Statistical Models for Inclusive Measurements of Student Engagement

Alison Rosenman

December 20, 2019
Version: Final
Advisor: John Dougherty
Abstract

In the technology industry and in academia, technologists frequently use the term "engagement" as one metric to measure a user's experience. Within the sub-field of Human-Computer Interaction (HCI), recently-published papers do not all agree on their definitions of this term. However, recent HCI research has realized some success in measuring and modelling student engagement. In this thesis, I survey the variety of approaches taken in such work. I also discuss how these studies surprisingly do not take into account students who use assistive technologies despite these technologies' notorious tendencies towards user disengagement. Ultimately, I propose future research towards modelling and quantifying engagement of all students in a way that takes into account the differences in the experiences of learning with and without assistive devices.
Acknowledgements

Many thanks to my advisor, Professor John Dougherty, for his support of this project, to Professor Sorelle Friedler for her help as the seminar advisor, and to Professor Rajesh Kumar for his feedback. Classmates Silvia Susai and Lamiaa Dakir each provided feedback at points throughout the process, so I thank them as well. I’d also like to thank the rest of Professor Dougherty’s research group—Kendall Robison, Sam Lowenstein, Mac Lyle, and Zachary Toner—for their continual support throughout the semester. Finally, I would also like to thank the good friends I had around me while I wrote this thesis.
## Contents

1 Introduction ..................................................... 2  
   1.1 Previous Work In This Area ................................. 2  
      1.1.1 Overview of Previous Work .......................... 2  
      1.1.2 Detail of Previous Work .............................. 4  
   1.2 Motivation .................................................. 4  
2 Literature Review ............................................. 6  
   2.1 Detail of Previous Work ................................. 6  
      2.1.1 Using Facial Expressions and Bayesian Networks to Model  
            Student Engagement Levels ............................ 8  
      2.1.2 Experience Sampling to Predict User Habits .......... 13  
      2.1.3 Robo-Teachers to Improve Engagement Using Brainwave Analysis 15  
   2.2 Conclusion ................................................. 19  
3 Statement of The Problem .................................... 20  
   3.1 Problem Introduction .................................... 20  
   3.2 Problem Approach ........................................ 21  
4 Conclusion .................................................... 23  

Bibliography ..................................................... 24
Introduction

The terms "user experience," "experience design," and "engagement" are common words in everyday talk about technology as such tenets of Human-Computer Interaction (HCI) increasingly influence the way that big technology companies design consumer products. But just as the technology industry throws around several loose definitions for these terms, HCI and other fields within Computer Science have ambiguous meanings and measures for them. Still, substantial amounts of work have been published on ways to measure "experience" and "engagement" in educational and occupational settings. In particular, researchers have developed statistical models to measure engagement levels of students completing classroom tasks from wildly different experiments on engagement and attention. The purpose of this paper is to summarize the multitude of studies which attempt to model educational engagement, compare these models, and ultimately propose an investigation to refine strategies for quantifying engagement so that they can be universally-applicable. Aside for improving statistical models for universal design, this investigation may also yield some worthwhile insight into how to define such important terms in the first place.

1.1 Previous Work In This Area

Before discussion of the papers I study in-depth in the literature review section, I survey the papers which first inspired me to pursue this topic. The breadth of interdisciplinary fields represented by the papers I survey in the following overview highlights the importance of the research I ultimately propose in Section 3 of this thesis.

1.1.1 Overview of Previous Work

In a 2018 study of engagement and HCI, Doherty and Doherty sift through the millions of publications in the Association for Computing Machinery Guide to Computing Literature in an attempt to focus in on how HCI researchers define the term that is so central to their work: engagement (K. Doherty and G. Doherty 2018). The study narrows first to the 1150 papers which have "engagement" in the title or as a keyword and then to the 351 results that specifically focused on digital technologies.
and which are not dissertations or authored books. Ultimately, Doherty and Doherty find that the majority of Computer Science papers which use the word "engagement" do not define it. What's worse, the papers which do define "engagement" altogether present hundreds of unique descriptions and definitions.

Despite the lack of consensus as to how to define "engagement" in computer science contexts, a variety of recent publications propose statistical models of classroom engagement created from observations as subject-students interact with a variety of systems. In particular, these studies often rely on Bayesian networks and other statistical models to come up with automatic detections of student engagement with electronic systems (Bosch et al. 2015)(Ting et al. 2013)(Millán et al. 2010)(Szafir and Mutlu 2012)(Kapoor and Horvitz 2008)(Whitehill et al. 2014). Thus, a problem of defining a key HCI term becomes a surprisingly computational one.

Though several such recent studies are successful at measuring and quantifying student engagement, their experiments and set-ups are each incredibly unique. Some studies involve automated recognition of facial expressions to gather real-time evaluations of student engagement in a way similar to how a teacher observing their students would (Bosch et al. 2015)(Whitehill et al. 2014). Another approach is to use data from brain signals to determine engagement on a subconscious level (Szafir and Mutlu 2012). And, still more research studies logs of users’ interactions with computer interfaces that are part of their normal classrooms and workplaces (Kapoor and Horvitz 2008)(Ting et al. 2013).

At the same time, research regarding assistive technology shows that such technologies are often rife with issues of disengagement due to social pressures and anxieties stemming from using uncommon and often clunky devices (Shinohara and Wobbrock 2011). Research in this area is often surprisingly qualitative rather than quantitative as compared to the above studies of engagement which do not indicate any inclusion of participants who require assistive technologies. This is perhaps an artifact of the often low-tech environment of data collection for disabled students (Marcu et al. 2013). Though there are risks of directly applying the engagement detection methods and their underlying statistical models to populations who use assistive technologies when these models were not trained on data from people included in this population, work from Shinohara and Wobbrock and others highlights the huge problem in current methods that people who rely on assistive technologies are often ignored (Guo et al. 2019)(Shinohara and Wobbrock 2011).

Though there has been much work published recently about the topics of of modelling engagement with computational methods and of engagement and assistive technology, I have found little work that attempts to blend the two problems together. At the end of this paper, I propose further research towards solving this problem.
1.1.2 Detail of Previous Work

Before assessing current engagement modelling efforts for universal applicability, I will survey past efforts within Computer Science to model student engagement. In particular, I will focus in on three novel methods for detecting and modelling engagement which take wildly different approaches. In two of these cases, experiments directly involve students. While the third is set in a workplace, it employs the least invasive technology, which I see possibly transferring over to classrooms to measure student engagement specifically in the future.

First, Bosch et al. 2015 presents a method for analyzing student affects, including engagement and "off-task-edness" using computer vision and machine learning methods to record students' facial expressions as they work on a computer-based physics class assignment. This paper finds a Bayesian network to be the most effective model of engagement after testing several other classifiers. Other methods tested in place of a Bayesian network include support vector machines, decision trees, and logistic regressions (Bosch et al. 2015). Second, I review Kapoor and Horvitz 2008, which uses experience sampling models (ESMs) to determine the engagement of Microsoft employees as they go about their normal work on their computers. Rather than analyzing facial expressions, this study instead issues probes to participants' computers which evaluate the participant's busyness levels and aversion to distraction. Finally, I review Szafir and Mutlu 2012, which tests a robotic teacher that is programmed to adapt to dips in student engagement with human-like cues. This work blends elements of the continual affect monitoring approach of Bosch et al. 2015 and the statistical models of Kapoor and Horvitz 2008 to successfully predict dips in engagement and update the robotic teacher's presentation of material accordingly and effectively, all in real time.

The uniqueness of each of these studies and others in this field demonstrate just how ripe for research the study of student engagement is. And, as such recent works which rely on fields such as psychology and education in addition to computer science, these papers have huge potential impacts on the world.

1.2 Motivation

The motivation for this work is twofold. First, it is important to in general understand the metric of "engagement" that is used to evaluate technology in a variety of Computer Science fields and how it might be measured. As previous work shows that engagement is seldom quantified, it is crucial to study the rather new methods which have shown promise in modelling engagement to highlight how such methods
should become more widely-used in academic papers and in the technology industry. And, because there are so many approaches for measuring engagement which are each so different, it is important to compare and contrast these methods to work towards a future standard of engagement modelling.

Second, and, most important, it is critical to understand ways in which engagement can be measured in educational settings so that educational technology can best benefit every student. As the past work I review will show, there are indeed ways to measure the engagement levels of students as they use technology to learn. But, this past work fails to measure engagement with assistive learning technologies, an area of technology often rife with issues of disengagement which stem from social anxieties of using such different technologies. Thus, a key motivation for this literature review is to bridge the often-qualitative studies that find issues of disengagement in assistive learning technologies with the highly-quantitative studies that model engagement with typical learning technologies. Ultimately, in this review, I will use my survey of prior work to propose future research to model engagement with assistive technologies.
Previous studies about engagement and educational technology provide glimpses at the ways in which student engagement can be measured and modeled. From brainwaves to facial expressions, past work has used an incredible variety of data sources to try to figure out when students are paying attention and when they are off-task. These studies ultimately present statistical models that each author finds successful at modelling engagement given their chosen data source.

At the same time, recent research about engagement with assistive technology is similarly able to come to conclusions about the ways in which engagement levels with these technologies can fall. But, unlike the body of work concerned solely with modelling engagement at all, research about engagement with assistive technologies often does not apply statistical models in the same way and are not so successful at quantifying engagement. Instead, reports on engagement where assistive technology is concerned rely heavily on qualitative data sources such as personal interviews and feedback on assistive products.

Why is it that the studies focused on assistive technology engagement are so much less quantitative than those which involve standard technologies? Or, on the other side, why do all these statistical studies not consider qualitative factors in their characterizations of engagement as well? In addition to providing a summary of recent publications regarding modelling engagement in education, this literature review is motivated by a desire to bridge the divide between engagement research for typical students and engagement research for students who require accessible tools.

2.1 Detail of Previous Work

Recent papers published in HCI present a variety of novel approaches towards quantifying student engagement with technology-based learning tools and systems. Though the experiments in these studies often rely on different technologies as means of collecting data about user engagement, the studies all ultimately present statistical models for engagement.
Underlying all work towards quantifying student engagement are several statistical models. Bayesian networks are a particularly common tool to model student engagement because uncertainty can be represented by probabilities implicit in these models, and efforts have been made to share such models with educators and within the field of education outside of pure computer science (Millán et al. 2010). Ting et al. find that the engagement levels of students using a physics simulation called INQPRO can be modelled from INQPRO’s usage logs alone (Ting et al. 2013). Though literature I review later in more detail involves more robust and complex data sources for use with Bayesian networks, it is quite powerful that Bayesian networks sufficiently model student engagement from classroom physics software alone as this study finds, though researchers note that their procedure required lots of data transformation steps between collection and analysis phases (Ting et al. 2013).

Past work concerned with modelling student engagement also often attempts to mimic the process a real schoolteacher would take to determine if students are paying attention or not. In particular, this often involves reading a student’s facial expression and determining engagement levels from them (Bosch et al. 2015)(Whitehill et al. 2014). Since this process is still being researched, there is not yet a standard way to map facial expressions to engagement levels. Whitehill et al. and Bosch et al. both experiment with multiple models, with Bosch et al. finding a Bayesian network the most effective model and Whitehill et al. finding a support vector machine (SVM) the most effective model (Bosch et al. 2015)(Whitehill et al. 2014). In addition, a related Haverford College thesis work uses the facial recognition system Affectiva, which also relies on an SVM model (Uriostegui 2016).

Finally, past work concerned with engagement and assistive technology takes a much different approach than the above-mentioned studies. From interviews of 20 people who rely on assistive technologies, social pressures and self-consciousnesses of using such technologies result in disengagement with these technologies in occupational and educational settings (Shinohara and Wobbrock 2011). Though this result is common among almost all of the study’s participants, it is important to recognize that this study is conducted entirely by qualitative interview, with no quantitative component put forth to measure the strains on engagement with assistive technologies, social or otherwise. Perhaps, this is because the complex and highly-individualized assistive technologies do not produce standardized usage metrics and data, following from general trends in special needs education environments in which data collection methods are often non-standardized, highly-individualized, and low-tech (Marcu et al. 2013).

In the following sections, I review three studies of engagement modelling which do not mention users of assistive technologies in their participant pools to first gain
insight into the current state of research in this area in general. Then, I use section 3 of this paper to propose further work which may ameliorate the divide between the quantitative-ness of non-accessible studies and qualitative-ness of Shinohara and Wobbrock 2011, all with Guo et al.’s concerns in mind.

### 2.1.1 Using Facial Expressions and Bayesian Networks to Model Student Engagement Levels

Bayesian networks have become a popular tool for researchers interested in modelling engagement. Older studies employ Bayesian networks to model subjects’ knowledge and focus at particular instances, but without modelling engagement over time (Ting et al. 2013)(Millán et al. 2010). However, Bosch et al. create a model of automatic, real-time student engagement using several statistical methods including Bayesian networks as well as logistic regressions. Ultimately, this study provides critical insight into the ability of statistics to model several affects involved with learning.

Simply put, a Bayesian network is a directed acyclic graph which represents the probabilities and dependencies of certain events. Given a series of events and their probabilities, and which events impact the likelihood of others, one can create a Bayesian network to model the various outcomes of the system. In the case of Bosch et al. 2015, a Bayesian network is used to predict a student’s affect given their facial expressions as they complete a school assignment.

The premise of Bosch et al.’s research is that human teachers can infer the emotions, thoughts, concerns, and needs of students given their facial expressions and use these inferences to determine which students are engaged and which students are off-task. So, analyzing facial expressions of students could help us create a model to predict the most probable affect from a student using real-time facial recognition technology. The study defines the five “affects” involved in learning which can be detected by facial recognition technology as "engaged," "bored," "confused," "delighted," and "frustrated," all of these as opposed to simply "off task".

Detecting engagement levels via facial features is a compelling opportunity because the presence of facial feature variation does not change based on learning environment and can be detected with relatively inexpensive equipment, since most computers have built-in cameras already (Bosch et al. 2015). Bosch et al., in describing their research as occurring “in the wild,” note that their facial feature recognition techniques must be careful to consider the variability of lighting, camera quality, and external distractions in the classrooms they experiment in, as compared to similar laboratory experiments which do not have these potential variables. Attempting to
measure student engagement in a normal classroom context, largely without any additional or special tools, makes this study unique in its field.

After detailing previous literature on both the facial recognition and engagement detection aspects of their work, Bosch et al. describe their procedure. Bosch et al. study the facial expressions of 137 8th and 9th grade students in a public school in the United States during a 55-minute class period which took place in the school’s computer lab. Researchers recorded students using an online physics education tool called Physics Playground to ultimately use FACET, a commercial facial feature extraction software, to characterize students’ facial expressions throughout their class. The researchers built separate models in their FACET system for each of the five affects, as well as for a combination of all five together.

Figure 2.1 depicts the overall flow of data, from a student’s facial expression recording to a final prediction of their engagement level, that is detailed in Bosch et al.’s study. As a student sits at a computer in their classroom, an inexpensive camera attached to the computer records their facial expressions. These recordings are then fed through the FACET system which provides probabilities for the presence of nineteen different kinds of facial muscle movement detections as well as the pose of the student's head. From all of these muscle movements and positions that FACET can detect, FACET generates probabilities for 78 distinct facial features/facial movement patterns. These features are then run through a classifier built by the researchers to ultimately predict the student’s affect, one possibility being "engaged". The researchers experiment with fourteen different classifiers including support vector machines (SVMs), regressions, and decision trees, which can be imagined as interchanged into the "classifier" component of Figure 2.1. Ultimately, Bosch et al. find Bayesian networks to be the most successful for modelling the affect of engagement.

In Figure 2.2, I depict an example of what it might look like to use a Bayesian network as the classifier in this experiment. This example is certainly simplified and imagined from what Bosch et al. actually use in their experiment; the details of their Bayesian network are not discussed. However, in general, this Bayesian network will have 78 nodes—one for each of the facial features FACET detects—which may or may not be linked depending on if training data shows there is a joint probability relationship that is important to determining engagement between any two nodes.

In the small example I illustrate in Figure 2.2, I assume that having eyes open, eyebrows raised, and perhaps a smile is indicative of engagement. Given the granularity of the facial movements that FACET detects, I have focused even this small example for just the right side of the face, since the features FACET can detect are detailed at the muscular movement level, though you can imagine similar
nodes for other facial muscle movements, such as those on the left side of the face, potentially linking to the nodes I have in my example. In this example, probabilities from Node C, which indicates the probability of the subject’s right eyebrow being raised in the recording, are determined based off of the results of Nodes A and Nodes B, which each have their own probabilities associated with them. In this example, it is more likely than not that the subject has their eyes open (hence the 0.9 under the "yes" of Node A) but slightly more likely that the subject is not engaging right-side smile muscles (hence the 0.6 under the "no" of Node B). The probabilities associated with Node C depend on the probabilities for nodes A and B. For example, I am assuming that someone with their eyes closed and with no smile would have a hard time raising an eyebrow, which is why the third line of the probability table for Node C is so extreme towards "no". Though all of the numbers and features in this example are made-up for illustration purposes, one can imagine that the real model used by Bosch et al. looks something like this, with many more nodes and more precise probabilities.

Ultimately, Bosch et al. find that a Bayesian network is the most accurate way to model several affects which they detect from the facial recognition data from the FACET system. Though Bayesian networks are the most accurate for detecting the affects of "engaged," "confused," and "frustrated", as well as the five-way combination of them all, the authors find that "off-task" is best determined using a logistic regression and "bored" is best determined using clustering. The authors note that accuracy should not be the only performance metric for the models, since the data was quite skewed from one group of students to the next. For example, suppose that in one class students were overall not likely to express 'delight,' and in another 'frustrated' was a very common affect. Simply trying to calculate accuracy in these two cases might be misleading—never marking students as delighted in the first example, or always marking students as frustrated in the second, would lead to
rather high accuracy percentages without the model truly capturing what is going on.

Instead of only relying on an accuracy measure, the authors calculate an area under the ROC curve (AUC) measurement to test their various models' effectiveness. In the field of statistical learning, a ROC curve is a graph which displays the error of a classifier (James et al. 2013). A ROC curve plots the true positive rate versus the false positive rate for a classifier. Thus, a successful model would have a ROC curve which follows along the upper-left corner and top of its graph, since this means true positives exceed false positives, as well as a high AUC, as the area under such a curve would be rather large. Looking at all the data across all the student groups which participated in the study, Bosch et al. calculate the AUC for the Bayesian model of the "engaged" affect to be 0.655 (Bosch et al. 2015). Had the model taken in no learning information and instead guessed by chance, we'd expect AUC to be 0.5 as the ROC would be a straight line. This suggests that the model is relatively successful, since true positives outweigh false positives.

Interestingly, of the five affects the study measures as well as the "off-task" category, the results for "engaged" vary the least with respect to the duration of the facial

---

1"ROC" is a historic name which stands for "receiver operating characteristic curve"
expression recordings used with FACET. The AUC calculation for modelling engagement with a Bayesian network hardly varies when the duration of facial expression recording used with the FACET software increases anywhere from 3 seconds to 12 seconds. This finding suggests that the Bayesian network is particularly useful at modelling engagement given facial expressions, as it has about the same success no matter how long the recording of a student’s face is. In other words, the consistency of the Bayesian network is reassuring. These results can be observed in Figure 2.3 which comes directly from Bosch et al.

Though these results are promising as an indication that student engagement—as well as other aspects and affects of classroom experiences with technology—can be modelled statistically, they are not without caveats. The authors note that 25% of the instances of facial movement that they capture could not be analyzed by FACET because of factors such as poor lighting, students having unusual positions, or rapid movement. It is surprising that the software and the researchers are unable to handle such irregular behavior and conditions which are incredibly frequent in classroom settings. Though other studies attempt similar experiments to Bosch et al., they do not attempt to pull them off in real classroom and thus do not present these same challenges (Whitehill et al. 2014). So, despite Bosch et al.’s challenges, this work shows promise towards modelling engagement in actual classrooms. Future work will hopefully reduce the instances of unusable data.

In all, Bosch et al. present an approach to measuring engagement that is rather intuitive; just as a human teacher can interpret the affect of a student from visual cues, so too can a Bayesian network, informed by FACET facial recognition software.
2.1.2 Experience Sampling to Predict User Habits

Next, we turn to research from Microsoft aimed at modeling the same affective states that Bosch et al. model with facial expression recognition software, this time, using experience sampling methods (ESMs). These methods involve some sort of self-reflection or reporting on research variables from participants, in this case, how engaged participants are in their work or if they are able to be interrupted. Studies involving ESMs typically are found in the psychology field, but have become increasingly useful in Computer Science as relatively non-invasive ways to evaluate technologies’ usefulness to users (Consolvo and Walker 2003).

Fig 2.4 from Kapoor and Horvitz 2008 depicts the overall methodology the study takes at using automated ESMs to determine user affect, in particular, user experience and engagement. Simply put, the ESM automation system Kapoor and Horvitz use involves initializing a model of user affect, issuing “probes”—in the case of this study, computer prompts for subjects to mark whether they are busy with their normal work or if they have time to be interrupted—then refining the model based on how the users responded to the probe. All of this is accomplished with BusyBody, a software previously developed by one of the authors and others at Microsoft. Designed to measure the cost of interrupting a user, BusyBody software probes users to self-assess how interrupt-able they are at the moment while also evaluating the context of this self-assessment from its own logs of the user’s availability (Horvitz et al. 2004). These logs are informed by things like the user’s calendar availability and how much stuff is running on the user’s desktop (Horvitz et al. 2004).
The study tested this procedure with four different rules for when to issue the probes: random, in which probes appear at random times; uncertainty, in which probes occur at what the system calculates are the most uncertain situations; decision-theoretic (DT), in which a probability of a probe being positive is evaluated to determine whether a probe appears or not; and decision-theoretic dynamic (DT-dyna), in which the probability model used for the decision-theoretic probe is extended to address potential variances of the user’s availability that occur in different contexts by allowing the model to remove or store previously-determined states for future use as contexts change (Kapoor and Horvitz 2008).

Kapoor and Horvitz had their colleagues at Microsoft install and run BusyBody over the course of a two week period as these colleagues otherwise went about their normal work, which often times took place on their computers. Results show that the model issued more probes on average via the random and uncertainty policies, and fewer for the DT and DT-dyna policies. In particular, the DT-dyna policy issued much fewer probes on average than the other three policies. This suggests that the DT-dyna approach to modeling user engagement is superior because it requires the least supervision.

Another metric that this paper uses to assess their system was the "learning curves" of classification performance and how it changes over time. Focusing in on the results of two subjects in particular, the study finds that the DT-dyna probing policy issues fewer probes on average while also increasing its average recognition accuracy. These two events happening together show that the model applied with the DT-dyna probing policy is most effective at collecting real-time data on a user's affective state, including their engagement with their work.

Finally, Kapoor and Horvitz 2008 includes results from a post-experiment survey which yield interesting insights into the ways in which the subjects prefer to have their engagement monitored. When asked to rate how annoying the probes were on a scale from one to ten, the participants whose probes followed the DT-dyna policy reported the least annoyance with an average of 4.9 (Kapoor and Horvitz 2008). Those with the DT policy reported the second-least annoyance with an average of 6.2, followed by those with the uncertainty policy at 6.4 and the random policy at 7.4 (Kapoor and Horvitz 2008). Again, we see that the DT-dyna probing policy has superior performance to the others. Not only does the more-complex probability model behind when to issue a DT-dyna probe result in fewer probes issues and increased average recognition accuracy, but it is also most effective at. Thus, this method enables engagement monitoring to integrate as seamlessly as possible into participants’ normal lives.
This study’s results which are strongly in favor of the DT-dyna probing policy have significant implications on the future of engagement modelling. In contrast with many of the other papers I review, Kapoor and Horvitz 2008 involves virtually no hardware to model the study’s participants’ engagement with their work, other than the users’ own computers to run BusyBody. As a trade-off for low hardware required to run the study, the successful model powering the BusyBody probes is also the most complex. The success of the DT-dyna probing policy in this paper suggests that we can create models for engagement with technology which are unobtrusive, but that they become much more complex the less variety of data we have.

It is worth mention that of the 44 Microsoft employees recruited for this study, only 37 ultimately were able to properly provide data for Kapoor and Horvitz to analyze. Originally, participants were split into four groups of 11, with each group receiving their own version of BusyBody set to run with one of the four probing policies. However, very quickly, these groups began to fall apart. Three subjects were unable to install BusyBody correctly. Another participant had a computer failure and quit the study, and another three subjects were too busy to provide researchers with the required log files from BusyBody by the end of the trial. Given that all recruited participants were software developers, researchers, program managers, or group managers at Microsoft, it is easy to see how repeating this study with less-technically-skilled subjects, for example, young students, may not be possible. On the other hand, the significant lack of new hardware needed for this study is hugely promising for replicating or expanding this study into classrooms.

Unlike the work from Bosch et al. which I detail in section 2.1.1, this study is not conducted in an educational setting. However, it presents a successful way to model engagement which involves rather non-invasive data collection; participants were able to go about their normal workday without much thought about the experiment until the probes came along. For this reason, ideas from this study may translate well into classroom contexts in the future to evaluate students’ engagement levels with online tasks and in computer-based courses.

### 2.1.3 Robo-Teachers to Improve Engagement Using Brainwave Analysis

For yet another attempt at measuring and modelling engagement, Szafir and Mutlu take the intuitive approach of Bosch et al. and automatic statistical modelling approach of Kapoor and Horvitz one step further by combining them together. Like Bosch et al. 2015, this study evaluates engagement in a human-like way, aiming to address student disengagement in ways similar to how any real teacher would. And, like Kapoor and Horvitz 2008, this study also builds up a statistical model of student engagement with technology.
Fig. 2.5: The experimental setup of Szafir and Mutlu 2012. The "adaptive embodied agent" is a robo-teacher replacement, while the user is a real student whose brainwaves are being measured by EEG, shorthanded by the headset drawing.

[Szafir and Mutlu 2012]

engagement that is continually updating to predict the most accurate engagement level at a particular moment in time. But, most uniquely, this study aims at detecting engagement on a sub-conscious level, using non-invasive yet rigorous brainwave analyses, and ultimately employs a robotic replacement for teachers.

Szafir and Mutlu’s study involves building and testing robotic teachers—"adaptive embodied agents," as the paper calls them—which tell stories and monitor students’ brainwaves as a proxy for engagement (Szafir and Mutlu 2012). These agents then adapt to declines in engagement levels with features such as louder volumes, slower playback of recorded speech, and even head nods and changes in eye contact, just as a human teacher would. Ultimately, the study measures success of its systems by the degree of engagement increase detected after the robo-teachers have to make an adaptation. Though this all sounds far-fetched, the robo-teachers ultimately have huge promise for improving the effectiveness of instruction, as their optimization of response time when a student’s engagement dips has been shown to increase the effectiveness of the teaching (Szafir and Mutlu 2012). The general setup for this experiment is depicted in Fig. 2.5.

More specifically, the paper details an experiment in which participants listen to a story from an adaptive embodied agent while wearing a low-cost electroencephalography (EEG) monitor headset and are then asked questions about the story they just heard. 30 native English speakers found on the University of Wisconsin-Madison’s campus participated, and each were previously unfamiliar with the subject matter of the story the agent told. While the adaptive embodied agent is telling the story, the EEG measurements are monitoring the participant’s engagement levels from the continuous stream of EEG readings.
Unlike other brainwave-based engagement studies, Szafir and Mutlu do not rely on machine learning models to extrapolate engagement levels from the EEG readings taken from participants. Past studies with brain-computer interfaces (BCIs) often utilize machine learning methods to monitor participants’ active brain controls of electronic systems to eventually "learn" when participants are engaged. Instead, this study uses only passive brain signals. In particular, the study uses formula 1 below to compute engagement signal $E$ from the $\alpha$, $\beta$, and $\theta$ brainwave readings that an EEG gives. This formula has been shown in previous work to be a good measurement for task engagement from these brain signals (Szafir and Mutlu 2012).

$$E = \frac{\beta}{\alpha + \theta}$$

(1)

Because of this formula, Szafir and Mutlu do not need to train a machine learning model with large amounts of information on active brain measurements to infer engagement, as much previous work with BCIs attempts, and can instead focus on using statistics to model these real-time brain signals which are not consciously controllable (Szafir and Mutlu 2012).

Participants in the study received one of three treatments: no intervention, in which engagement levels as determined by EEG readings made no impact on how the story was read to them; randomly-timed intervention, in which the robotic agent made random adaptations meant to boost engagement; and, most importantly, the adaptive interventions, in which the robotic agent made cues to boost engagement based on calculated dips in engagement from the EEG signals. When the engagement levels determined from this formula 1 decline, the agent in the adaptive intervention mode is programmed to respond with cues such as louder volume and different body language (or, to be clear, robotic mimicry of body language!) to encourage the participant to become more engaged again. To determine when the engagement levels have dropped significantly enough for the agent to need to intervene, Szafir and Mutlu use two thresholds. The first compares the more recent engagement signals to all of the engagement signals seen so far using formula 2 below, where $x$ represents a 15-second time frame and $y$ represents the entire signal seen so far (Szafir and Mutlu 2012). Basically, this formula says to intervene if the most-recent 15-second window’s engagement signal is less than the entire observed signal and is declining. The second threshold uses a Least-Squares Regression to minimize a constantly-updating average of the engagement levels seen so far for the participant. So long as no cue has been issued in the last 15 seconds, the system issues a cue if either formula 2 indicates a cue is needed (outputs 1) or if the Least-Squares Regression
Regression finds a drop in the average engagement at the time (Szafir and Mutlu 2012). 

$$Cue(E) = \begin{cases} 
1 & : \frac{dE(x)}{dt} < \frac{dE(y)}{dt}, \frac{dE(x)}{dt} < 0 \\
0 & : otherwise 
\end{cases} \quad (2)$$

In all, results from the study confirm the hypothesis that participants who received targeted engagement cues that were triggered by drops in their engagement signals could recall the story they heard better than other participants (Szafir and Mutlu 2012). That is, the things like speaking louder and gesturing more at times of reduced engagement seemed to be effective for participants’ overall learning. Of the 14 post-story questions asked, participants who received the adaptive agents averaged over 12 questions correct, whereas the other two groups saw much lower averages. These results are depicted in Figure 2.6.

In addition the number of correct comprehension questions, the study also measured the subjective results of how participants rated the embodied agents in terms of motivation and rapport. Overall, female participants who heard the story from the adaptive agent rated the agent more positively than the participants who heard the story from the random or non-adaptive agents (Szafir and Mutlu 2012). Though I personally am hesitant to agree with the authors’ reasoning that this result came about because female participants were more likely to find the agent "cute" than male participants were, perhaps this was the case (Szafir and Mutlu 2012). Further research should certainly rely on theories of gender and education, as well as larger sample sizes, to investigate these claims in more detail.
In general, Szafir and Mutlu show that adaptive teaching methods are the best way to maintain engagement. This result is similar to Kapoor and Horvitz’s paper discussed previously, in which the dynamic decision-theoretic probing policy provided the best results over regular or random probing policies. In both cases, authors find not only that it is possible to measure when and how useful intervention will be to assessing engagement, but that the measurement which involves the most context and adaptability to the particular participant is the most successful.

This paper also provides strong evidence for educational engagement as both a psychological and computational problem to solve for. Unlike the other papers I have reviewed, Szafir and Mutlu use biological data sources to assess engagement. Ultimately, this third novel approach to modelling student engagement is hardware-dependent—authors note the low-cost of the EEG headbands, but not of the embodied agents—and thus possibly not feasible for replication in typical classrooms soon. But, its insights are valuable for further experimentation of engagement in psychology-focused contexts.

### 2.2 Conclusion

In summary, HCI research has taken a variety of paths towards solving the problem of quantifying and modelling engagement in classroom and occupational settings. Data from users’ faces, self-evaluated busyness levels, and even brain waves are viable for determining whether the user is engaged or not, and, in some cases, how much. Additionally, these studies represent a range of hardware-intensive versus software-intensive procedures towards modelling engagement. These studies I survey are often careful to note that their procedures rely on easy-to-obtain, inexpensive technologies and thus show promise for replication in other classrooms and workplaces. However, neither the three studies I survey in detail nor other papers I encounter in this research area mention the inclusion of participants who rely on assistive technologies for their schooling or jobs. This is despite the fact that over six million of students receive accommodations for disabilities in the United States alone, and nearly one in ten of those students relies on assistive technologies and services in their classroom (Quinn et al. 2009)(U.S. Department of Education and Services 2004).
Statement of The Problem

From my review of past literature related to modelling student engagement in HCI contexts, there is a noticeable lack of mention of how engagement with assistive technologies can also be measured. This is surprising, given that such technologies are used by large amounts of people and are known to be prone to engagement losses for a variety of reasons (Shinohara and Wobbrock 2011).

3.1 Problem Introduction

The papers I survey represent significant innovation in the HCI field towards quantifying engagement. For modelling something which can seem so intangible, this is certainly an impressive feat. Yet these studies each fail to ultimately be as universally-applicable in the domain of education as they attempt to be. With their claims of using low-cost and noninvasive devices for data collection, Bosch et al. and Szafrir and Mutlu portray their studies as having the potential to be extended for use in a typical American classroom (Bosch et al. 2015), (Szafrir and Mutlu 2012). And with Kapoor and Horvitz’s use of users’ own computers and consideration for the methods ranked least annoying by participants/with the fewest probes issued for the duration of the two-week experiment, it is clear that the authors seek a low-friction and scalable method for modelling engagement in everyday environments like workplaces and schools (Kapoor and Horvitz 2008).

Despite these apparent considerations for reuse in non-experimental environments, these studies do not address the new considerations which will need to be accounted for when measuring engagement in such places. Unlike participants in each of these studies, many students utilize assistive technologies which likely impact the ways they interact with their classroom environments, including their facial expressions, focus on tasks, and brain activity—all things the previous studies rely on as training data. In the United States alone, upwards of 6,500,000 students were enrolled in special education services in 2003, and this number has likely grown (Quinn et al. 2009). While the United States Department of Education does not distinguish between those who utilize assistive technology and those who receive assistive services, it estimates that between eight and nine percent of disabled students, or over half a million students by the 2003 count, receive either assistive technology or
services (U.S. Department of Education and Services 2004). However, noting that assistive technology use is practically unstudied by the United States Government, Quinn et al. find that around 31% of students with disabilities they study rely on assistive technologies in the classroom (Quinn et al. 2009). Whatever the percentage truly is, the fact is that assistive technologies are uncommon but not at all rare in typical classrooms.

To truly understand and model student engagement, HCI research must better incorporate engagement levels of students who use assistive technologies. This is both because these students are in general not yet represented in past research in this area and, most importantly, because assistive technologies come with high likelihoods of engagement issues due to misconceptions and social implications of using such technologies (Shinohara and Wobbrock 2011). In particular, electronic assistive devices, while appreciated for their ability to give users new access to information, often are disused due to their clunkiness and outdated technology as compared to state-of-the-art electronics (Shinohara and Wobbrock 2011).

### 3.2 Problem Approach

I propose a replication of one of the three studies I review, this time including participants who use assistive technologies, to assess these past studies’ true progress towards universally-applicable engagement modelling. Blind and low-vision participants make up a significant proportion of the interviews in Shinohara and Wobbrock 2011 which are conducted to understand the ways in which assistive technologies can lead to disengagement with the work and schoolwork they’re supposed to help with. These interviewees mention a plethora of tools they use daily to read, write, and navigate the world, from screen readers and braille slates and displays to white canes and guide dogs. Though these technologies are likely to throw off FACET as used in Bosch et al. 2015, these users will face even greater difficulties with the other two experimental set-ups, as they are unable to respond to the visual cues of the embodied agents in Szafr and Mutlu 2012 or the computer probes in the BusyBody system used in Kapoor and Horvitz 2008. For these reasons, I propose a replication of the Bosch et al. 2015 study with the aim of extending engagement modelling to those who use assistive technologies.

More specifically, this work will involve fine-tuning the models which draw from facial recognition data to describe engagement to account for the differences in facial expressions which come from blind and low-vision students. This will also involve testing these models on a new group of non-blind students for comparison. While it might seem like facial expressions will be much less useful data points from low- and
no-vision participants, as these students rely on these expressions much less in their daily communications, past research on vision abilities and facial expression refutes this idea (Valente et al. 2018). A first goal will be to verify that facial expressions are a viable data source for engagement for blind and low-vision students. Already, Bosch et al. find difficulty with the classroom setting’s variable lighting and students’ often unpredictable movements. So, when an image of a student now includes an assistive technology device, and these students potentially do not communicate as visually as full-sighted students, it is possible that it will be harder for FACET to recognize fine movements of the participant’s face or for these facial movements to be a good proxy for engagement. The models that Bosch et al. use to transform FACET output into an engagement measurement must be tweaked to account for the variances in facial features extracted in the presence of assistive technologies. If this is not possible, this will be evidence that Bosch et al.’s method is unable to apply universally and will hopefully spur further research in other directions.
Conclusion

From the work I survey, it is remarkable that the HCI field’s understanding of engagement boils down to some function of facial expressions, brain waves, and self-perceived busyness levels. While the uncertainty as to one single definition for "engagement" presented in K. Doherty and G. Doherty 2018 remains, the papers I review are significant attempts towards narrowing in on what engagement even is. Throughout the creation of this literature review, it has been exciting to find that computational methods are successful at modelling something so intangible as student engagement. While there is still a lot of work to go towards ensuring such measures can eventually benefit all students, these past studies show promise for identifying engagement losses and intervening so as to provide more effective education techniques. I hope my proposed work towards the goal of universally-applicable engagement modelling increases future knowledge of effective education methods for all students.
Bibliography


