ocumpaugh, J. S. (2018). The Implications of a Subtle Difference in the Calculation of Affect Dynamics. *In 26th International Conference for Computers in Education*.

Karumbaiah, S., Baker, R. S., Tao, Y., & Liu, Z. (2022). How does Students' Affect in Virtual Learning Relate to Their Outcomes? A Systematic Review. *International Learning Analytics & Knowledge Conference.*

Kearns, M. (2021). *Foundations of Machine Learning*. CIS 399. University of Pennsylvania.

Khachatryan, G. A., Romashov, A. V., Khachatryan, A. R., Gaudino, S. J., Khachatryan, J. M., Guarian, K. R., & Yufa, N. V. (2014). Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education*, 24(3), 333-382.

Kimble, G. A. (1987). The scientific value of undergraduate research participation. *American Psychologist*, 42(3), 267- 268

Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, emotion, and well-being: Good feelings in Japan and the United States. *Cognition & Emotion*, 14(1), 93-124.

Kizilcec, R. F., & Lee, H. (2020). Algorithmic Fairness in Education. ArXiv E-Prints.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30-43.

Koedinger, K. R., & Corbett, A. (2006). *Cognitive tutors: Technology bringing learning sciences to the classroom*.

Kort, B., Reilly, R., & Picard, R. W. (2001, August). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In Proceedings *IEEE international conference on advanced learning technologies* (pp. 43-46). IEEE.

Kuppens, P. (2015). It's about time: A special section on affect dynamics. *Emotion Rev*., 7(4), 297-300.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in US schools. *Educational researcher,* 35(7), 3-12.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 182.

Lee, H., & Kizilcec, R. F. (2020). Evaluation of Fairness Trade-offs in Predicting Student Success. *ArXiv E-Prints.*

Lohr, S. (2018). Facial recognition is accurate, if you're a white guy. *New York Times*, 9, 8.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education.*

Madnani, N., Loukina, A., Von Davier, A., Burstein, J., & Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. *In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 41-52).

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016, May). AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. *In Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems* (pp. 3723-3726).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv E-Prints.*

Melis, E., Goguadze, G., Libbrecht, P., & Ullrich, C. (2009). Culturally Adapted Mathematics Education with ActiveMath. *AI & Society*, 24(3), 251–265.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

O'Reilly-Shah, V. N., Gentry, K. R., Walters, A. M., Zivot, J., Anderson, C. T., & Tighe, P. J. (2020). Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. *British Journal of Anaesthesia*, 125(6), 843–846.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487–501.

Ocumpaugh, J., Baker, R. S., Karumbaiah, S., Crossley, S. A., & Labrum, M. (2020, July). Affective Sequences and Student Actions Within Reasoning Mind. *In International Conference on Artificial Intelligence in Education* (pp. 437-447). Springer, Cham.

Ogan, A., Walker, E., Baker, R. S., Rebolledo Mendez, G., Jimenez Castro, M., Laurentino, T., & De Carvalho, A. (2012, May). Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1381-1390).

Ogan, A., Yarzebinski, E., Fernández, P., & Casas, I. (2015, June). Cognitive tutor use in Chile: Understanding classroom and lab culture. *In International Conference on Artificial Intelligence in Education* (pp. 318-327). Springer, Cham.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards Understanding How to Assess Help-Seeking Behavior Across Cultures. *International Journal of Artificial Intelligence in Education*, 25(2), 229–248.

Okur, E., Aslan, S., Alyuz, N., Esme, A.A., Baker, R.S. (2018) Role of Socio-Cultural Differences in Labeling Students' Affective States. *Proceedings of the 19th International Conference on Artificial Intelligence in Education*, 367-380.

Padron-Rivera, G., Joaquin-Salas, C., Patoni-Nieves, J. L., & Bravo-Perez, J. C. (2018, June). Patterns in poor learning engagement in students while they are solving mathematics exercises in an affective tutoring system related to frustration. *In Mexican Conference on Pattern Recognition* (pp. 169-177). Springer, Cham.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2), 127-144.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, 12(3), 1–30.

Pardos, Z. A., & Heffernan, N. T. (2010, June). Modeling individualization in a bayesian networks implementation of knowledge tracing. *In International Conference on User Modeling, Adaptation, and Personalization* (pp. 255-266). Springer, Berlin, Heidelberg.

Pardos, Z. A. (2017). Big data in education and the models that love them. *Current opinion in behavioral sciences*, 18, 107-113.

Rebolledo-Mendez, G., Huerta-Pacheco, N. S., Baker, R. S., & du Boulay, B. (2021). Meta-affective behaviour within an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education*, 1-22.

Reich, J., & Ito, M. (2017). From good intentions to real outcomes: Equity by design in learning technologies. *Digital Media and Learning Research Hub*.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, 14(2), 249-255.

Richardson, J. T. (2004). Methodological issues in questionnaire-based research on student learning in higher education. *Educational psychology review*, 16(4), 347-358.

Rodrigo, M., Baker, R., & Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: comparison to prior research in the USA. *Teachers College Record*, 115(10), 1-27.

Rosser, P. (1987). *Sex bias in college admissions tests: Why women lose out.*

Roth, L. (2009). Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. *Canadian Journal of Communication,* 34(1).

Ryan, A. M., Shim, S. S., Lampkins-uThando, S. A., Kiefer, S. M., & Thompson, G. N. (2009). Do gender differences in help avoidance vary by ethnicity? An examination of

African American and European American students during early adolescence. *Developmental Psychology*, 45(4), 1152–1163.

Sap M, Card D, Gabriel S, Choi Y, Smith NA. (2019). The risk of racial bias in hate speech detection. *In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp 1668–1678

Santelices, M. V., & Wilson, M. (2010). Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning. *Harvard Educational Review*, 80(1), 106–134.

Self, J. (1999). The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. *International journal of artificial intelligence in education*, 10(3-4), 350-364.

Shute, V. J., & Psotka, J. (1994). *Intelligent Tutoring Systems: Past, Present, and Future*. Armstrong Lab Brooks AFB TX Human Resources Directorate.

Soundarajan, S., & Clausen, D. L. (2018). Equal Protection Under the Algorithm: A Legal-Inspired Framework for Identifying Discrimination in Machine Learning. *Proceedings of the 35th International Conference on Machine Learning*.

Strmic-Pawl, H. V, Jackson, B. A., & Garner, S. (2018). Race Counts: Racial and Ethnic Data on the U.S. Census and the Implications for Tracking Inequality. *Sociology of Race and Ethnicity*, 4(1), 1–13.

Suresh, H., & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. *ArXiv E-Prints.*

Tsai, J., Levenson, R. (1997) Cultural influences on emotional responding: Chinese Am. & European Am. dating couples during interpersonal conflict. *J. Cross-Cultural Psych*., 28(5), 600-25.

Uchida, Y., Townsend, S., Rose Markus, H., Bergsieker, H. (2009). Emotions as within or between people? Cultural variation in lay theories of emotion expression and inference. *Personality and Social Psychology Bulletin*, 35(11), 1427-1439.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197-221.

Wang, Y., & Beck, J. (2013, July). Class vs. student in a bayesian network student model. *In International Conference on Artificial Intelligence in Education* (pp. 151-160). Springer, Berlin, Heidelberg.

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, 2(2), 5-13.

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving Retention: Predicting at- Risk Students by Analysing Clicking Behaviour in a Virtual Learning Environment. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 145–149.

Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D. G., Dolan, R., & Christopherson, R. M. (2010). The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (ITS'10) (pp. 327–337). Springer Berlin Heidelberg.

Yang, T. Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active Learning for Student Affect Detection. *International Educational Data Mining Society*.

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *Proceedings of The 13th International Conference on Educational Data Mining* (EDM 2020), 292–301.

Yudelson, M. V., Fancsali, S. E., Ritter, S., Berman, S. R., Nixon, T., & Joshi, A. (2014). Better Data Beat Big Data. *Proceedings of the 7th International Conference on Educational Data Mining*, 205–208.