



University of Pennsylvania
ScholarlyCommons

Center for Benefit-Cost Studies of Education

Graduate School of Education


8-2015

Efficiency of Automated Detectors of Learner Engagement and Affect Compared with Traditional Observation Methods

Fiona Hollands

Ipek Bakir

Follow this and additional works at: <https://repository.upenn.edu/cbcse>

 Part of the [Economics Commons](#), [Educational Assessment, Evaluation, and Research Commons](#), and the [Education Economics Commons](#)

Hollands, Fiona and Bakir, Ipek, "Efficiency of Automated Detectors of Learner Engagement and Affect Compared with Traditional Observation Methods" (2015). *Center for Benefit-Cost Studies of Education*. 4. <https://repository.upenn.edu/cbcse/4>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/cbcse/4>
For more information, please contact repository@pobox.upenn.edu.

Efficiency of Automated Detectors of Learner Engagement and Affect Compared with Traditional Observation Methods

Abstract

This report investigates the costs of developing automated detectors of student affect and engagement and applying them at scale to the log files of students using educational software. We compare these costs and the accuracy of the computer-based observations with those of more traditional observation methods for detecting student engagement and affect. We discuss the potential for automated detectors to contribute to the development of adaptive and responsive educational software.

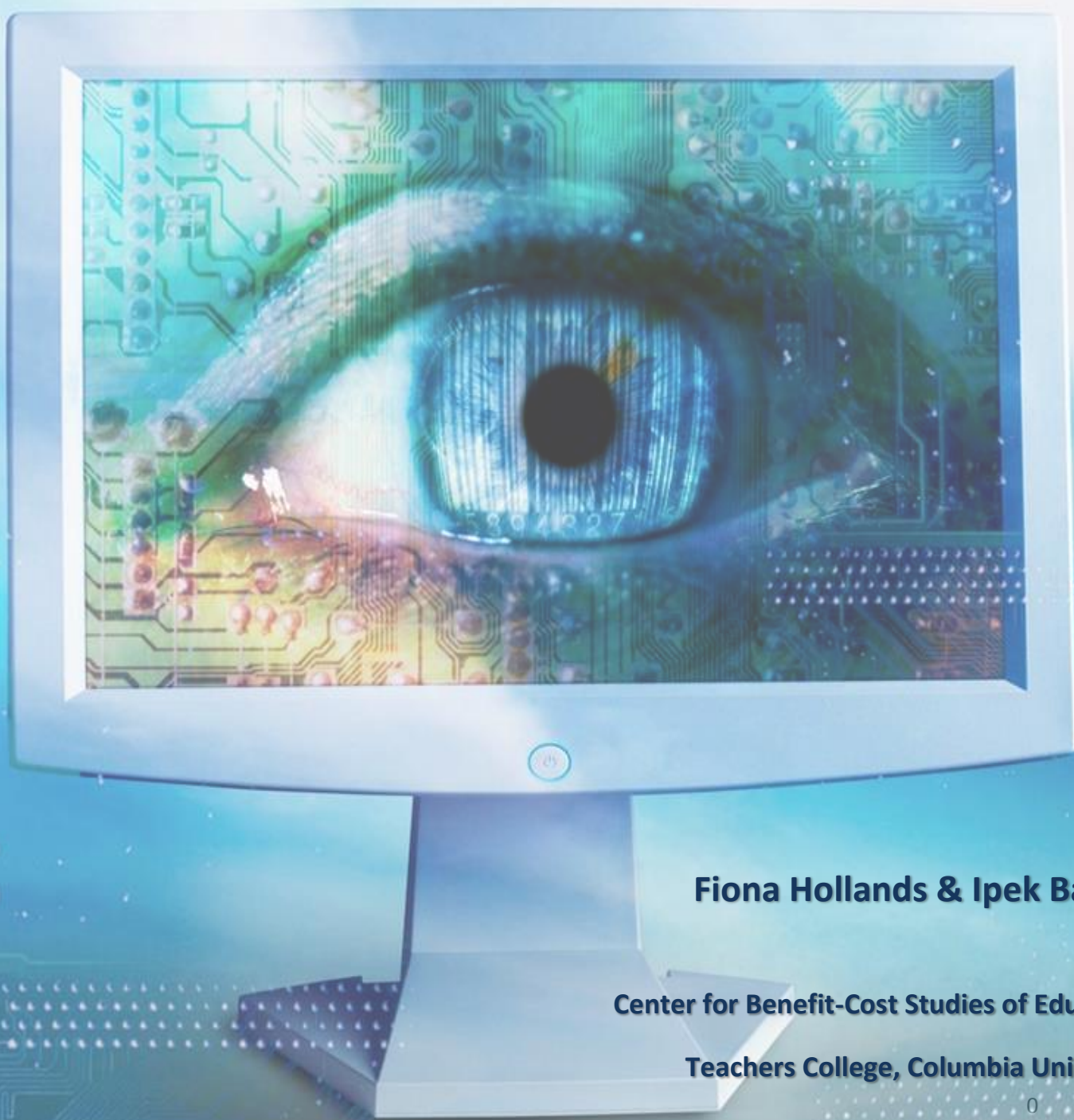
Keywords

educational technology

Disciplines

Economics | Educational Assessment, Evaluation, and Research | Education Economics

Efficiency of Automated Detectors of Learner Engagement and Affect Compared with Traditional Observation Methods



Fiona Hollands & Ipek Bakir

Center for Benefit-Cost Studies of Education

Teachers College, Columbia University

**0
August 2015**

Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods

Hollands, F.M., & Bakir, I. (2015)

fmh7@tc.columbia.edu

Center for Benefit-Cost Studies of Education, Teachers College, Columbia University

www.cbcse.org

Contents

Summary 2

Introduction 4

Methods..... 9

 Selecting observation studies 9

 Identifying ingredients 9

 Associating costs with ingredients..... 11

 Interobserver agreement as a proxy for accuracy of observations..... 11

Results..... 13

 Classroom observations using pen and paper 13

 Classroom observations using an electronic recording device..... 15

 Video analysis..... 17

 Automated detectors of engagement and affect 18

Discussion and recommendations 20

References 23

 Appendix A: Interviewees 28

 Appendix B: Ingredients and cost tables 29

Summary

Researchers and educators have consistently sought to identify factors that influence educational outcomes in the classroom and, wherever feasible, modify them to optimize the impact of educational experiences. One factor that has repeatedly been tied to academic achievement is student engagement with learning activities. If student engagement with the subject matter is important to produce learning, it is necessary to gauge the extent to which learners are engaged, to isolate factors affecting engagement, and to find ways to alter those factors to increase engagement. Detecting student engagement has historically been carried out by observing students in the field or watching video-taped learning sessions. More recently, computer scientists have developed detectors that can recognize student affect and engagement using activity patterns recorded on educational software servers.

As the use of technology for delivering instruction grows, opportunities arise to develop educational software and intelligent tutoring systems (ITSs) that adapt to individual student performance by altering the student's trajectory through the content and activities. The eventual goal is affordable personalization of learning at scale. If automated detectors of engagement and affect can be built into the software itself, the possibility arises of real-time automated responsiveness to the student's emotional state as well as to her academic performance. Given the substantial resources required to build detectors and adaptiveness into a software program, the question arises as to whether this strategy is economically viable within the price range generally tolerated for educational software. If such software is to be widely affordable to schools, the most cost-effective development strategies must be adopted and they must be applied to software used at scale.

To investigate the economic viability of investing in the development of automated detectors, we used the ingredients method to estimate and compare the costs of each of four methods of collecting observation data on student affect and engagement: classroom observations recorded using a pen and paper protocol, classroom observations recorded using a smartphone application, video analysis, and automated detectors. We provide several different cost metrics: overall cost of the study, cost per affect or engagement label assigned, cost per student observed, and cost per hour of observation.

Results indicated that costs of collecting observation data on learner engagement and affect vary widely from as little as a penny per observation label when using automated detectors applied to ASSISTments log files, to as much as \$7.36 per label for a classroom observation using a pen and paper protocol. Costs per student ranged from \$23 for automated detectors applied to ASSISTments log files or a classroom observation using a smartphone application, to \$558 per student when trained judges analyzed videos of learners. Costs per hour of observation ranged from \$4 when using automated detectors applied to ASSISTments log files to \$1,804 for a classroom observation recorded using a smartphone application. Overall study costs ranged from a few thousand dollars for classroom observations to almost \$88,000 for the development of automated detectors for ASSISTments and their application to ASSISTments log files.

Developing automated detectors of affect and engagement requires a significant upfront investment. Our cost results were reasonably consistent across two sets of detectors developed for two different ITSs: \$13,490 for each of six detectors for ASSISTments, and \$12,460 for each of four detectors for Inq-ITS. Applying the detectors to student log files costs several thousand dollars, comparable with the costs of the classroom observation studies we analyzed. However, given the ease with which the detectors can be applied to many hours of log files for many students, they can yield several hundred thousand to several million observation labels at a cost of 1-28 cents per label, \$23-\$47 per student, and \$4-\$50 per hour, with the magnitude of cost being inversely related to the scale of application.

While the low costs of applying automated detectors at scale are clearly attractive, accuracy of these detectors is less compelling. Agreement between machine-assigned labels and human coder labels averaged around 0.35 across all detectors we investigated, falling into Landis & Koch's (1977) "fair agreement" range. If automated detectors are to be built for large scale applications with thousands of learners in order to create responsive and adaptive learning environments, starting with more accurate data may lead to better academic outcomes for users due to a more appropriately responsive computer system.

We conclude that for small-scale studies of engagement and affect, in-person classroom observations recorded using either pen and paper or a smartphone application are the least costly and the most reliable. For large-scale studies, automated detectors are vastly less costly per unit of data collected but are currently low in reliability. As automated detectors become more reliable in assessing learners' affect and engagement, we expect they will be embedded in the software itself so that the learner's state can be detected real-time and the software will respond accordingly with messages, talking agents, or different activities, just as a live teacher might change pace or activity if she sees students yawning or looking puzzled.

Introduction

Importance of learner engagement and affect. Researchers and educators have consistently sought to identify factors that influence educational outcomes in the classroom and, wherever feasible, modify them to optimize the impact of educational experiences. For example, Carroll's (1963) influential model of school learning postulated five factors that influence academic achievement: the student's aptitude or time needed to learn a task, the student's ability to understand instruction, the quality of instruction, the opportunity to learn, and the student's "perseverance-in-learning" (p. 728). Carroll defined perseverance-in-learning or persistence as the amount of time the learner is willing to engage actively in learning. He described it as a function of motivation or desire to learn and of emotional variables such as frustration. [Reyes and Fennema \(1981\)](#) claimed that the most important educational influences in the mathematics classroom are teacher-student interactions and student engagement with the subject. Karweit and Slavin (1981) investigated the relationship between four different measures of time used in the classroom - scheduled time, actual instructional time, engaged time, and engaged rate - with mathematics achievement and found that the engagement measures were the most strongly related to achievement.

[Fredericks et al. \(2011\)](#) illustrate how the definition of engagement has evolved over the last 30 years, extending beyond the initial focus on behaviors such as participation and time on task to incorporate emotional or affective aspects, and "cognitive engagement" aspects. The latter include the student's investment in learning, perseverance in the face of challenge, use of deep as opposed to superficial learning strategies, and self-regulation. Numerous studies have linked student engagement at the classroom level or more broadly in the school community with educational outcomes. Fredericks, Blumenfeld, and Paris (2004), and Marks (2000) claim that engaged students are more likely to earn better grades and to perform well on standardized tests. Finn (1989) outlined a trajectory of disengagement leading to dropping out of school. Gobel (2008) observed college students in Japan using software to learn English and, while he found students to be on-task more than he expected (76% of the observations), he determined that off-task behaviors such as inactivity, surfing the internet, checking email, reading a book or magazine, or time spent gaming the system¹ were negatively correlated with students' gains on listening and reading tests. [Baker, Corbette, Koedinger, and Wagner \(2004\)](#) demonstrated a clear relationship between misuse of intelligent tutoring systems (ITSs) by students and the amount of learning that occurred. Those students who frequently "gamed" the system learned 30% less than students who used the software as intended.

Measuring engagement and affect. In order to positively influence learning outcomes, malleable factors must be measured and strategies devised to improve them. If student engagement with the subject matter is important to produce learning, it is necessary to gauge the extent to which learners are engaged, to isolate factors affecting engagement, and to find ways to alter those factors to increase engagement. Carroll (1963) suggested that the most direct evidence available for validly assessing perseverance would come from observation of the amount of time the student is actively engaged in learning. However, he asserted that, at that time, measurements of perseverance were "practically non-existent" (p.731). Since then, many protocols have been developed for the assessment of teacher and

¹ Baker, Corbett, Koedinger, and Wagner (2004) describe student activities that constitute "gaming the system" while working on an intelligent tutoring system (ITS) including asking for help multiple times until the ITS provides the correct answer, entering responses swiftly and systematically without working through the questions, and selecting every alternative in a list of multiple choice responses.

student activity in classrooms (see Simon & Boyer Eds., 1970) and specifically of student engagement (see Volpe, DiPerna, Hintze, & Shapiro, 2005; Fredericks et al., 2011). Some measures of engagement rely on student self-reports, some on teacher reports, and some on observational measures. Fredericks et al. (2004) report that studies of student engagement generally attempt to capture one or two dimensions of student engagement but that ideally all three - behavior, emotion, and cognition - should be measured.

[Nock and Kurtz \(2005\)](#) discuss advantages and disadvantages of direct observation procedures compared with other methods such as rating scales completed by teachers, by parents, or by the students themselves. They argue that direct observation is more objective, more precise for evaluating specific target behaviors, and more externally valid as it assesses behavior as it is actually occurring in the school context. On the other hand, they note that direct observation is more costly in terms of time, money, and resources because a qualified observer must be in the classroom for sustained periods of time. Furthermore, travel is often involved and the observer must be trained in the use of the observation protocol. Additional limitations of direct observation include the possibility that students being observed may act differently in the presence of an observer, that the observer may suffer from perceptual biases or from observer drift as time progresses, and that the observation only captures behaviors that occur during the observation period. Rating scales can offer a longer term view of a student's behavior. [Hintze, Volpe, and Shapiro \(2002\)](#) assert that systematic direct observation of students provides one of the most useful strategies for establishing links between assessment and intervention. An alternative to direct observation in the classroom is video-taping students individually with a webcam (e.g., [D'Mello, Taylor, Davidson, & Graesser, 2008](#)) or as a group with a mounted or handheld video recorder. The video footage is viewed and coded ex-post.

Before the widespread availability of handheld electronic devices, classroom observations were generally recorded using pen and paper observation protocols. For example, Reyes and Fennema (1981) adapted an instrument created by Romberg, Small, Carnahan, and Cookson (1979) to produce an observation protocol for evaluating student engaged time while learning mathematics. Observers using the protocol watched students sequentially in a classroom, recording an observation code every 30 seconds on a bubble sheet. Coding options included absent, engaged, off-task, and six additional codes to capture the kind of activity such as peer interaction, or engagement in a process-oriented or product-oriented mathematical task. Shapiro's (1996, 2010) Behavioral Observation of Students in Schools (BOSS) requires observers to code a student's behavior every 15 seconds over a 15-minute observation period. Coding options include active engagement, passive engagement, off-task motor, off-task passive, and off-task verbal. While earlier applications of BOSS involved recording observation codes with pen and paper, recordings can now be made electronically using a \$30 iPhone or iPad application.

More recently developed observation protocols are usually associated with electronic data collection procedures. For example, data collected by observers using the [Baker Rodrigo Ocumpaugh Monitoring Protocol](#) (BROMP) are entered directly into a smartphone using a freely available Android application, the Human Affect Recording Tool ([HART](#)) (see Ocumpaugh et al., 2015). BROMP facilitates a momentary time sampling technique at 20-second intervals, allowing simultaneous collection of data on student engagement and affect. The developers of BROMP assume that these constructs are orthogonal at least to some extent (see [Ocumpaugh, Baker, & Rodrigo, 2015](#)). HART offers a variety of customizable coding schemes but common behavioral categories include several forms of on-task and off-task activity, and gaming the system. These build on coding schemes developed by [Karweit and Slavin \(1982\)](#), and Lloyd and Loper (1986). HART affective categories include boredom, confusion, delight, engaged concentration, frustration and surprise. These were derived from work by [D'Mello, Picard, and Graesser \(2007\)](#) who hypothesized that the affective categories of boredom, confusion, frustration, eureka

experiences, and flow or engagement are more prevalent among learners working in computing environments than the commonly used basic emotions identified by Ekman and Friesen (1976, 1978): anger, fear, happiness, sadness, disgust, and surprise. In addition to expediting data collection, HART synchronizes field observations to internet time so that BROMP data can be precisely synchronized to the log file² data from educational software. This allows researchers to compare the user's observed state of affect and engagement with her specific actions in the software.

D'Mello, Duckworth, and Dieterle (under review) describe state-of-the art approaches to assessing student cognition, affect, and motivation during learning activities. These "AAA approaches" use "advanced computational techniques for the analytic measurement of fine-grained components of engagement in a fully automated fashion" (p.4). Computer-based assessments of engagement derived from sensor signals such as keystrokes, log files, facial or eye movements, posture, or electrodermal activity offer the advantage of objectivity and reliability compared with human assessments. While all "AAA" approaches require some initial labor-intensive data collection by humans, once machine-learning models have been built to detect patterns of behavior associated with specific states of affect or engagement, they can be applied at scale to new student data collected by automated sensors with low to negligible marginal costs.

D'Mello et al. (under review) distinguish between sensor-free, sensor-light, and sensor-heavy detection methods. They provide several examples of studies which implement sensor-free measurement of engagement by relying on the log files of students working on computer-based activities (D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008; [Pardos, Baker, San Pedro, Gowda, & Gowda, 2013](#); [Bixler & D'Mello, 2013](#); [Baker et al., 2012](#); and Sabourin, Mott, & Lester, 2011). Sensor-light approaches use inexpensive, ubiquitous, and relatively unobtrusive devices such as webcams or microphones to collect signals (e.g., Whitehill, Serpell, Lin, Foster, & Movellan, 2011). Sensor-heavy approaches involve expensive equipment such as eye trackers, pressure pads, and physiological sensing devices (e.g., Kapoor & Picard, 2005) which are hard to use in the field at scale. Software is used to "read" and automatically categorize the signals collected by the various sensors.

Automated sensor-free detectors are essentially sequences of computer code that are used to detect patterns of user activity in the log files that are generated by educational software platforms. These detectors are specific to the software and are developed in multiple stages. Initially, field observations are conducted and the learner's states of engagement and affect are recorded by human coders while the learner uses the software in question. These observations may also be made ex-post from video recordings. The resulting observation labels are subsequently synchronized with the user log files to match the time of the observation label with the actual keystrokes recorded in the log file. Patterns of keystrokes that are associated with particular states of engagement or affect are identified. For example, it may be the case that students who are confused repeat certain steps more frequently, students who are off-task register longer pauses, and students who are gaming the system enter answers without following intermediary steps. These patterns are used to develop programming code that can recognize the same patterns in log files collected from other learners using the same software. The detector automatically assigns a corresponding engagement or affect label to the log file data at regular intervals, usually every 20 seconds.

² Log files are time-stamped lists of events that are automatically generated by servers when users interact with software or a web site. They reflect the user's every activity (or lack thereof), pages visited, resources accessed and so on.

Accuracy of observations. Key questions for any method of assessing learner engagement and affect are the extent to which observations are reliable and valid. Ary and Suen (1983) document that, when duration of a behavior is of interest, momentary time sampling with intervals set at the shortest possible length will yield the best estimate of actual duration. With respect to reliability of data obtained through direct observation, Suen and Ary (1989) propose an evaluation of both interobserver agreement and intraobserver reliability. [Hintze \(2005\)](#) reviewed existing measures of interobserver agreement and argued that coefficient kappa, an estimate of agreement between two or more observers corrected for chance, is the most robust. To capture both interobserver agreement and intraobserver reliability, he recommends calculating an intraclass correlation coefficient that allows an evaluation of systematic variance across subjects and across observers. He also describes the application of Generalizability Theory, developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972), to observation data. This approach assesses the degree to which a set of measurements for one person generalize to a larger set of measurements for the same person. In an application of Generalizability Theory, Hintze and Matthews (2004) concluded that adequate levels of reliability with regard to learner engagement could not be attained by observing a student for 15 minutes twice per day over two weeks. They estimated that students would need to be observed four times per day over four weeks. This is clearly more time than planned in many instances of classroom observation. In practice, coefficient kappa is the most widely used measure of reliability of direct observation data (Hintze, 2005).

Validity of observation measures is difficult to assess for indirectly observable constructs such as affect and engagement because a definitive “ground truth” cannot be established. Even when acceptable levels of interobserver agreement with respect to a learner’s state of affect or engagement can be obtained, these observer judgments do not coincide well with the learner’s self-assessments. D’Mello (in press) reviewed interobserver reliability in affective computing studies and found an average kappa of 0.39 indicating only fair agreement between observers (based on Landis & Koch, 1977). Graesser, McDaniel, Chipman, Witherspoon, D’Mello, and Gholson (2006) found only slight agreement (kappa=0.12) between learners’ self-assessments of affect and the judgments of trained judges. To address validity of observations, Hintze (2005) recommends evaluating whether the data gathered on learner states correlate with other known measures of the construct being observed, whether they can predict future behavior, whether they can discriminate between groups of known status, and whether they are sensitive to changes in the learning environment. D’Mello, Duckworth, and Dieterle (under review) suggest that advanced, automated, analytic measures need to establish predictive validity, for example the ability to predict outcomes such as GPA or college graduation, and to establish external validity or generalizability to new students with different demographics. Ocumpaugh, Baker, Gowda, Heffernan, and Heffernan (2014) found that automated detectors of affect trained on a population of students from one demographic grouping did not generalize well to populations drawn from other groupings. They suggest that affective states may be susceptible to cultural variation and recommend verifying population validity of automated measures before applying them at scale.

Improving engagement. In order to improve student engagement levels in learning activities, it is necessary not only to detect disengagement, but to understand the causes well enough to be able to design corrective responses. Fredericks et al. (2004) find that engagement is higher in classrooms with supportive teachers and peers, and when students are presented challenging and authentic tasks, structure, and choice in learning activities. As the use of technology for delivering instruction has grown, opportunities have arisen to develop educational software and ITSs that adapt to individual student performance by altering the student’s trajectory through the content and activities. These efforts to automate the tailoring of instructional experiences to individual students remain relatively unsophisticated, but the eventual goal is affordable personalization of learning at scale. Researchers

have also been taking advantage of educational technology platforms to experiment with strategies to hold students' attention and keep them engaged in the learning materials. D'Mello, Craig, Fike, and Graesser (2009) developed two different embodied pedagogical agents to respond to learners' cognitive-affective states while working with the AutoTutor ITS. The "Supportive" AutoTutor is formal, empathetic, and encouraging, while the "Shakeup" AutoTutor is unconventional and attributes any detected emotions directly to the learner. Rebolledo Mendez, Du Boulay and Luckin (2005) added motivational elements to the Ecolab ITS and found that learners receiving affective feedback that varied according to the perceived cause of demotivation performed better than those receiving only cognitive feedback on their performance. Arroyo, Woolf, Royer, and Tai (2009) investigated the reaction of female students to the gender of an embedded pedagogical agent providing affective feedback in Wayang Outpost, an adaptive software program teaching math. They found that female learners exposed to embedded male "learning companions" showed more positive emotions, attitudes and learning than those exposed to pedagogical agents that provided the same feedback with a female voice.

If automated detectors of engagement and affect can be built into the software itself, the possibility arises of real-time automated responsiveness to the student's emotional state as well as her academic performance. Given the substantial programming and instructional design resources required to build detectors and adaptiveness into any one software program, the question arises as to whether this strategy is economically viable within the price range generally tolerated for educational software. If such adaptive and responsive software is to be widely affordable to schools, the most cost-effective development strategies must be adopted and they must be applied to software programs or ITSs used at scale. Cost-effective strategies in this context would be those in which the least amount of resources are used to develop responsive ITSs that lead to the greatest improvement in student learning.

Assessing costs of detectors of engagement and affect. The standard methodology for estimating costs for the purposes of economic evaluations of educational interventions is the "ingredients method" developed by [Levin \(1975\)](#) and further refined by Levin and McEwan (2001). This approach estimates the opportunity cost of all resource components required to implement the intervention. It has been applied to a wide range of educational interventions including computer-assisted instruction ([Levin & Woo, 1981](#); [Levin, Glass, & Meister, 1987](#)), blended learning programs ([Hollands, 2012](#)) and massive open online courses ([Hollands & Tirthali, 2014](#)). We set out to test the hypothesis that developing automated detectors of affect and engagement requires a high level of investment but that, if they can be applied at scale to the log files of many learners, they will produce and process observation data at a lower cost per observation label than more traditional observation methods. Large datasets of observation labels aligned to student activity and performance in the learning environment will be invaluable in the development of adaptive and responsive software.

We applied the ingredients approach to estimate the costs of developing automated detectors of affect and engagement and to apply the detectors to student log files. We compared these costs with the costs of collecting learner engagement and affect data using more traditional observation methods. The four methods we compared are: classroom observations using pen and paper observation protocols; classroom observations using a smartphone to record observations; video-taping learners and analyzing the video ex-post; and automated detectors applied to educational software user log files. We also consider the "accuracy" of the observation data. Because it is difficult to establish ground truth for observations of engagement and affect, interrater agreement, which assesses reliability, usually serves as a proxy for validity. We report coefficient kappa where available. We compare the four methods with respect to overall cost of observation studies, cost per student, cost per hour of observation, and cost

per “observation label,” where a label is defined as a single record of engagement or affect. In all cases but one the learners were observed while using computer-based educational programs.

Methods

We first reviewed the literature on learner engagement and affect to assess what methods are commonly used for detecting learner states in educational settings. We determined that the most ubiquitous methods are classroom observations recorded using pen and paper, a smartphone, a tablet, or a computer. Video analysis is also fairly common. Physiological detectors are used rarely and most often in lab situations rather than in typical classrooms due to their high costs and the difficulty of transporting and setting up the equipment in the field. Most recently, there has been a growing use of automated detectors applied to the log files generated when learners engage with computer software.

Selecting observation studies

For each of the most common observation methods, we aimed to investigate the costs of implementing at least two studies of learner affect or engagement in order to assess the potential variability in implementation costs. Our selection criteria for studies to include were:

- i) the study collected data on regular learners;
- ii) the data collected included records of learner engagement and/or affective state at intervals of 60 seconds or less;
- iii) the study was recent enough (i.e., not more than 10-12 years old) so that we could interview the researchers and reasonably expect them to recall the details of implementation to allow for acceptable accuracy in our cost estimations.

We focused on real studies in which the observation method was implemented so that we could tie the resource requirements to the number of students observed and the amount of data collected. The studies we selected and the observation codes used in each case are summarized in Table 1.

Identifying ingredients

We followed the methods laid out by Levin and McEwan (2001) to estimate the costs of educational interventions. Levin and McEwan’s ingredients approach requires the identification of all resources utilized in the implementation of an intervention and an accounting of their opportunity costs. The opportunity cost of a resource is its value as estimated by the foregone next best alternative use, which is typically represented by a market price. Note that the costs of implementing a study are therefore different from how a study is financed as many costs are not directly funded. The aim of our cost analyses was to estimate the cost of replicating the specific implementation of each study in order to collect the quantity and quality of observation data reported. In situations where the study took place in a regularly scheduled classroom setting, we considered only costs above and beyond the resources that contributed to the regular instructional activities. That is, we identified the incremental costs. For example, we did not count the costs of the classroom facility or the classroom teacher’s time because these costs would be incurred regardless of the study’s existence.

We first used information from the methods sections of each of the published studies to develop a list of ingredients (personnel, materials and equipment, facilities, or other inputs) required to implement the method of collecting observation data. We included any resources required to customize the observation instrument to the learning environment being studied, to train the observers, to set up

logistics for the observations, to collect and to summarize the data. Subsequently, for each study we contacted one or more of the authors to invite their participation in an interview to provide further details on implementation of the study. Upon receipt of a positive response, we created an extensive customized interview protocol for each person to confirm details we had already gleaned and to gather further information on personnel qualifications, work experience, and amount of effort. Personnel typically account for 70%-80% of the costs of educational interventions (Levin, 1975) and therefore merit particular attention. We also asked about types of equipment, materials, and facilities utilized and the amount of use for the study implementation, transportation needs, and so on. We included questions about the quality and quantity of the data collected over the reported periods of observation. We focused only on the resources required to collect the engagement and affect data and to process them to the point of presentation in table format. In instances where the first interviewee was not able to answer all of our questions, we interviewed additional members of the study team.

We conducted a total of 15 interviews with 11 different people, each listed in Appendix A. Nine of the interviewees were researchers or computer programmers and two were information technology personnel who could help us assess the technology resource requirements. Interviews were conducted face-to-face, by telephone, or by Skype between October 2014 and January 2015. Interviews ranged in length from 35 to 128 minutes and averaged 71 minutes in length. Most interviews were recorded. Follow-up questions or clarifications were answered via email. At the end of the study this report was circulated to the interviewees for comment.

Information from the interviews was used to finalize our ingredients list for each study implementation. We calculated the amount of each ingredient used and, based on our qualitative descriptions of each item, we identified a national average U.S. price for the ingredient sourced from a publicly available survey. National prices were used in order to make the costs directly comparable across studies. All prices were converted to 2014 dollars for consistency. Each ingredient, the amount used to implement the study, and the price were entered into the CBCSE Cost Tool Kit, a set of Excel spreadsheets developed for the purpose of estimating costs of educational programs (an online version of this tool kit is available at <http://www.cbcsecosttoolkit.org/>). The studies were all less than one year in duration so no discounting was necessary. A total cost of each implementation was calculated and divided by the number of students observed, the number of hours of observation time, and the number of observation labels collected.

For personnel ingredients we obtained national average salaries from surveys such as those issued by College and University Professional Association for Human Resources (CUPA-HR). Using the amount of time spent on the study as reported by interviewees, we calculated the appropriate percentage of total salary and added benefits using national average rates published by the Bureau of Labor Statistics. For materials and equipment costs such as computers, software, video recorders, and smartphones, we found market prices from national online distributors. Costs of durable items were spread over the number of years they are typically expected to last, for example, three years for computers. We calculated the costs of each item by multiplying the price by the fraction of available time it was used. Travel costs for observers and other researchers included the amount of time spent traveling (calculated as personnel time as above) to and from observation sites from a local residence or hotel, and costs of transportation. Car mileage allowance was obtained from the Internal Revenue Service. In situations where a trainer or observer traveled by air, we used an average U.S. domestic itinerary fare for flights from the Bureau of Transportation Statistics. We used hotel and per diem rates published by the General Services Administration.

Associating costs with ingredients

Current market rates such as national average rental rates are not typically available for school and university buildings so for facilities prices we used construction costs adjusted for costs of land, development, furnishings, and equipment, and amortized over 30 years. For example, for postsecondary office space, we found a national average construction cost per square foot in the Annual Construction Report published by the College Planning and Management magazine. We updated this cost per square foot by 33% to account for costs of land, development, furnishings, and equipment (based on College Planning and Management magazine, 2011) and amortized the costs over 30 years to obtain the equivalent of a market price per square foot per year. We asked interviewees to estimate the size of the office spaces they used for the study and the amount of use for relevant portions of the study. The cost of the space was obtained by multiplying the price per square foot per year by the number of square feet, and the fraction of time used per year. We used an interest rate of 3% for amortization, approximating the yield of 30-year U.S. Treasury Bonds.

Interobserver agreement as a proxy for accuracy of observations

Most studies we identified provided a report of interobserver agreement in the form of kappa statistics (see Cohen, 1960). According to Landis and Koch (1977), a kappa of 0.41 – 0.60 indicates moderate agreement between observers, a kappa of 0.61 – 0.80 indicates substantial agreement, and above this level is considered near perfect agreement. For the classroom observation studies we reviewed, the kappa statistics report the agreement levels between two observers. Agreement levels are expected to vary depending on the observer's amount of training and practice, and also whether observers stop periodically to discuss their judgments (see D'Mello in press). For the study that involved peer judgments of a learner's affective state from video, the kappa statistic reported agreement with the learners' self-assessments. As discussed earlier, agreement between self and an observer is invariably low. For the studies involving automated detectors, the kappas reported for the detectors indicate agreement between computer-based judgments and human coder judgments. The studies we analyzed differ in the specific constructs that were coded and it is important to note that some learner states are harder than others to judge accurately (Lehman, Matthews, D'Mello, & Person, 2008). For example, D'Mello (personal communication, July 20th, 2015) observes that differentiating engagement from a neutral state is "an extremely difficult discrimination." Among the studies in our sample, this distinction was attempted only by Graesser et al. and D'Mello et al. when assessing learner states from video. In Table 2 we show the average kappa across the various constructs observed in each study.

Table 1. Summary of Studies and Coding Options

Method and study	Learning activity	Coding options	Duration of each coding interval	Frequency of coding
Classroom observation using pen and paper				
Hintze & Matthews, 2004	Math and ELA	<i>Behavior:</i> on/off task (+/-)	Momentary	Every 15 secs
Gobel, 2008	DynEd	<i>Behavior:</i> on-task, on-task teacher/peer help, off-task non-software, off-task software help, off-task inactive, off-task gaming	60 seconds	Every 60 secs
Classroom observation using smartphone application (HART)				
Ocuppaugh et al., 2011	Reasoning Mind	<i>Behavior:</i> on task, on task conversation, off task, gaming, other <i>Affect:</i> boredom, confusion, delight, engaged concentration, frustration, other	20 seconds	Every 20 secs
Pardos et al., 2013	ASSISTments	<i>Behavior:</i> off-task, gaming, other <i>Affect:</i> boredom, frustration, engaged concentration, confusion	20 seconds	Every 20 secs
Paquette et al., 2014	Inq-ITS	<i>Affect:</i> boredom, frustration, engaged concentration, confusion, "?" (other)	20 seconds	Every 20 secs
Video analysis				
Self-judgments Graesser et al., 2006	AutoTutor	<i>Affect:</i> boredom, confusion, delight, flow, frustration, neutral, surprise	Momentary	Every 20 secs
Peer judgments Graesser et al., 2006	AutoTutor	<i>Affect:</i> boredom, confusion, delight, flow, frustration, neutral, surprise	Momentary	Every 20 secs
Trained judge judgments Graesser et al., 2006	AutoTutor	<i>Affect:</i> boredom, confusion, delight, flow, frustration, neutral, surprise	Momentary	Every 20 secs
Teacher judgments D'Mello et al., 2008	AutoTutor	<i>Affect:</i> boredom, confusion, delight, flow, frustration, neutral, surprise	Momentary	Every 20 secs
Automated detectors				
Paquette et al., 2014	Inq-ITS	<i>Affect:</i> boredom, frustration, engaged concentration, confusion	20 seconds	Every 20 secs
Pardos et al., 2013; San Pedro et al., 2013	ASSISTments	<i>Behavior:</i> off-task, gaming the system; <i>Affect:</i> boredom, confusion, engaged concentration, frustration	20 seconds	Every 20 secs

Results

Table 2 summarizes our estimated cost results for each method of collecting engagement and/or affect data. We review our findings regarding each study in detail below and provide tables showing ingredients and costs for each study individually in Appendix B.

Classroom observations using pen and paper

We estimated costs for two different studies in which data on student engagement were collected through classroom observations using pen and paper protocols. In the first study, fifth grade students were observed in math and English language arts classes. In the second study, college students were observed using DynEd intelligent tutoring software to learn English.

Observing Math and ELA: *The generalizability of systematic direct observations across time and setting: a preliminary investigation of the psychometrics of behavioral observation (Hintze & Matthews, 2004)*. The purpose of this study was to assess the reliability and validity of systematic direct observation across time and setting. Fourteen fifth-grade students in the north east U.S. were observed by graduate psychology students during math and English language arts classes and were coded as either on-task or off-task. The observers used a modified version of Shapiro's (1996) Behavioral Observation of Students in Schools. During each of 18 one-hour observation sessions, 3-4 students were observed, each for a 15-minute stretch using momentary time samples at 15-second intervals, yielding 60 data points per student over the 15 minutes. Each student was observed twice per day on each of 9 days. Five observers collectively spent 63 hours of observation time and recorded student on/off task behavior for a total of 245 fifteen-minute sessions (some students were absent for a few sessions). Fifty-five of the observation sessions were conducted by two observers to allow for inter-rater reliability checks. Kappa indices ranged from 0.31 to 0.93 for the 55 sessions, with an average of 0.65.

Ingredients used for gathering observation data in the Hintze and Matthews (2004) study and associated costs are shown in Table B1. The observers' time accounted for 60% of the costs and personnel time for training accounted for another 35%. Costs per on/off task label collected every 15 seconds were 42 cents, costs per hour of observation were \$100, and costs per student observed were \$449.

Observing DynEd: *Student off-task behavior and motivation in the CALL classroom (Gobel, 2008)*. In this study students were observed while using DynEd intelligent tutoring software to learn English in a computer-assisted language learning (CALL) classroom at a large, private university in Japan. The purpose of the study was to determine whether students' on-task or off-task behavior correlated with gains on listening and reading tests. A total of 30 mostly male students, selected at random from three classes of 50 students each, were observed and coded using 6 categories of on-task or off-task behavior while using the software in regularly scheduled sessions over a period of 4 weeks. The categories were: on-task, on-task teacher/peer help, off-task non-software, off-task software help, off-task inactive, and off-task gaming. During each class session, 10 students were observed sequentially for one minute at a time over a period of 60 minutes. The observer used pen and paper to record on a grid judgments regarding student engagement. Assessments were based on a visual observation of the student and also by viewing the student's activity in the DynEd software via a master console that could access any computer in the CALL classroom at any time. A total of 720 one-minute observations were conducted over 12 class sessions. As only one observer conducted the study, no inter-rater reliability data are available.

Table 2. Summary Table of Costs of Observation Methods

Method and study	Total cost	Hrs of observation	# of students observed	Observed time per student	Cost per student	Cost per hour	# of labels	Cost per label	Kappa index
Classroom observation with pen and paper									
Hintze & Matthews, 2004	\$6,286	63	14	270 mins	\$449	\$100	15,120	\$0.42	0.65
Gobel, 2008	\$5,302	12	30	24 mins	\$177	\$442	720	\$7.36	nm
Classroom observation with smartphone application (HART)									
Ocuppaugh et al., 2011	\$3,609	2	130	1.5 mins	\$28	\$1,804	569	\$6.34	0.68
Pardos et al., 2013	\$6,325	17	229	9 mins	\$28	\$372	6,150	\$1.03	0.79
Paquette et al., 2014	\$7,551	23	326	4.25 mins	\$23	\$328	4,155	\$1.82	0.64
Video analysis									
Self-judgments Graesser et al., 2006	\$11,548	15	28	30 mins	\$412	\$770	2,688	\$4.30	na
Peer judgments Graesser et al., 2006	\$11,548	15	28	30 mins	\$412	\$770	2,688	\$4.30	0.06*
Trained judge judgments Graesser et al., 2006	\$15,621	15	28	30 mins	\$558	\$1,041	2,688	\$5.81	0.31
Teacher judgments D'Mello et al., 2008	\$11,898	15	28	30 mins	\$425	\$793	2,688	\$4.43	0.12
Automated detectors									
Paquette et al., 2014	\$56,476	1,139	1,196	57 mins	\$47	\$50	204,960	\$0.28	0.35**
Pardos et al. 2013, & San Pedro et al., 2013	\$87,576	19,511	3,747	625 mins	\$23	\$4	7,023,776	\$0.01	0.34**

nm = not measured; na = not applicable. *This kappa indicates agreement with self-judgments. ** These kappas indicate agreement between automated detector assessment and human coders.

Ingredients used in the Gobel (2008) study and associated costs are shown in Table B2. The observer's time accounted for 85% of the costs. Costs per on/off task label collected every 60 seconds were \$7.36, costs per student were \$177, and costs per hour of observation were \$442. Costs per label and costs per hour were much higher than in the Hintze and Matthews (2004) study because the observer was a university professor as opposed to graduate students who were paid by the hour and received no benefits. Additionally, only one label was recorded per minute in the Gobel study compared with four per minute in the Hintze and Matthews study. If labels were collected every 15 seconds in the Gobel study, the costs per label would fall to \$1.84. If a graduate student conducted the observations and collected four labels per minute, costs would fall to \$1.15 per label. Costs per student were, however, lower for Gobel's study because twice as many students were observed and each for less total time (24 minutes vs. 270 minutes).

Classroom observations using an electronic recording device

We estimated observation costs for three studies in which data on student engagement and/or affect were collected through classroom observations using an electronic recording device. In the first study, elementary school students were engaged in the use of Reasoning Mind mathematics software; in the second study, middle school students were using another computer-based math program, ASSISTments; and in the third study, eighth-grade students were observed using Inq-ITS, an inquiry-based science software program.

Observing Reasoning Mind: *Field Observations of Engagement in Reasoning Mind (Ocumpaugh, Baker, Gaudino, Labrum, & Dezen Dorf, 2011)*. In this study, field observations were conducted to evaluate student engagement and affect while working on Reasoning Mind software. Reasoning Mind is a game and problem-solving based software package that teaches mathematics to elementary school students. Certified observers used BROMP to record student engagement and affective state. Behavior states coded were: on-task, on-task conversation, off-task, or gaming. Affective states coded were: boredom, confusion, delight, engaged concentration, or frustration. Students were observed in two classrooms from each of three schools in the Texas Gulf Coast region. Two schools were urban with around 25 students per class and one school was a suburban charter with approximately 15 students per class. The total number of students observed was 130. During each observation session the observer watched each student in the class sequentially for 20 seconds and recorded a judgment of affective state and of behavior state simultaneously at the end of the 20 seconds. Judgments were recorded using a smartphone application, the Human Affect Recording Tool (HART). If more than one behavior or affective state was observed during the 20 seconds, only the first was recorded. In situations that were ambiguous or if the student left the room, "Other" was recorded. Trainee coders were also present and inter-rater reliability recorded was $\kappa = 0.58-0.72$ for affect and $\kappa = 0.63-0.79$ for behavior. However, only the trainer data were included in the analysis. Accordingly, we did not include the trainees in our cost estimate. The researchers found that observed students were on task 82% of the time and in a state of engaged concentration 71% of the time.

Ingredients used in this study and associated costs are shown in Table B3. Over half of the costs were attributable to training the observer in the use of BROMP. The observer's observation time accounted for 25% of the costs, and costs of air travel, hotel, and per diem accounted for 23%. Costs per label (one affect and one behavior label collected every 20 seconds) were \$6.34. Observations labeled "Other" were not included in this estimate. Costs per student were low at \$28 as 130 students were each observed for a total of only 1.5 minutes, but costs per hour of observation time were very high at \$1,804 because all costs were spread over just 2 hours of total observation time for the study.

BROMP training costs are further broken down in Table B4. Training in how to assess student affect and behavior and record it with the HART application lasts two days and is usually conducted one-on-one until an acceptable level of interobserver agreement is attained between trainer and trainee during practice observations. Training often involves travel costs for the trainer. In our analyses we attribute all costs of BROMP training to the one study being analyzed. However, if the observers used BROMP in multiple observation studies, the costs could be spread across the number of instances.

Observing ASSISTments: *Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes.* (Pardos, Baker, San Pedro, Gowda, & Gowda, 2013). The purpose of this study was to analyze student behavior when using ASSISTments, a web-based tutoring platform for 7th-12th grade mathematics, and to use these data to develop automated detectors of engagement and affect which could be used to predict end-of-year learning outcomes. Here we address only initial collection of observation data. We report on the development of automated detectors for ASSISTments in a later section. Using the BROMP method, field observations of student affect and engagement were conducted by two trained observers over three days with 229 students at an urban middle school in Massachusetts. Judgments were recorded using HART. Students in the classroom were observed serially for 20-second intervals and codes were recorded for behavior (off-task behavior, gaming, other behavior) and affective state (boredom, frustration, engaged concentration, or confusion). Engaged concentration was observed 65% of the time and off-task behavior 22% of the time. Inter-rater reliability was assessed for 51 of the total 6,150 coding instances and was high for affect codes ($\kappa=0.86$) and acceptable for behavior codes ($\kappa=0.72$).

Ingredients used in this study and associated costs are shown in Table B5. As with the [Ocumpaugh et al. \(2011\)](#) study, over half of the costs (59% in this case) were attributable to training the observers in the use of BROMP. The remaining costs were almost all attributable to observation time. The total observation costs for this study (\$6,325) were almost twice those for the Ocumpaugh et al. study (\$3,609) because two trained observers collected the data over three days rather than one observer working over three days. The number of students observed in Pardos et al. (2013) was almost twice the number observed by Ocumpaugh et al. and each one was observed for a total of nine minutes rather than 1.5 minutes. The costs per student were the same in both studies at \$28. However, the costs per label (one affect and one behavior label collected every 20 seconds) were six times lower in the Pardos et al. study at \$1.03 and the costs per hour of observation were almost five times lower at \$372. These two results reflect economies of scale as the costs of training are spread over more data points and more hours of observation. Each of the two observers was able to collect over five times the amount of data as the one person observing students using Reasoning Mind. This increase in efficiency may be partially explained by the fact that students observed while using ASSISTments were all located in one school while the Reasoning Mind observer needed to travel between three schools.

Observing Inq-ITS: *Sensor-free affect detection for a simulation-based science inquiry learning environment* (Paquette, Baker, San Pedro, Gobert, Rossi, Nakama, & Kauffman-Rogoff, 2014). In this study four expert field observers coded student affective states while the students used Inq-ITS, a web-based, inquiry-oriented environment offering interactive simulations in physical, life, and earth science topics. Observations were conducted across 11 eighth-grade classrooms in three schools in Massachusetts. The observers used the BROMP method and entered codes in a Google Android device using the HART application. Observers collected 4,155 affect labels. Coding options were: boredom, frustration, engaged concentration, confusion, or “?” for indeterminate or other. Of these 4,155 labels, 22% were coded as engaged concentration, 3% as boredom, and 1% each as confusion and frustration. Interobserver agreement was assessed for three pairs of observers and an average kappa of 0.64 was reported.

Ingredients used in this study and associated costs are shown in Table B6. Almost 70% of the costs were attributable to training the four observers in the use of BROMP and 29% to their observation time. Two observers conducted observations over two days and the other two only collected data for one day each. Costs per label (one affect label collected every 20 seconds) were \$1.82. This is higher than the cost per label for the ASSISTments observations partly because only one affect label was collected every 20 seconds whereas Pardos et al. (2013) collected both an affect and an engagement label every 20 seconds. Additionally, the four observers of ASSISTments traveled between three schools and were able to collect fewer data points than two observers working intensively at one school. Furthermore, because of relatively high training costs, the use of four trained observers each averaging 1.5 days of observations was less efficient than two trained observers each conducting three days of observation. Costs per hour of observation were \$328 and costs per student were \$23 reflecting some economies of scale as more students were observed over more hours than in the studies of ASSISTments and Reasoning Mind.

Video analysis

To estimate costs of assessing student affect using video analysis, we used two related studies that compared the reliability of affect judgments made by learners themselves, by peers, by teachers, and by trained judges. Judgments were made by viewing a collection of half-hour long video-tapes of each of 28 college-level learners interacting with AutoTutor, a software program that teaches computer literacy topics. While we accounted for costs of the lab and equipment used for this study as it was conducted outside of regular classroom time, we did not assign any incremental value to the students' time as participation in such studies was required as part of their degree programs. The learners' faces were video-taped and their screen activities were recorded using Camtasia screen-capture software. Subsequently, one of the following affective states was coded every 20 seconds: boredom, confusion, delight, flow, frustration, neutral, or surprise. In total, among the 28 students, 2,688 coded states were recorded. In the first study (Graesser et al., 2006), self-judgments were compared with those of peers and trained judges. In the second study (D'Mello, Taylor, et al., 2008), self-judgments were compared to those of master teachers.

Observing AutoTutor: *Detection of emotions during learning with AutoTutor (Graesser, McDaniel, Chipman, Witherspoon, D'Mello, & Gholson, 2006)*. In this study, self-judgments of the AutoTutor learners' affective states were compared to the judgments of peers and of trained judges. The AutoTutor learners were asked to review the video-tapes of themselves and code their own affective states at 20-second intervals of the replayed video. Subsequently, the learners were each asked to judge the affective states of a video-taped peer, also at 20-second intervals. Finally, a pair of judges trained in the Facial Action Coding System (Ekman & Friesen, 1976; 1978) each coded the videos. Graesser et al. found the highest agreement between the two trained judges ($\kappa = 0.31$) but that agreement between self-judgments and trained judge judgments was low, averaging $\kappa = 0.12$. Self-judgments almost never matched with peer judgments ($\kappa = 0.06$).

Ingredients used in this study and associated costs are shown in Table B7. In order to compare the efficiency of different judges, we first estimated the data collection costs that applied to all situations equally and then added the costs associated with each set of affect judges. Data collection costs accounted for 81% of the total costs when self or peer judgments were used and 60% of the costs in the case of trained judges. Costs of self-judgments and peer judgments were the same as the time and personnel involvement were equivalent in the two situations. Total costs for data collection and self- or peer judgments were \$412 per student observed, \$770 per hour of observation time, or \$4.30 per affect label assigned every 20 seconds. Total costs when trained judges were involved were higher due to the

time spent on FACS training and the greater cost of the trained observers' time: \$558 per student observed, \$1,041 per hour of observation time, or \$5.81 per affect label.

Observing AutoTutor: Self versus teacher judgments of learner emotions during a tutoring session with AutoTutor (D'Mello, Taylor, Davidson, & Graesser, 2008). In this study, self-judgments of the AutoTutor learners' affective states were compared to the judgments of two middle school master teachers. The teachers coded only half of each video due to time constraints. The researchers compared the inter-judge reliability for the two teachers and for each teacher against the student self-judgments. They found that the teacher judgments did not match well with each other ($\kappa = 0.123$), and matched even less well with the students' self-judgments ($\kappa_{\text{Teacher 1-student}} = 0.076$; $\kappa_{\text{Teacher 2-student}} = 0.027$). They concluded that even accomplished teachers do not accurately assess the affective states of learners.

Ingredients used in this study and associated costs are also shown in Table B7. Data collection costs accounted for 78% of the total costs when teacher judgments were used. As before, total costs for data collection and self-judgments were \$412 per student observed, \$770 per hour of observation time, or \$4.30 per affect label assigned every 20 seconds. Total costs when teachers were involved were higher due to the greater costs of their time: \$425 per student observed, \$793 per hour of observation time, or \$4.43 per affect label. However, because the teachers did not undergo FACS training, the costs were lower than for trained judges.

Automated detectors of engagement and affect

We investigated the costs of developing automated detectors of affect and engagement for ASSISTments based on Pardos et al. (2013) and of developing automated detectors of affect for Inq-ITS based on Paquette et al. (2014). In the case of Inq-ITS, four detectors were built to detect each of the following affective states: boredom, frustration, engaged concentration, and confusion. For ASSISTments, six detectors were built. Four of these detected the affective states of boredom, frustration, engaged concentration, and confusion. Two detected behavioral states: off-task, and gaming the system. We also include the costs of applying the detectors to new log files based on [San Pedro, Baker, Bowers, and Heffernan \(2013\)](#) to render the equivalent of an observation study in which the data are collected and summarized in table format. This allows comparability of the costs of observation with those of the other methods we analyzed.

The first step in the development of automated detectors is to collect in-person observation labels either through direct classroom observations or video analysis. Interviewees estimated that several hundred observation labels are needed to develop a detector, for example, Sujith Gowda suggested 800 or more to build an ASSISTments affect detector. This first step is documented above in the section titled **Classroom observations using an electronic recording device**. In this section we address the second and third steps of building the detectors and applying them to new data. Costs of all steps were combined for a total cost for the development and application of automated detectors of affect and engagement.

Automated detectors of ASSISTments: Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. (Pardos, Baker, San Pedro, Gowda, & Gowda, 2013); Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. (San Pedro, Baker, Bowers, & Heffernan, 2013). Pardos et al. describe the process of building the six detectors used for assessing student affect and engagement while working with ASSISTments. San Pedro et al. describe how the detectors were applied to the action log files of 3,747 middle school students from three districts in New England. The students used ASSISTments

systematically throughout the school year and log files were collected mostly from school years 2004-2005 through 2006-2007, with a few from the following two school years. Pardos et al. reported kappa statistics for the six detectors which averaged 0.344. This represents the degree to which the engagement and affect labels assigned by the machine-learning model matched the labels assigned by the human coders in the initial data collection phase.

Ingredients and their associated costs for the data collection step were reported in Table B5 as discussed above. Table B8 reports the ingredients and associated costs for the second step of building the detectors. These amounted to \$74,620. Personnel costs for programmers accounted for 90% of the costs of detector development. Adding together costs of the first and second steps, total costs to collect observation data and build the six detectors for ASSISTments were \$80,950 or \$13,490 per detector.

To estimate the costs of applying the detectors to a new set of student log files to assess affect and engagement, we calculated the labor costs for ten days' worth of a research programmer's time and one hour per day of a supervising programmer's time. This was based on Sujith Gowda's estimate of 5-15 days to apply detectors to new log files, depending on the size of the files. Adding costs of facilities, equipment, and materials, the costs for applying the detectors to new log files were \$6,630, bringing the total costs of developing and applying the six detectors to \$87,580.

San Pedro reported that the total time logged for the 3,747 students was 19,510 hours (personal communication, April 2nd, 2015). With one affect and one engagement label assigned every 20 seconds, over 7 million observation labels were obtained from the log files. The cost of "observing" each student for affect and engagement was therefore \$23, the cost per hour of observation was \$4, and the cost per observation label was just over a penny. Clearly, while the detectors were costly to develop initially, the ease with which they can be applied at scale renders the costs per label and costs per hour of observation significantly lower than other data collection methods. This illustrates the economies of scale achieved in applying the detectors to massive amounts of data. Costs per student were around the same as for collecting observation data using HART, but the automated detectors "observed" 625 minutes on average per student while the HART observations observed students for between 1.5 and 9 minutes.

In calendar year 2014, 61,609 students used ASSISTments, logging a total of 14,757,331 hours or 240 hours per student (Yutao Wang, personal communication, April 5th, 2015). If the detectors were applied to all these log files and we assume that the cost of applying the detectors to this amount of data increased ten-fold from \$6,630 to \$66,300 (conservatively allowing for around 3 months of data processing time), the costs of observing each student for engagement and affect would fall to around \$2.40 per student and a penny per hour. Over five billion affect and engagement labels would be assigned at a cost of less than one hundredth of a penny per label.

Automated detectors of Inq-ITS: Sensor-free affect detection for a simulation-based science inquiry learning environment (Paquette, Baker, San Pedro, Gobert, Rossi, Nakama, & Kauffman-Rogoff, 2014).

The collection of in-person observation data on student affect while using Inq-ITS was described earlier. Subsequently, multiple computer programmers were involved in developing the automated detectors. Their tasks included cleaning the data files, synchronizing the observation labels with the Inq-ITS log files so that affect labels could be matched to user keystroke patterns, identifying patterns in the data that appeared to indicate a particular affective state ("feature engineering"), writing the machine learning algorithms to identify and count the instances of each pattern in the log files, and finally applying the detectors to new log file data to obtain machine-generated predictions of students' affective state based on their keystrokes.

Ingredients and their associated costs for the data collection step were reported in Table B6 as discussed above. Table B9 reports the ingredients and associated costs for the second step of building the detectors. Personnel costs accounted for 98% of the costs of detector development, with the programmer who built the detectors accounting for the largest share of costs. Added together, the costs to collect observation data and to build the four detectors of affect for Inq-ITS were \$49,850 or \$12,460 per detector.

Paquette et al. (2014) did not report a specific application of the detectors to new log files so we calculated cost per student and cost per label by assuming that the detectors could be applied to all Inq-ITS learner log files collected over two academic years (2012-13 and 2013-14). Over these two years, 1,196 students used Inq-ITS for a total of 68,320 minutes or 1,139 hours - just under an hour per student (Ryan Baker, personal communication March 11th, 2015). Applying the affect detectors to these log files at 20-second intervals would yield almost 205,000 observation labels (68,320 minutes x three 20-second intervals per minute = 204,960 labels). We assumed that the costs of applying the Inq-ITS detectors to new log files were the same as the costs estimated for applying the ASSISTments detectors to new log files (\$6,630). This assumption is conservative as it is probable that the costs would be lower given the smaller amount of data. Under this assumption, the total costs of developing and applying the Inq-ITS detectors were \$56,480. The costs of “observing” each student for affect were \$47, the costs per hour of observation were \$50, and the costs per observation label were 28 cents. In terms of accuracy, Paquette et al. report an average kappa statistic across the four detectors of 0.354. This represents the degree to which the affect labels assigned by the machine-learning model matched the affect labels assigned by the human coders in the initial data collection phase.

Discussion and recommendations

We reported cost estimates for each of four methods of collecting observation data on student affect and engagement: classroom observations recorded using a pen and paper protocol, classroom observations recorded using a smartphone application, video analysis, and automated detectors. We provide several different cost metrics: overall cost of the study, cost per affect or engagement label assigned, cost per student observed, and cost per hour of observation. Results indicated that costs of collecting observation data on learner engagement and affect vary widely from as little as a penny per observation label when using automated detectors applied to ASSISTments log files, to as much as \$7.36 per label for a classroom observation using a pen and paper protocol. Costs per student ranged from \$23 for automated detectors applied to ASSISTments log files or a classroom observation using HART, to \$558 per student when trained judges analyzed videos of learners. Costs per hour ranged from \$4 when using automated detectors applied to ASSISTments log files to \$1,804 for a classroom observation recorded using a smartphone application (although this particular study appeared to be an outlier as explained below). Overall study costs ranged from a few thousand dollars for classroom observations to almost \$88,000 for the development of automated detectors for ASSISTments and their application to ASSISTments log files.

Within each of the four observation methods we considered, results varied substantially depending on factors such as the number of students and schools involved, the total observation time planned, the effort required to develop an observation instrument, the amount of training required for the observers, the types of personnel involved, and whether travel to the observation site was necessary.

One study that involved classroom observations with a pen and paper protocol (Hintze & Matthews, 2004) yielded relatively low costs per label and low costs per hour of observation compared with other methods (\$0.42 and \$100 respectively). This was because graduate students collected the observation data, an existing observation protocol was used with only minor modifications, supervision requirements were negligible, training costs were fairly low because the students were trained together and for only half a day, and travel costs were minimal. However, because each learner was observed for a substantial amount of time (270 minutes), costs per student were the second highest among all studies at \$449. The second study we analyzed that involved classroom observations with a pen and paper protocol (Gobel, 2008) yielded the highest cost per observation label (\$7.36). This was primarily because it involved a professor conducting the observation and observation labels were assigned only every 60 seconds rather than every 15 or 20 seconds as in the other studies we analyzed.

We analyzed three studies in which classroom observations were conducted with a smartphone application (HART) being used to record the observation labels. Costs per student were similar across the three studies (\$23-\$28) and among the lowest across all methods because students were observed for only a few minutes each in total. The Ocumpaugh et al. (2011) study yielded a high cost per label (\$6.34) and the highest cost per hour of observation across all methods (\$1,804) because it collected the fewest labels and total observation time was the lowest at only 2 hours. Given the significant costs of BROMP training and air travel, this study suffered from diseconomies of scale. The other two studies in this category, Pardos et al. (2013) and Paquette et al. (2014), collected several thousand observation labels each over 17-23 hours and yielded among the lowest costs per label (\$1.03 and \$1.82 respectively), and per hour of observation (\$372 and \$328 respectively). Costs per label for Paquette et al. were 75% higher primarily because for each coding interval only one label was assigned for affect while Pardos et al. assigned one for affect and one for behavior at each coding interval, doubling the yield of labels.

Studies that involved classroom observations as opposed to video analysis or automated detectors were the lowest cost overall, ranging from around \$3,500-\$7,500. Inter-rater reliability was more or less comparable for observations recorded using a pen and paper protocol and those recorded using a smartphone application. All of them fell into Landis and Koch's (1977) "substantial agreement" range, with one achieving a kappa at the top of this range, most likely because the observers were more experienced in the use of the observation protocol.

The studies that involved video analysis were more costly overall than the classroom observations, ranging between \$11,500 and \$15,500, with costs increasing as judgments of affect were made by teachers instead of students and then by trained judges instead of teachers. The costs per label were in the middle of the range across all methods but the costs per student and costs per hour of observation were close to the highest as relatively few students were observed. The inter-rater reliability for each of the video analysis studies was low, falling into Landis and Koch's (1977) "slight" or "fair" agreement range. This may be partially explained by the fact that these studies included a "neutral" construct which, according to D'Mello (personal communication, July 20th, 2015), is hard to assess accurately. Other studies involving video analysis have reported substantial interobserver agreement for constructs that are easier to assess such as happiness, frustration, and anxiety (see Lehman et al. 2008).

Developing automated detectors of affect and engagement requires a significant upfront investment. Our cost results were reasonably consistent across two sets of detectors developed for two different ITSs: \$13,490 for each of six detectors for ASSISTments and \$12,460 for each of four detectors for Inq-ITS. Applying the detectors to student log files costs several thousand dollars, comparable with the costs of the classroom observation studies we analyzed. However, given the ease with which the detectors

can be applied to many hours of log files for many students, they can yield several hundred thousand to several million observation labels at a cost of 1-28 cents per label, \$23-\$47 per student, and \$4-\$50 per hour, with the magnitude of cost being inversely related to the scale of application.

While the low costs of applying automated detectors at scale are clearly attractive, accuracy of these detectors is less compelling. Agreement between the machine-assigned labels and the human coder labels averaged around 0.35 across all detectors, falling into Landis & Koch's (1977) "fair agreement" range. One strategy we recommend trying in order to improve the detectors' accuracy is to collect the initial observation data using two experienced observers who display a high level of interobserver agreement and subsequently only use the observation labels for which they show agreement to develop the automated detectors. Furthermore, given Hintze and Matthews' (2004) suggestion that students need to be observed four times per day for 15 minutes over four weeks in order to assure that the assessment reflects the learner's behavior in general, more extensive initial data collection per student should yield more reliable assessments of student affect and engagement while using an ITS. Additionally, given Ocumpaugh et al.'s (2014) finding that automated detectors developed using data from a population of students belonging to one demographic grouping did not generalize well to populations drawn from other groupings, we recommend further investigation of whether detectors need to be built specific to a population. This strategy would likely be more costly than building a universal set of detectors using data collected across several populations, but it may yield higher accuracy in assigning states of affect and engagement.

An unresolved issue with respect to any observation method is the question of how well it can truly assess engagement and affect, that is, how close the method can get to ground truth with respect to the learner's state. D'Mello suggested to us that the closest one might get to ground truth is by using a combination of physiological sensors and self-assessments to capture a predictable response to a contrived stimulus. While this would be prohibitively costly for most purposes, if automated detectors are to be built for large scale applications with thousands of learners in order to create responsive and adaptive learning environments, starting with more accurate data may lead to better academic outcomes for users due to a more appropriately responsive computer system.

We conclude that for small-scale studies of engagement and affect, in-person classroom observations recorded using either pen and paper or a smartphone application are the least costly and the most reliable. For large-scale studies, automated detectors are vastly less costly per unit of data collected but are currently low in reliability. As automated detectors become more reliable in assessing learners' affect and engagement, we expect they will be embedded in the software itself so that the learner's state can be detected real-time and the software will respond accordingly with messages, talking agents, or different activities, just as a live teacher might change pace or activity if she sees students yawning or looking puzzled.

⌘ ⌘ ⌘ ⌘ ⌘ ⌘ ⌘ ⌘

References

- Arroyo, I., Woolf, B. P., Royer, J. M., & Tai, M. (2009). Affective gendered learning companions. In *AIED* (Vol. 200, pp. 41-48). doi: 10.3233/978-1-60750-028-5-41
- Ary, D., & Suen, H.K. (1983). The use of momentary time sampling to assess both frequency and duration of behavior. *Journal of Behavioral Assessment*, 5(2), 143-150.
- Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the Cognitive Tutor classroom: when students “game the system.” *Proceedings of ACM CHI 2004: Computer-Human Interaction* (pp. 383–390). New York, USA. Retrieved from <http://www.columbia.edu/~rsb2162/p383-baker-rev.pdf>
- Baker, R. S. J., Kalka, J., Aleven, V., Rossi, L., Gowda, S. M., Wagner, A. Z., . . . Ocumpaugh, J. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In K. Yacef, O. Zaïane, H. HersHKovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 126-133): International Educational Data Mining Society. Retrieved from <http://www.columbia.edu/~rsb2162/EDM%20Affect%20Detection%20V22%20final.pdf>
- Bixler, R., & D'Mello, S. (2013). Detecting engagement and boredom during writing with keystroke analysis, task appraisals, and stable traits. *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI 2013)* (pp. 225-234). New York, NY: ACM. Retrieved from http://delivery.acm.org/10.1145/2450000/2449426/p225-bixler.pdf?ip=160.39.78.237&id=2449426&acc=ACTIVE%20SERVICE&key=7777116298C9657D%2ECCAFA7F43E96773E%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=693996014&CFTOKEN=64686623&_acm_ =1437149254_131f8b2aa6c79676c434fe4743e667ad
- Carroll, J.B. (1963). A model of school learning. *Teachers College Record*, 64(8), 723-731.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- College Planning and Management (June, 2011). *Living on campus. Trends and analysis*. College Planning & Management.
- Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.
- D'Mello, S. K. (in press). On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*.
- D'Mello, S., Craig, S., Fike, K., & Graesser, A. (2009). Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In *Human-computer interaction. Ambient, ubiquitous and intelligent interaction* (pp. 595-604). Berlin, Heidelberg: Springer.

- D'Mello, S. K., Taylor, R., Davidson, K., & Graesser, A. (2008). Self versus teacher judgments of learner emotions during a tutoring session with AutoTutor. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems* (pp. 9-18). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://www.researchgate.net/profile/Roger_Taylor/publication/225380577_Self_Versus_Teacher_Judgments_of_Learner_Emotions_During_a_Tutoring_Session_with_AutoTutor/links/02e7e520999338cedc000000.pdf
- D'Mello, S. K., Picard, R. W., & Graesser, A. C. (2007). Towards an affect-sensitive AutoTutor. *Special issue on Intelligent Educational Systems – IEEE Intelligent Systems*, 22(4), 53-61. Retrieved from <http://affect.media.mit.edu/pdfs/07.dmello-et-al.pdf>
- D'Mello, S., Duckworth, A., & Dieterle, E. (under review). Advanced, analytic, automated measures of person-oriented components of engagement during learning.
- D'Mello, S. K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1-2), 45-80.
- Ekman, P., & Friesen, W.V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1(1), 56-75.
- Ekman, P. & Friesen, W.V. (1978). *Investigator's guide to the Facial Action Coding System, Part II*. Palo Alto, CA: Consulting Psychologists Press.
- Finn, J. (1989). Withdrawing from school. *Review of Educational Research*, 59(2), 117-142.
- Fredericks, J.A., Blumenfeld, P.C., & Paris, A.H. (2004). School engagement: potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109.
- Fredericks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). *Measuring student engagement in upper elementary through high school: a description of 21 instruments*. (Issues & Answers Report, REL 2011–No. 098). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/rel_2011098.pdf
- Gobel, P. (2008). Student off-task behavior and motivation in the CALL classroom. *International Journal of Pedagogies and Learning*, 4(4), 4-18.
- Graesser, A.C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, (pp. 285-290), Vancouver, Canada.
- Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: a preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, 33(2), 258-279.

- Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, 34(4), 507.
Retrieved from http://www.researchgate.net/profile/John_Hintze/publication/238546006_Psychometrics_of_Direct_Observation/links/54d388ab0cf2b0c6146da9d6.pdf
- Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2002). Best practices in the systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 2, pp. 993-1006). Bethesda, MD: National Association of School Psychologists. Retrieved from <http://www.emporia.edu/~persingj/systematicobservation.pdf>
- Hollands, F.M. (2012). Using cost-effectiveness analysis to evaluate School of One. Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.
Retrieved from <http://cbcse.org/wordpress/wp-content/uploads/2013/09/2012-Hollands-CEA-to-evaluate-School-of-one.pdf>
- Hollands, F. M., & Tirthali, D. (2014). Resource requirements and costs of developing and delivering MOOCs. *The International Review of Research in Open and Distance Learning*.15(5), 113-132.
Retrieved from <http://www.irrodl.org/index.php/irrodl/article/view/1901>
- Kapoor, A., & Picard, R. W. (2005, November). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 677-682). ACM.
- Karweit, N., & Slavin, R. E. (1981). Measurement and modeling choice in studies of time and learning. *American Educational Research Journal*, 18(2), 157-171.
- Karweit, N., & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Educational Psychology*, 74(6), 844. Retrieved from <http://psycnet.apa.org/journals/edu/74/6/844.pdf>
- Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.
- Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008, January). What are you feeling? Investigating student affective states during expert human tutoring sessions. In *Intelligent Tutoring Systems* (pp. 50-59). Springer Berlin Heidelberg.
- Levin, H. M. (1975). Cost-effectiveness analysis in evaluation research. In M. Guttentag & E. Struening (Eds.), *Handbook of evaluation research* (Vol. 2, pp. 89-122). Beverly Hills, CA: Sage. Retrieved from http://cbcse.org/wordpress/wp-content/uploads/2013/05/Levin1975_CEA-in-Evaluation-Research.pdf
- Levin, H. M., Glass, G. V., & Meister, G. (1987). Cost-effectiveness of computer-assisted instruction, *Evaluation Review*, 11(1), 50-72. Retrieved from <http://cbcse.org/wordpress/wp-content/uploads/2013/02/1987-Levin-Computer-Assisted-Instruction.pdf>
- Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

- Levin, H. M. & Woo, L. (1981). The costs of computer-assisted instruction. *Economics of Education Review*, 1(1), 1-25. Retrieved from <http://cbcse.org/wordpress/wp-content/uploads/2013/05/Levin-and-Woo-1981.pdf>
- Lloyd, J.W., & Loper, A.B. (1986). Measurement and evaluation of task related learning behavior: attention to task and metacognition. *School Psychology Review*, 15(3), 336-345.
- Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle and high school years. *American Educational Research Journal*, 37(1), 153–184.
- Nock, M. K., & Kurtz, S. (2005). Direct behavioral observation in school settings: bridging science to practice. *Cognitive and Behavioral Practice*, 12(3), 359–370. Retrieved from http://www.researchgate.net/publication/222705078_Direct_behavioral_observation_in_school_settings_Bringing_science_to_practice
- Ocuppaugh, J., Baker, R. S. J. d., Gaudino, S., Labrum, M. J., & Dezendorf, T. (2011) Field observations of engagement in Reasoning Mind. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, (pp. 624-627), Memphis, USA. Retrieved from http://www.columbia.edu/~rsb2162/aied2013_submission_37.pdf
- Ocuppaugh, 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.
- Ocuppaugh, J., Baker, R.S., & Rodrigo, M.M.T. (2015). *Baker Rodrigo Ocuppaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual*. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences. Retrieved from <http://www.columbia.edu/~rsb2162/bromp.html>
- Ocuppaugh, J., Baker, R. S., Rodrigo, M. M. T., Salvi, A., van Velsen, M., Aghababayan, A., & Martin, T. (2015). HART: the human affect recording tool. ACM Special Interest Group on the Design of Communication (SIGDOC). University of Limerick, Ireland. Retrieved from <http://www.columbia.edu/~rsb2162/SigDoc2015.pdf>
- Paquette, L., Baker, R. S. J. D., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., & Kauffman-Rogoff, Z. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. In *Intelligent Tutoring Systems* (pp. 1-10). Springer International Publishing.
- Pardos, Z., Baker, R., San Pedro, M., Gowda, S., & Gowda, S. (2013). Affective states and state tests: investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (pp. 117–124). Retrieved from <http://www.columbia.edu/~mzs2106/research/LAK2013.pdf>
- Rebolledo Mendez, G., du Boulay, B., & Luckin, R. (2005). “Be bold and take a challenge”: Could motivational strategies improve help-seeking? In *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (pp. 459-466).
- Reyes, L., & Fennema, E. (1981). *Classroom processes observer manual*. U.S. Department of Education, National Center for Educational Statistics. Buffalo, NY. ERIC Report ED224793. Retrieved from <http://files.eric.ed.gov/fulltext/ED224793.pdf>

- Romberg, T. A., Small, M., Carnahan, R., & Cookson, C. (1979). *Observer's manual, coordinated study# 1, 1978-1980*. Madison: Wisconsin Research and Development Center for Individualized Schooling.
- Sabourin, J., Mott, B., & Lester, J. C. (2011). Modeling learner affect with theoretically grounded dynamic Bayesian networks. In *Affective computing and intelligent interaction* (pp. 286-295). Springer Berlin Heidelberg.
- San Pedro, M., Baker, R., Bowers, A., & Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 177–184). Retrieved from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_26.pdf
- Shapiro, E. S. (1996). *Academic skills problems workbook*. New York: Guilford.
- Shapiro, E. S. (2010). *Academic skills problems fourth edition workbook*. New York: Guilford Press.
- Simon, A., & Boyer, E. G. (Eds.). (1970). *Mirrors for behavior, an anthology of classroom observation instruments, 1970 supplement, Vols. A and B*. Philadelphia, PA: Research for Better Schools, Inc.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.
- Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: a review of seven coding schemes. *School Psychology Review, 34*, 454-473.
- Whitehill, J., Serpell, Z., Foster, A., Lin, Y. C., Pearson, B., Bartlett, M., & Movellan, J. (2011, June). Towards an optimal affect-sensitive instructional system of cognitive skills. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference* (pp. 20-25).

RR RR RR RR RR RR RR

Appendix A: Interviewees

We are grateful to the following individuals who agreed to be interviewed to provide information for our cost analyses:

Ryan Baker	Associate Professor	Teachers College, Columbia University
Sidney D'Mello	Assistant Professor	University of Notre Dame
Peter Gobel	Professor	Kyoto Sangyo University, Japan
Adam B. Goldstein	Software Engineer	MeYou Health
Sujith Gowda	Research Programmer	Metacog Inc.
John Hintze	Professor	University of Massachusetts, Amherst
Jaclyn Ocumpaugh	Postdoctoral Fellow	Teachers College, Columbia University
Luc Paquette	Postdoctoral Research Associate	Teachers College, Columbia University
George Schuessler	Director of Academic Technology	Teachers College, Columbia University
M.T.Torres	Director of Network Systems	Teachers College, Columbia University
Ermal Toto	Senior Software Engineer	Worcester Polytechnic Institute

Appendix B: Ingredients and cost tables

Notes: We do not include time spent on any relevant Institutional Review Board application and approval process. All prices are expressed in 2014 U.S. dollars. Ingredient category cost totals may differ slightly from the sum of ingredient costs in each category due to rounding.

Table B1. Ingredients and costs for Hintze & Matthews (2004) observation study

Classroom observation of math/ELA with pen and paper observation protocol. Includes a half day of training with 2 trainers and 5 trainee observers. Each of 5 observers collected data in 18 class sessions over 9 days.

Categories/ingredients	Cost	% of Total*
Personnel	\$5,960	95%
Trainer I	\$509	
Trainer II	\$611	
Training time for observers	\$420	
Researcher for analysis of observations	\$636	
Observers	\$3,783	
Facilities	\$66	1%
Office space for data analysis	\$27	
Program space for training	\$39	
Materials and equipment	\$162	3%
Computer and Excel for data analysis	\$1	
Clipboards for training	\$10	
Handheld recording device for timing intervals	\$87	
Copies of BOSS paper recording forms	\$20	
Pencils	\$14	
Stopwatch to create the interval recording	\$5	
Training video	\$21	
Video recorder, cassette, and player	\$3	
Other inputs	\$99	2%
Car mileage for trainers and observers	\$99	
Total cost	\$6,286	100%

*May not add to 100% due to rounding.

Table B2. Ingredients and costs for implementing the observation study: *Student off-task behavior and motivation in the CALL classroom* (Gobel, 2008)

The subjects were university students learning English with DynEd software. A pen and paper observation protocol was used.

Categories/ingredients	Cost	% of Total
Personnel	\$5,086	96%
Observer/researcher	\$4,513	
Trainer	\$282	
University lecturers for time on study design	\$291	
Facilities	\$150	3%
Computer Assisted Language Learning (CALL) classroom	\$120	
Office space for analysis	\$28	
Copies of observation grid	\$1	
Materials and equipment	\$67	1%
Computer and Excel for data analysis	\$1	
Classroom computers and extra monitor for control console	\$31	
Classroom management software	\$1	
DynEd English Language software	\$32	
Total cost	\$5,302	100%

Note: Costs of the classroom facilities and equipment were only counted for a small amount of training time on the basis that the costs of the classroom during the observation time were not attributable to the study but to the regular costs of classroom instruction.

Table B3. Ingredients and costs of a field observation of Reasoning Mind using the HART smartphone application

Observations were conducted over three days.

Categories/Ingredients	Cost	% of Total
Personnel	\$915	25%
Observer	\$878	
Analyst to summarize data	\$37	
Facilities	\$1	0%
Office for analyst	\$1	
Materials and equipment	\$3	0%
Android device, USB cable, data plan, battery	\$3	
Computer, Internet access, Excel	\$0	
Other inputs	\$2,690	75%
Car mileage for transportation	\$76	
Air travel, hotel and per diem for observer	\$741	
BROMP training for observer*	\$1,873	
Total cost	\$3,609	100%

*See Table B4 for a breakdown of these costs

Table B4. Ingredients and costs of a two-day training session for one observer in the use of BROMP (Baker Rodrigo Ocumpaugh Monitoring Protocol)

Assumes one trainer and two days of training.

Categories/ingredients	Cost	% of Total*
Personnel	\$1,093	58%
Trainee	\$496	
Trainer	\$585	
Manual editor	\$7	
Manual writer	\$4	
Facilities	\$2	0%
Training room	\$2	
Materials and equipment	\$26	1%
Android devices, battery replacements, USB cable	\$4	
Laptop with Excel, Internet, email, Google Drive	\$22	
Clipboard for phone and paper	\$0	
Computer for manual writing	\$0	
HART (data collection phone app)	-	
Other inputs	\$752	40%
Air travel for trainer/observer	\$396	
Hotel for trainer/observer	\$230	
Car mileage for transport	\$34	
Per diems	\$92	
Total cost	\$1,873	100%

*May not add to 100% due to rounding.

Table B5. Ingredients and costs of a field observation of ASSISTments using the HART smartphone application

Three days of observations by 2 observers

Categories/Ingredients	Cost	% of Total
Personnel	\$2,540	40%
Analyst to summarize data	\$37	
Observer I	\$1,758	
Observer II	\$745	
Facilities	\$1	0%
Office for analyst	\$1	
Materials and equipment	\$6	0%
Android devices, USB cable, battery	\$6	
Computer to analyze data, Excel, Internet, email, Google Drive	\$0	
Other inputs	\$3,778	60%
Car mileage for transportation	\$32	
Prior BROMP Training for trainer	\$3,746	
Total cost	\$6,325	100%

Table B6. Ingredients and costs of a field observation of Inq-ITS using the HART smartphone application

Categories/Ingredients	Cost	% of Total*
Personnel	\$2,227	29%
Observer I	\$1,018	
Observer II	\$585	
Observer III	\$293	
Observer IV	\$293	
Analyst to summarize collected data	\$37	
Facilities	\$1	0%
Office for analyst summarizing collected data	\$1	
Materials and equipment	\$6	0%
Android devices, USB cable, battery, HART	\$6	
Computer with Internet, email, Google Drive, Excel	\$0	
Other inputs	\$5,317	70%
Car mileage for travel to 3 schools	\$81	
BROMP training for trainer and observers	\$5,236	
Total cost	\$7,551	100%

**May not add to 100% due to rounding.*

Table B7. Ingredients and costs of assessing student affect using video analysis. Data collection costs apply to all judgment situations with judgment costs being additional. For example, the costs of the study that relied on teacher judgments of student affect were \$9,307 plus \$2,591.

Cost Categories/Ingredients	Data Collection	Self- or peer judgment	Teacher judgment	Expert judgment
Personnel	\$7,454	\$1,859	\$2,279	\$4,666
Program supervisors	\$2,445			
AutoTutor - researcher time to secure license	\$37			
AutoTutor - lawyer time to review license	\$97			
Researcher A	\$2,179	\$660		
Researcher B	\$768	\$293		
Researcher C	\$604	\$128		
Undergraduate researchers	\$1,324	\$778	\$147	
Teacher time for coding			\$2,132	
Trained coders – coding time				\$1,934
Trained coders – FACS certification				\$2,733
Facilities	\$538	\$378	\$253	\$1,014
Lab for data collection	\$473			
Office space for meetings	\$65			
Lab time for training, AutoTutor familiarization, and coding		\$378	\$253	\$1,014
Materials and equipment	\$1,315	\$5	\$3	\$633
Computer (with camera, Internet, email, Google Drive, Excel) for training, data collection, coding, analysis, AutoTutor familiarization	\$10	\$5	\$3	\$37
Camtasia Studio software	\$299			
Emotion Annotation Tool	\$1,006			
Mirror				\$7
FACS training manual (CD)				\$590
Other inputs	\$0	\$0	\$56	\$0
Car mileage			\$56	
Total cost	\$9,307	\$2,241	\$2,591	\$6,313

Table B8. Ingredients and costs of building six automated detectors of affect and engagement for ASSISTments

Note: for total costs of development of detectors add data collection costs of \$6,325 from Table B5 above.

Categories/Ingredients	Cost	% of Total
Personnel	\$67,400	90%
Research programmer	\$39,620	
Supervising programmer	\$8,114	
Programmer II	\$16,944	
Programmer III	\$2,555	
Programmers for features brainstorming session	\$168	
Facilities	\$6,520	9%
Lab space for programmers	\$6,287	
Office space for supervising programmer	\$233	
Materials and equipment	\$695	1%
Computer, internet access, and Excel for all programmers and supervisor	\$586	
Refreshments for brainstorming session	\$109	
Other inputs	\$5	0%
Car mileage to 1 school for synchronization	\$5	
Total cost	\$74,621	100%

Table B9. Ingredients and costs of building four automated detectors of affect for Inq-ITS

Note: for total costs of development of detectors add data collection costs of \$7,551 from Table B6 above.

Categories/Ingredients	Cost	% of Total*
Personnel	\$41,325	98%
Supervisor programmer	\$3,065	
PhD students for feature brainstorming	\$252	
MA students for feature brainstorming	\$63	
Research software engineer for correlator	\$2,674	
Programmer for detector building	\$33,993	
Programmer II for synchronization of Inq-ITS log files to HART	\$875	
Research associate supervising detector building work	\$402	
Facilities	\$723	2%
Office space for programmers and supervisor	\$715	
Space for feature brainstorming event	\$9	
Materials and equipment	\$246	1%
Computers with Excel, Internet, email, Google Drive	\$164	
Refreshments for feature brainstorming event	\$83	
Other inputs	\$5	0%
Car mileage to 1 school for synchronization	\$5	
Total cost	\$42,300	100%

*May not add to 100% due to rounding.

XX XX XX XX XX XX XX