Cost-Effective, Actionable Engagement Detection at Scale

Ryan S. Baker, Jaclyn Ocumpaugh
Teachers College, Columbia University
525 W. 120th St., Box 118
New York, NY 10027 USA
1-212-678-8329 and 1-212-678-3854
baker2@exchange.tc.columbia.edu, jo2424@columbia.edu

ABSTRACT
Costs of educational measures and interventions have important real-world implications, made more pertinent when used at scale. Traditional measures of engagement (e.g., video and field observations) scale linearly, so that expanding from 10 classrooms to 100 can incur 10 times the cost. By contrast, the cost of applying an automated sensor-free detector of student engagement is independent of the size of the data set. While the development and validation of such detectors requires an initial investment, once this cost is amortized across large data sets, the cost per student/hour is quite modest. In addition, these detectors can be reused each year at minimal additional cost. In this paper, we provide a formal cost analysis of automated detectors of engagement for ASSISTments.

Keywords
Affective computing, sensor-free detection, ASSISTments, STEM, student engagement, cost-effectiveness

1. INTRODUCTION
Automated sensor-free detectors of student engagement are now available for several systems [3, 6], shifting the debate from whether this detection is possible to a discussion of the upper limits of its performance and generalizability (see [5]). Automated detectors have been used to drive interventions [1] and in discovery with models analyses [6, 8]. As researchers and policy-makers seek to identify the teaching methods and online learning systems that promote greater engagement, studies at considerable scale have become a priority [2]. Unfortunately, this scale is often achieved at considerable cost. An alternate option is to use EDM models on log files. With appropriately validated detectors, engagement among large numbers of students can be gauged rapidly, and as students continue using the same learning system, extensive, individualized data can be applied to interventions and to long-term predictions through discovery with models.

In this paper, we examine the cost of developing detectors of 7 constructs of student engagement for ASSISTments, outlining current expenses and applications. We include a brief description of their applicability towards discovery with models research, a technique that leverages such existing models, substantially increasing their worth.

2. ASSISTments
Detectors in this study were developed for ASSISTments [7], a formative assessment system that provides scaffolded math instruction and targeted hints. ASSISTments was developed at Worcester Polytechnic Institute and is available to educators at no cost. Typically, students spend 1 regular class period per week using ASSISTments, and some also use it for homework [3]. Currently, approximately 60,000 students use ASSISTments in schools throughout the Northeastern United States.

3. Methods
3.1 Overview of Detector Construction
For this study, we consider the cost of producing detectors for 7 different constructs, including 4 affective indicators of engagement (boredom, confusion, engaged concentration, and frustration) and 3 behavioral indicators of engagement (gaming the system, off-task behavior, and carelessness). As reported in previous research, different methods were used to obtain the ground truth labels used to develop these detectors.

For the 4 affective detectors and for 2 of the behavioral detectors (gaming the system and off-task behavior), ground truth labels were generated by BROMP-certified field observers [4]. Observers spent 379 hours in field, obtaining 5,564 observations of 590 students at 6 different schools. These observations were then used to train separate detectors for each construct, each of which was cross-validated at the student-level to ensure generalizability to new populations (e.g., [6]). Research shows that rural students' affect manifested differently in their interactions with ASSISTments compared to urban and suburban students, so affect detectors were constructed to reflect these demographic differences [5].

The construction of the carelessness detector was different than the other six detectors as no fieldwork was required. Instead, programmers used Bayesian Knowledge Tracing (BKT) algorithms to calculate Contextual Slip (e.g., [1]). Each time a student makes an incorrect answer on a problem, the probability that a student is making a careless error is calculated based on his or her previous performance on the same skill [8].

3.2 Calculation of Cost
Two major sources of funding were used to create the ASSISTments detectors. The first was a National Science Foundation award to the Pittsburgh Science of Learning Center for $100,000, used to develop initial detectors for ASSISTments and 4 other systems. The second, a grant from the Bill & Melinda Gates Foundation for $277,044 funded further enhancement and validation of the ASSISTments detectors and models for 2 other systems. Roughly, this means that the initial investment for cross-

validated models of all seven constructs in ASSISTments (also tested across 3 populations) totaled $117,348 ($16,050 per construct or $7,823 per detector). A list of these detectors, their algorithms, and their performance metrics are provided in Table 1.

Table 1. Table captions should be placed above the table

Detector	Algorithm	Kappa	A'	r
Boredom (Urban)	Jrip	0.23	0.6	*na*
Boredom (Rural)	K*	0.24	0.7	*na*
Boredom (Suburban)	REPTree	0.19	0.7	*na*
Confusion (Urban)	J48	0.27	0.7	*na*
Confusion (Rural)	JRip	0.14	0.6	*na*
Confusion (Suburban)	REPTree	0.38	0.7	*na*
Engaged Concentration (Urban)	K*	0.36	0.7	*na*
Engaged Concentration (Rural)	REPTree	0.37	0.7	*na*
Engaged Concentration	J48	0.27	0.6	*na*
Frustration (Urban)	REPTree	0.29	0.7	*na*
Frustration (Rural)	JRip	0.2	0.6	*na*
Frustration (Suburban)	REPTree	0.17	0.6	*na*
Gaming the System	K*	0.37	0.8	*na*
Off-Task Behavior	REP-Tree	0.51	0.8	*na*
Carelessness	Linear	*na*	*na*	0.50

However, since opportunities to apply both interventions and discovery with analyses are predicated on the number of labels (not just the number of detectors), the cost per label is perhaps a better indicator of the value of this research. At present, these detectors have been applied to 231,543 hours of ASSISTments data produced by 54,401 students. For BROMP-trained detectors, a label has been applied to every 20 seconds of interaction within the system (41,677,740 intervals x 6 constructs = 250,066,440 labels), and carelessness labels have been applied to every problem incorrectly completed during that time (3,163,616 labels). This puts the current cost per label well under 1 penny ($0.00046/label), a price that will continue to drop over the years as these detectors are applied to new data.

The cumulative cost per student/hour, a calculation important to allocating financial resources in education, is also extremely low ($1.97) and becomes even lower when calculated as a cost per construct ($1.97/7 = $0.28). By comparison, even if we (incorrectly) assumed a single observer, paid the 2014 federal minimum wage of $7.25/hour, could replicate the granularity of this data, the cost would top $1.6 million. In reality, the minimum rate of a trained observer is likely closer to $25/hour (plus approximately 37% in benefits), totaling almost $8,000,000, and a 1-1 coder-student ratio would be needed. Such conditions would likely destroy the value of any data collected since classroom conditions would be so disrupted as to make any data meaningless, and using video coding to attempt to replicate this level of granularity would incur even further expenses.

4. DISCUSSION/IMPLICATIONS

Cost is not the sole criteria for evaluating educational research, but it is a necessary consideration when developing resources to be implemented at scale. In this report, we have discussed the costs involved in developing detectors for 7 measures of student engagement, demonstrating that their relative cost is quite low, particularly when amortized across the use of the detectors to label data. These costs will drop further still in the coming years.

Further research with these detectors demonstrates the long-term predictive prognostic power of these constructs, which can predict standardized test scores [6] and college attendance several years later [8], showing the importance of this granular data. Currently, we are working to make these predictions more accessible to educators, providing them with actionable reports about who most needs intervention and which student behaviors are most problematic. As such, these models are likely to be of value both for research and for practice, and as these detectors scale easily, interventions based on them will also.

5. ACKNOWLEDGMENTS

Thanks to the Bill & Melinda Gates Foundation and the National Science Foundation (#SBE-0836012) for support, and thanks to Fiona Hollands for useful suggestions. (All errors are ours.)

6. REFERENCES

[1] Cooper, D. G., Arroyo, I., & Woolf, B. P. (2011). Actionable affective processing for automatic tutor interventions. In *New Perspectives on Affect and Learning Technologies* (pp. 127-140). Springer New York.

[2] Corbett, A.T., and Anderson, J.R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4, 4, 253-278.

[3] D'Mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, *18*(1-2), 45-80.

[4] Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. *Bill & Melinda Gates Foundation*.

[5] Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. *J. Research on Comp. in Education*, *41*, 3, 331.

[6] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*.

[7] Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. (2012) *Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual*. Technical Report. New York, NY: EdLab.

[8] Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2013) Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proc. of the 3rd International Conference on Learning Analytics and Knowledge*, 117-124.

[9] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A. and Rasmussen, K.P. (2005). The Assistment project: Blending assessment and assisting. In *Proc. AIED 2005*, 555-562.

[10] San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (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*, 177-184.