The Dynamics of Affect in Online Learning

Ryan Baker
University of Pennsylvania
@BakerEDMLab

Thank you

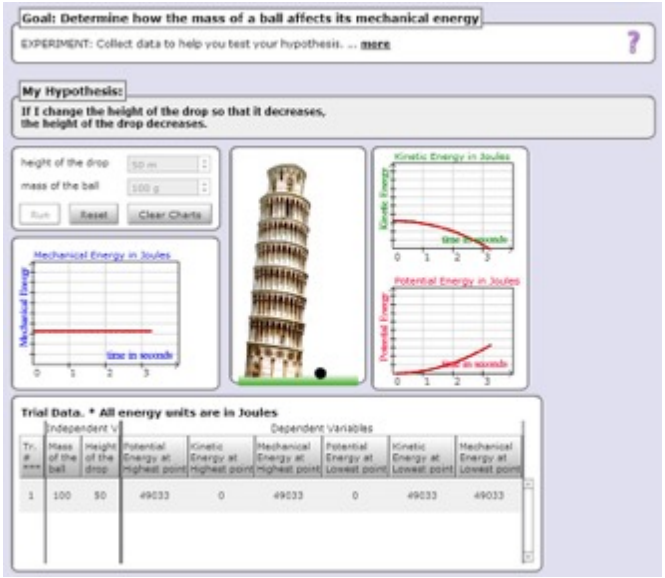
- For welcoming me here today

Please interrupt
whenever you would like

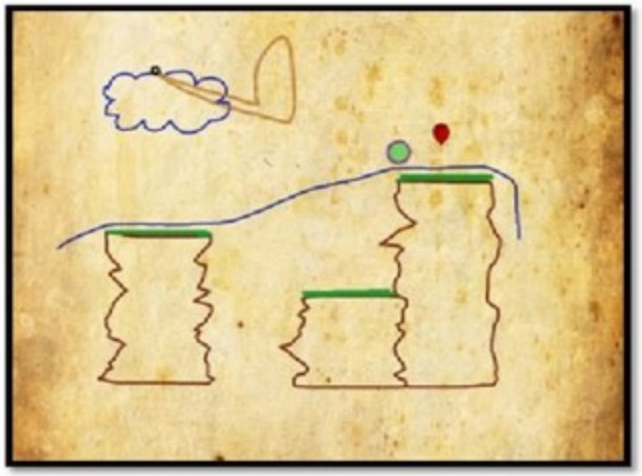
In recent years,

- More and more learning occurs in interactive online environments and MOOCs

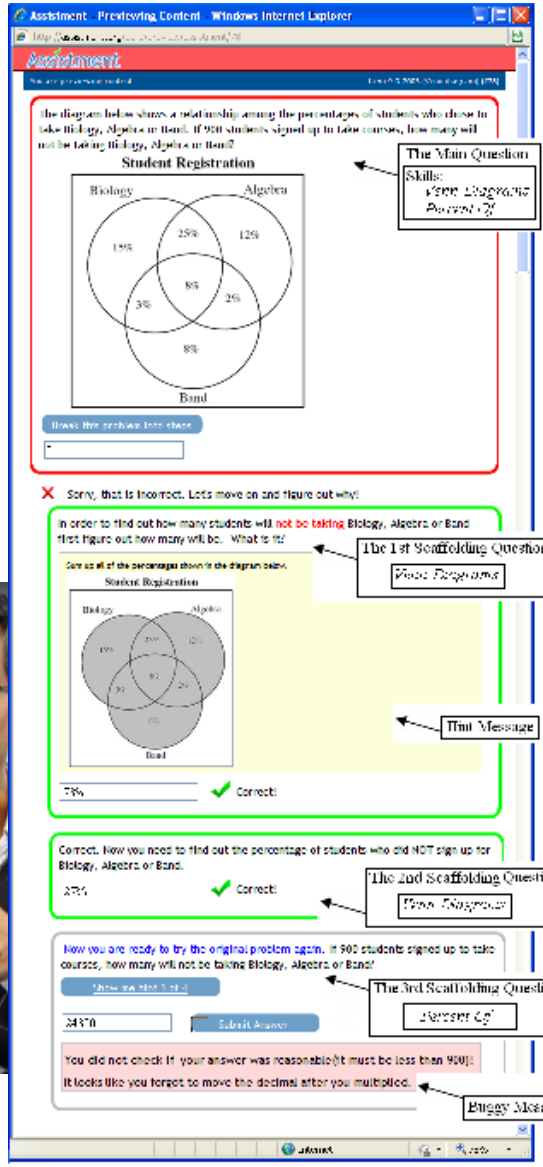
- Millions of learners a year


Goal: Determine how the mass of a ball affects its mechanical energy
EXPERIMENT: Collect data to help you test your hypothesis. ... more
My Hypothesis:
If I change the height of the drop so that it decreases, the height of the drop decreases.
height of the drop
mass of the ball
Run
Reset
Clear Charts
Kinetic Energy in Joules
Potential Energy in Joules
Mechanical Energy in Joules
time in seconds
Trial Data. * All energy units are in Joules
Independent Variables
Dependent Variables
Tr. #
Mass of the ball
Height of the drop
Potential Energy at Highest point
Kinetic Energy at Highest point
Mechanical Energy at Highest point
Potential Energy at Lowest point
Kinetic Energy at Lowest point
Mechanical Energy at Lowest point
1 100 50 49033 0 49033 0 49033 49033


Settings
Quit

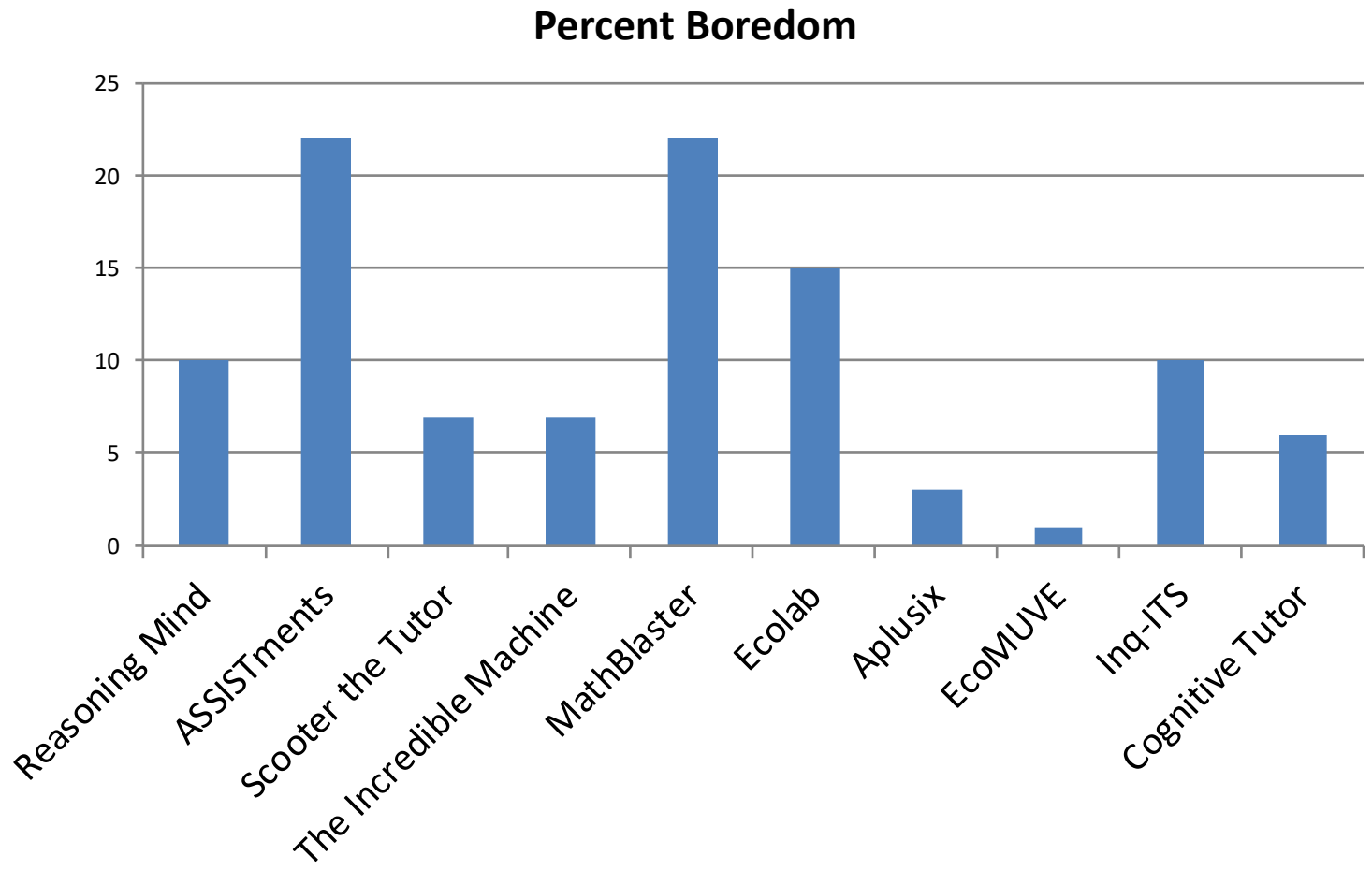



Additive Manufacturing for Innovative Design and Production
KARTIK MANGUDI VARADARAJAN
Assistant Professor of Orthopaedic Surgery
MORE VIDEOS


Assessment - Previewing Content - Windows Internet Explorer
Assessment
The diagram below shows a relationship among the percentages of students who chose to take Biology, Algebra or Band. If 900 students signed up to take courses, how many will not be taking Biology, Algebra or Band?
Student Registration
Biology
Algebra
Band
15%
25%
12%
8%
3%
2%
8%
The Main Question
Skills:
Parent Q
Sorry, that is incorrect. Let's move on and figure out why!
In order to find out how many students will not be taking Biology, Algebra or Band first figure out how many will be. What is it?
The 1st Scaffolding Question
Hint Message
Correct!
Correct. Now you need to find out the percentage of students who did NOT sign up for Biology, Algebra or Band.
Correct!
The 2nd Scaffolding Question
Now you are ready to try the original problem again. If 900 students signed up to take courses, how many will not be taking Biology, Algebra or Band?
The 3rd Scaffolding Question
Submit Answer
You did not check if your answer was reasonable. It must be less than 900.
It looks like you forgot to move the decimal after you multiplied.
Buggy Message

And some systems and courses can be very engaging

But there is considerable variation in engagingness



Important Because…

- Affect and engagement in online learning predict student outcomes, even several years later (e.g. San Pedro et al., 2013, 2015)

Our group has developed measures…

- That are
  - **Automated**: Able to make assessments about students in real-time, with no human in the loop

Our group has developed measures…

- That are
  - **Automated**: Able to make assessments about students in real-time, with no human in the loop
  - **Fine-grained**: Able to make assessments about students second-by-second

Our group has developed measures…

- That are
  - **Automated**: Able to make assessments about students in real-time, with no human in the loop
  - **Fine-grained**: Able to make assessments about students second-by-second
  - **Validated**: Demonstrated to apply to new students and new contexts

Detectors Built For

- ASSISTments
- Science ASSISTments/InqITS
- EcoMUVE
- SQL-Tutor
- Aplusix
- BlueJ
- Cognitive Tutors for Math, Genetics
- Reasoning Mind
- vMedic
- Newton’s Playground
- Betty’s Brain

Opportunities: Improvements to Practice

- Can we develop systems that recognize when a student is becoming disengaged, and adapt to improve engagement?
- Can we assess which materials are less engaging, to drive re-design?
- Can we determine which students are less engaged, to provide predictive analytics?

Opportunities: Basic research

- What are the dynamics of student affect and engagement over time?
- What is the duration of different affective states?
- Which affective states and forms of engagement matter in different contexts?
- Which affective states and forms of engagement matter for the long-term?

Basic research influences practice!

- What are the dynamics of student affect and engagement over time?
  - Which shifts should we expect? Which shifts do we have a greater chance to influence?
- What is the duration of different affective states?
  - Which affective states form “vicious cycles” which are hard to disrupt?
- Which affective states and forms of engagement matter in different contexts?
  - Drives design in a specific environment
- Which affective states and forms of engagement matter for the long-term?
  - Focus on what matters for the long-term

How they work

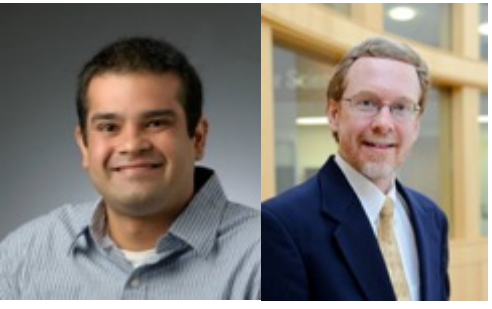
- Detect engagement and affect solely from student interactions with software

- Sensors raise privacy, political, cost, and equity concerns that we'd prefer to sidestep

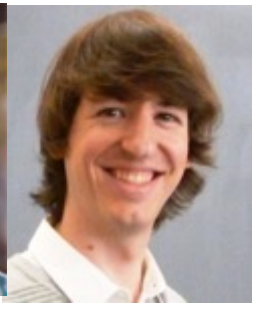
(But see)

- Our work to integrate interaction-based and sensor-based detectors
  (Bosch et al., 2015a, 2015b, 2016a, 2016b; D'Mello et al., 2016; Kai et al., 2015; Paquette et al., 2015)

- With D'Mello's group and Lester's group







Brief Summary of that work

- Interaction-based detectors either better (Paquette et al., 2015) or not as good but have additive value (Bosch et al., 2016)

- Interaction-based detectors usable in many situations when video-based detectors ineffective (Bosch et al., 2015)

Primary Constructs we Model

Off-Task Behavior

- When a student completely disengages from the learning environment and task to engage in an unrelated behavior

Gaming the System

- Intentionally misusing educational software to complete problems and advance without learning (Baker et al., 2004)

- Systematic guessing
- Hint abuse

Careless Errors

- Making errors despite knowing the relevant skills or concepts

- When 6*9 equals 42

Affect

- Engaged Concentration
  - positive focused concentration towards the task
  - related to *flow* (Csikszentmihalyi, 1990)
- Boredom
- Frustration
- Confusion

Method

1. Get human assessments of engagement and disengagement, synchronized to log files from educational software
2. Use data mining to develop models that can replicate those human judgments, using just log files

Building automated detectors: Our classic approach

- Synchronize log data to field observations
- Distill meaningful data features for learning environment
  - based on qualitative study of log files, experiences of field observers, and past experience with other data sets
- Develop automated detector using classification algorithms
- Validate detector for new students/new lessons/new populations

Classical machine learning or deep learning?

- Most of our work has involved classical machine learning algorithms (Baker et al., 2008, 2010, 2013; Paquette et al., 2014; Pardos et al., 2014; DeFalco et al., 2018; Jiang et al., 2018)
  - Decision Trees (J48)
  - Decision Rules (JRip)
  - Functional Classification (Step Regression)
  - Instance-Based Classification (K*)

Classical machine learning or deep learning?

- Some of our recent work has attempted to use "deep learning" (recurrent neural networks) (Botelho et al., 2017; Bosch et al., 2018)
  - Initial appearance of much better performance in one system; unstable across student populations
  - About the same as classical machine learning in the other case

Use of detectors

- What are the dynamics of student affect and engagement over time?
- What is the duration of different affective states?
- Which affective states and forms of engagement matter in different contexts?
- Which affective states and forms of engagement matter for the long-term?

Previous work

- Lots of research into which affective states precede and follow each other over time
- Started with (D'Mello et al., 2007; Baker et al., 2007)
- Dozens of publications since then
- This work has mostly involved sequences of field observations or self-reports
  - Limited amounts of data
  - Relatively long gaps between two observations of same student

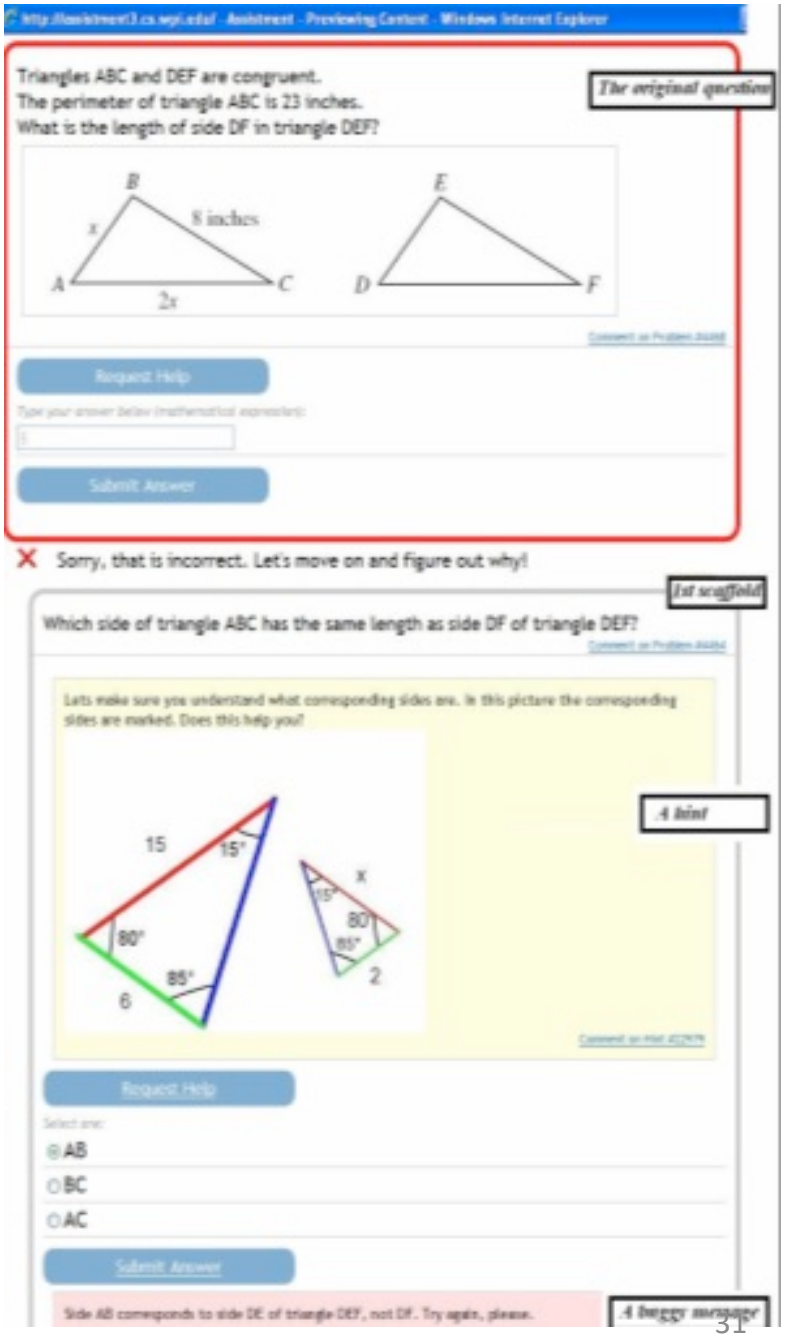
Recent work

- (Botelho, Baker, Ocumpaugh, & Heffernan, 2018) applied automated detectors to larger data set
  - Context: ASSISTments platform

ASSISTments

- Web-based mathematics tutor
- Primarily for middle school math

- Gives student mathematics questions
- Offers multi-step hints to struggling student
- If student makes error, student is given scaffolding that breaks the original questions down into sub-steps



Over 50,000 kids a year

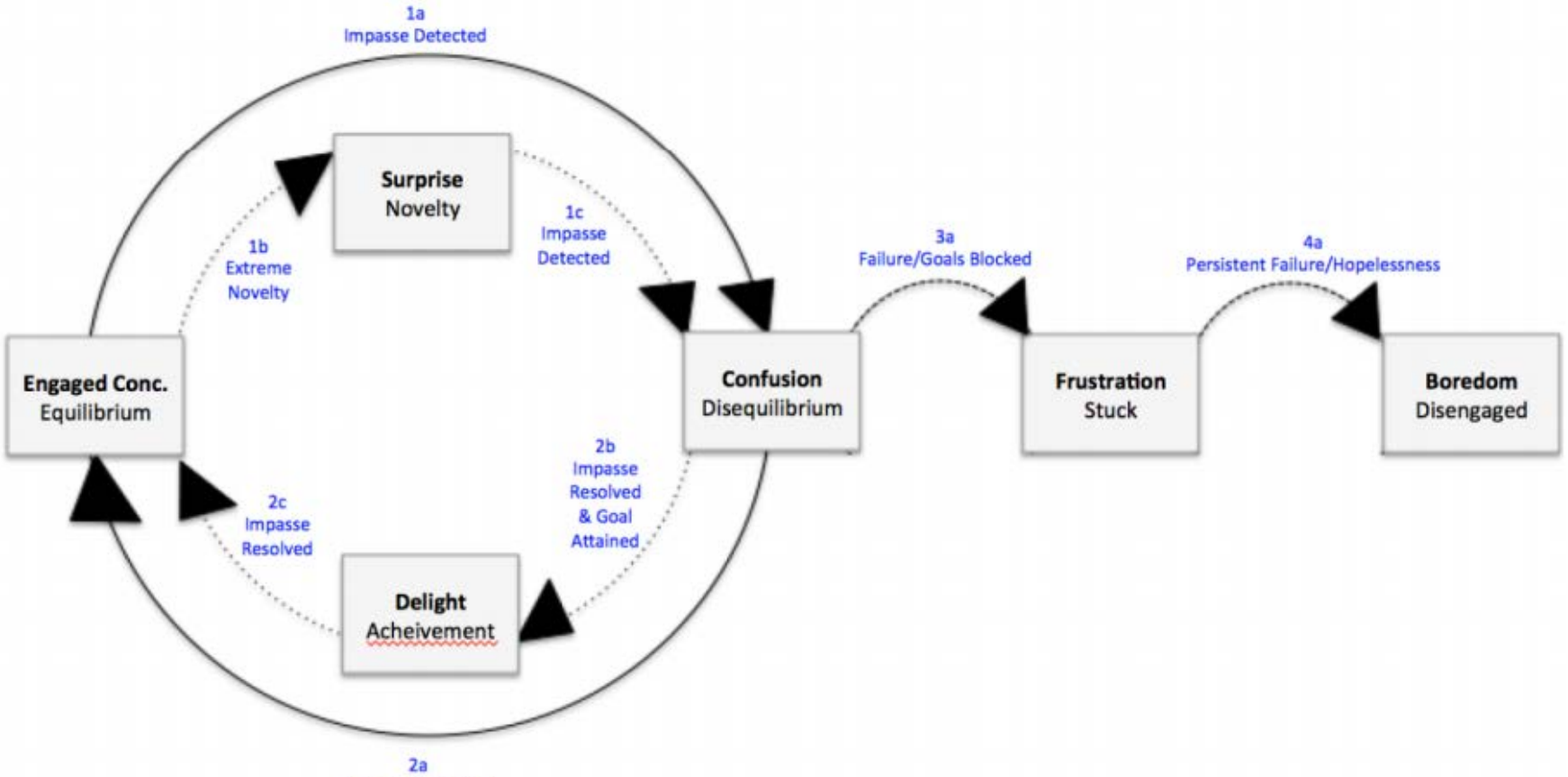
Data and analysis

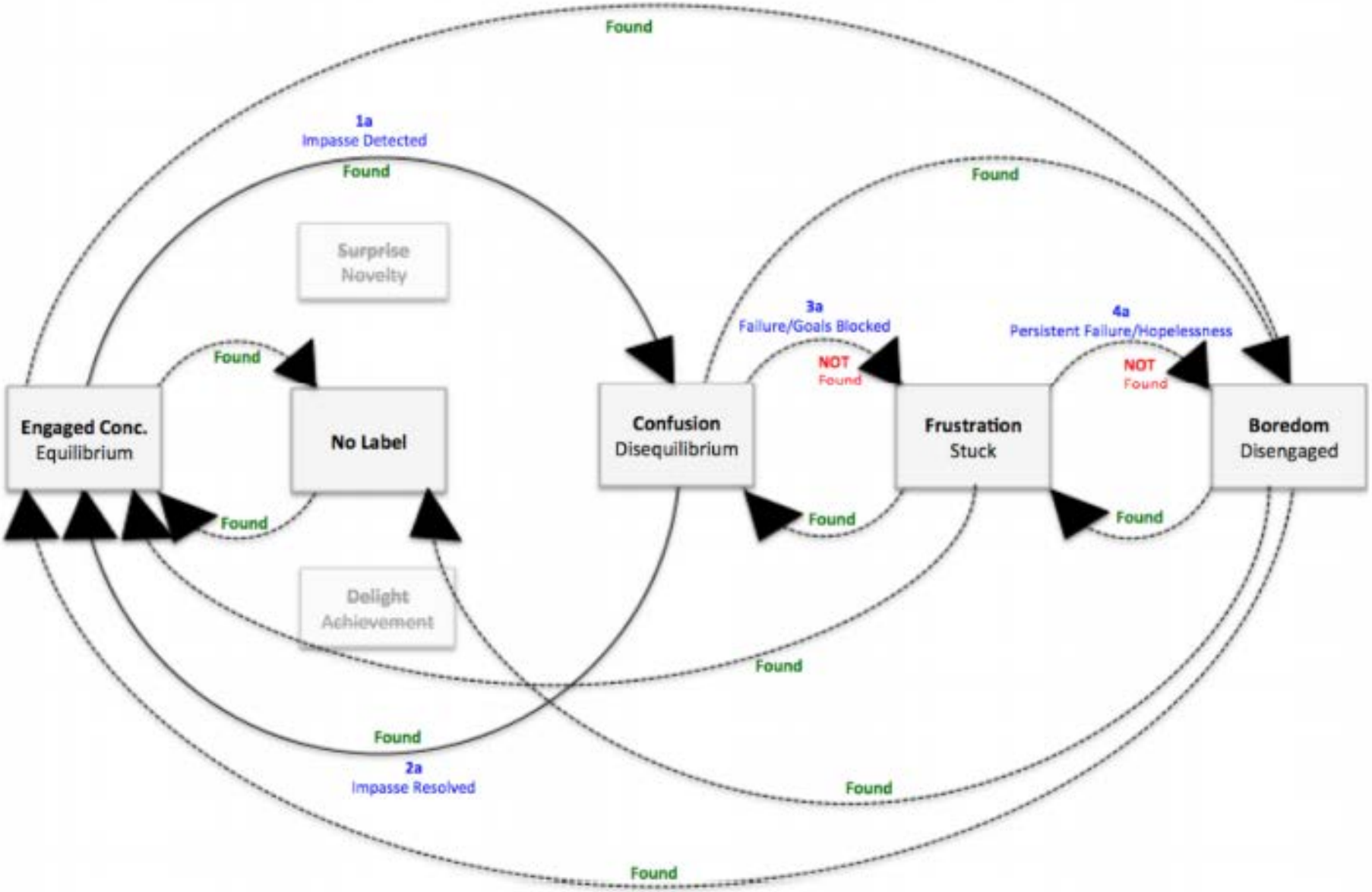
- 48,276 20-second segments of affect by 838 students
- Looking to see whether a transition from affective state P to affective state N occurs statistically significantly more often than would be suggested by affective state N's base rate
- D'Mello's (2007) ***L***

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

Data and analysis

- Compare findings to D'Mello & Graesser's (2012) theoretical model of affective dynamics




Found
1a
Impasse Detected
Found
Found
Surprise
Novelty
3a
Failure/Goals Blocked
4a
Persistent Failure/Hopelessness
Found
NOT
Found
NOT
Found
Engaged Conc.
Equilibrium
No Label
Confusion
Disequilibrium
Frustration
Stuck
Boredom
Disengaged
Found
Found
Found
Delight
Achievement
Found
Found
2a
Impasse Resolved
Found
Found

Use of detectors

- What are the dynamics of student affect and engagement over time?
- What is the duration of different affective states?
- Which affective states and forms of engagement matter in different contexts?
- Which affective states and forms of engagement matter for the long-term?

Previous work

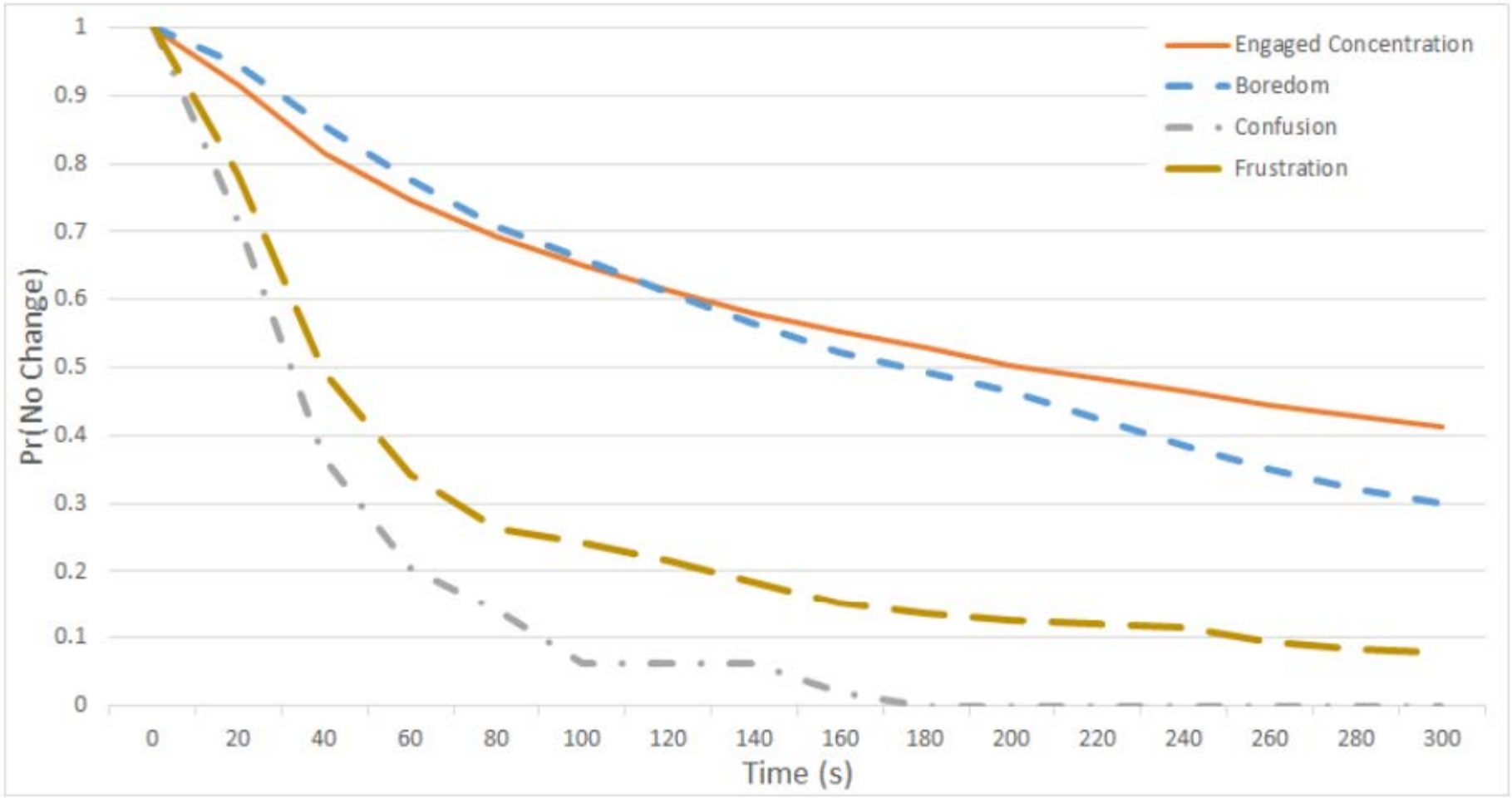
- Relatively limited

- One lab study over short durations by D'Mello & Graesser (2011)

Recent work

- (Botelho, Baker, Ocumpaugh, & Heffernan, 2018) analyzed duration of affect on same larger ASSISTments data set

- Affect much more persistent in classroom setting than in earlier lab study


Engaged Concentration
Boredom
Confusion
Frustration
Pr(No Change)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Time (s)

Use of detectors

- What are the dynamics of student affect and engagement over time?
- What is the duration of different affective states?
- Which affective states and forms of engagement matter in different contexts?
- Which affective states and forms of engagement matter for the long-term?

High consistency for behavioral disengagement

- Gaming the system associated with negative learning outcomes in several studies (Baker et al., 2004; Cocea et al., 2009; Pardos et al., 2014; Fancsali, 2015)
- Carelessness associated with negative learning outcomes in several studies (San Pedro et al., 2013; Pardos et al., 2014 ; Fancsali, 2015)
- Off-task not particularly associated with negative learning outcomes in online learning (Baker et al., 2004; Cocea et al., 2009; Pardos et al., 2014; Fancsali, 2015) with one notable exception (Kostyuk et al., 2018)

A lot of variation in affect

- College student lab studies (Craig et al., 2004)
  - Boredom negatively associated with outcomes
  - Engaged concentration and confusion positively associated with outcomes
- College programming (Rodrigo et al., 2009)
  - Boredom and confusion negatively associated with outcomes
  - Engaged concentration positively associated with outcomes
- Middle school math (Pardos et al., 2014)
  - Boredom negatively associated with outcomes
  - Engaged concentration positively associated with outcomes
- Stats MOOC (Dillon et al., 2016)
  - Confusion, frustration, anxiety, and hope (???) negatively associated with outcomes
- Military cadets (DeFalco et al., 2018)
  - Frustration negatively associated with outcomes
  - No correlation for boredom

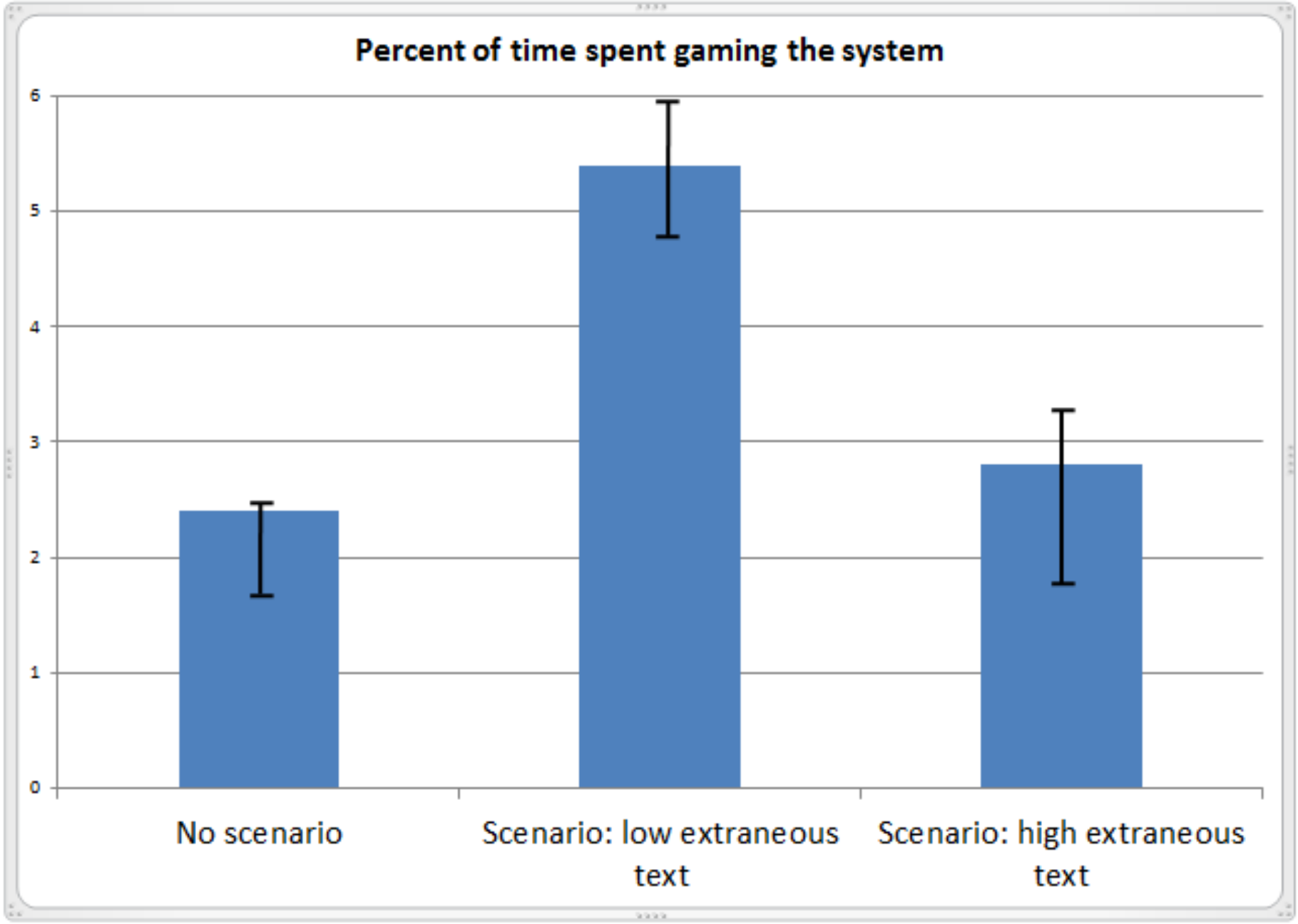
What about confusion?

- Liu et al. (2011) found that brief confusion associated with positive learning outcomes and extended confusion associated with negative learning outcomes

Design of curricular materials

- How does design of curricular materials impacts disengagement and affect? (Baker et al., 2009; Doddanarra et al., 2013)
  - Very concrete problems good for affect & engagement
  - Very abstract problems good for affect & engagement
  - In between not so good

  - Context: Cognitive Tutor/MATHia


Percent of time spent gaming the system
6
5
4
3
2
1
0
No scenario
Scenario: low extraneous text
Scenario: high extraneous text

Other Features

- Ineffective hints -> More gaming
- Abstract hints -> More gaming
- Unclear UI -> More gaming

Use of detectors

- What are the dynamics of student affect and engagement over time?
- What is the duration of different affective states?
- Which affective states and forms of engagement matter in different contexts?
- Which affective states and forms of engagement matter for the long-term?

Engagement and Standardized Exam Score (Pardos et al., 2013, 2014)

- Detectors applied to whole year of data for 1,393 students who used ASSISTments

- Gaming the system (r = -0.36)
- Engaged concentration (r = +0.36)
- Boredom (r = -0.2)

- First two similar magnitude to correlation between cigarette smoking and lifespan

College Attendance

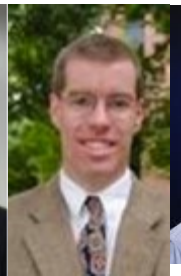
(San Pedro, Baker, Bowers, & Heffernan, 2013)

- Apply detectors to data from 2004-2007
- The detectors can predict
- Whether a student will go to college or not, ~6 years later
  - 69% of the time for new students

Predict College Attendance (San Pedro et al., 2013)

- Student knowledge, engaged concentration, carelessness associated with going to college
- Gaming the system, boredom, confusion associated with not going to college

- Overall model A’ = 0.69


Note

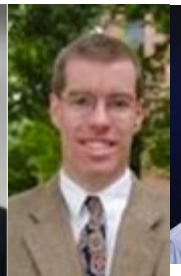
- Carelessness positively associated with college until you control for student knowledge
- Then associated with not going to college

- Carelessness is the disengaged behavior of generally successful students (cf. Clements, 1982)

Predict Selective College Attendance (San Pedro et al., 2013)

- Student knowledge, engaged concentration, carelessness associated with going to selective college
- Gaming the system, boredom associated with not going to selective college

- Overall model A’ = 0.76


Predict STEM Major in college (San Pedro et al., 2014)

- Student knowledge, carelessness associated with STEM major
- Gaming the system associated with non-STEM major (D= 0.573)

- Overall model A’ = 0.68

Another Example

- Student engagement within a MOOC on data science can predict whether the student will eventually submit a scientific paper in the field (Wang et al., 2017)

- Forum lurkers are more likely to submit a scientific paper than forum posters!
  - Even though forum posters are more likely to complete the course

Summary

How do we use this information?

- Advance the Science of Learning
- Empower Teachers and Guidance Counselors
- Automated Intervention/Individualization

Advancing the Science of Learning

- Many scientific discoveries enabled by these methods

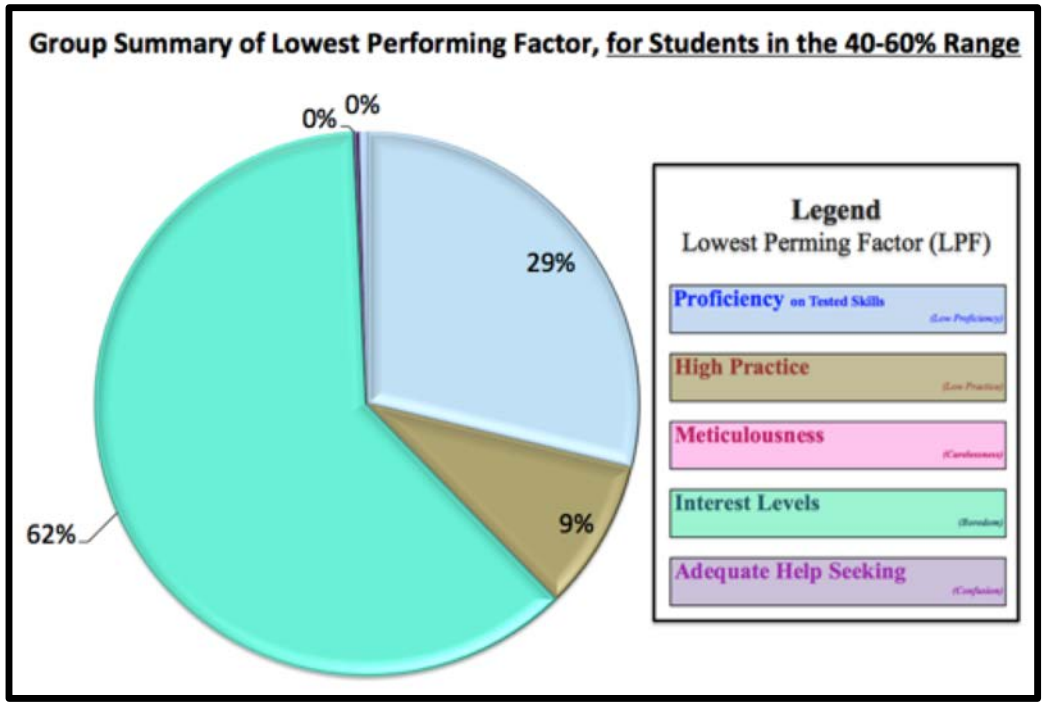
- You've seen a sample from my research group today

Empower Instructors, Guidance Counselors, Course Designers

- With data on long-term student trajectories
  - Along with each student's risk factors

Guidance Counselor Reports (Ocumpaugh et al., 2017)

	College Forecast	Lowest Performing Category (excludes Sincere Effort, which was not used to calculate CF)	Proficiency on Tested Skills (Low Proficiency)	High Practice (Low Practice)	Meticulousness (Carelessness)	Interest Levels (Boredom)	Comprehension (Confusion)	Sincere Effort* (Gaming the System) *not used to predict CF
	CF	LPC	P	HP	M	I	C	SE
Alice Bly	80-100%	I	+	+	+	+	+	-
Arthur McBride	60-80%	HP	+	+	+	+	+	+
Flora West	60-80%	M	+	+	-	avg.	avg.	-
Ira Hayes	80-100%	I	+	+	+	+	+	+
Jack Davey Black	60-80%	M	+	+	+	+	+	+
Lamar Houston	60-80%	C	+	+	+	+	+	+
Nettie Moore	80-100%	I	+	+	+	+	+	+
Tim Angel	60-80%	HP	+	+	+	+	+	+
Charlie Patton	40-60%	C	avg.	avg.	+	avg.	avg.	avg.
Dusty Blackcoat	40-60%	P	avg.	avg.	+	+	+	avg.
Frankie Lee	40-60%	P	-	avg.	+	avg.	avg.	-
Hollis Brown	40-60%	P	avg.	avg.	+	avg.	avg.	+
Joe Diamond	40-60%	M	+	avg.	-	+	+	+
Jim Jones	40-60%	P	-	avg.	+	avg.	+	+
Joey DelRey	40-60%	P	avg.	avg.	+	avg.	+	avg.
John Harding	40-60%	M	+	avg.	-	+	+	+
Lenny Bruce	40-60%	HP	avg.	avg.	+	avg.	avg.	avg.
Maggie Farmer	40-60%	M	+	+	-	avg.	avg.	avg.
Maria Washington	40-60%	P	-	avg.	+	avg.	avg.	avg.
Sara J. McMillan	40-60%	C	avg.	avg.	-	+	+	+
Scarlet Pueblo	40-60%	P	avg.	avg.	+	+	+	-
Willie McTell	40-60%	I	avg.	avg.	+	-	avg.	avg.
Duquesne Whistle	0-20%**	M	+	avg.	-	-	-	avg.
Hazel Love	0-20%**	M	-	-	-	-	-	-
Henry Lee	20-40%	M	+	avg.	-	-	-	-
Sally Sue Brown	20-40%	M	+	avg.	-	-	-	-


Group Summary of Lowest Performing Factor, for Students in the 40-60% Range
0%
0%
29%
9%
62%
Legend
Lowest Perming Factor (LPF)
Proficiency on Tested Skills
High Practice
Meticulousness
Interest Levels
Adequate Help Seeking

Reports to Regional Coordinators

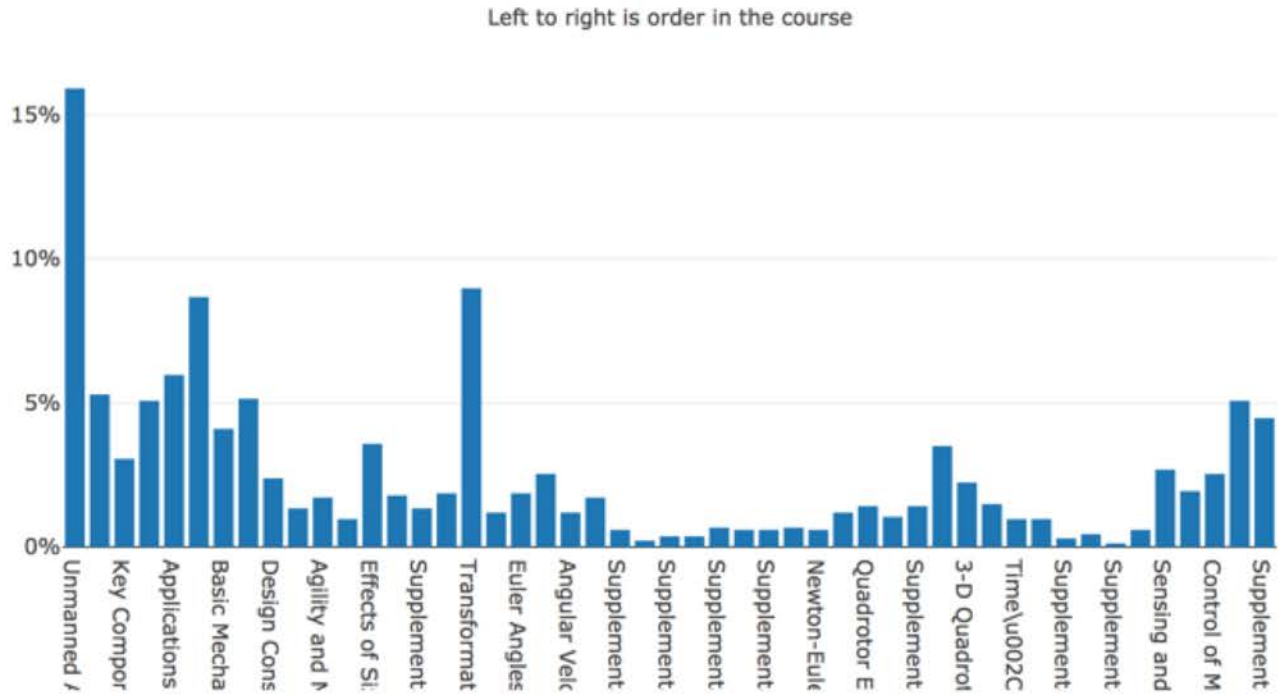
- Another online curriculum we work with, Reasoning Mind, deployed reports on student engagement to regional coordinators prior to their acquisition by another company

- Allowed them to target teachers for additional support and professional development

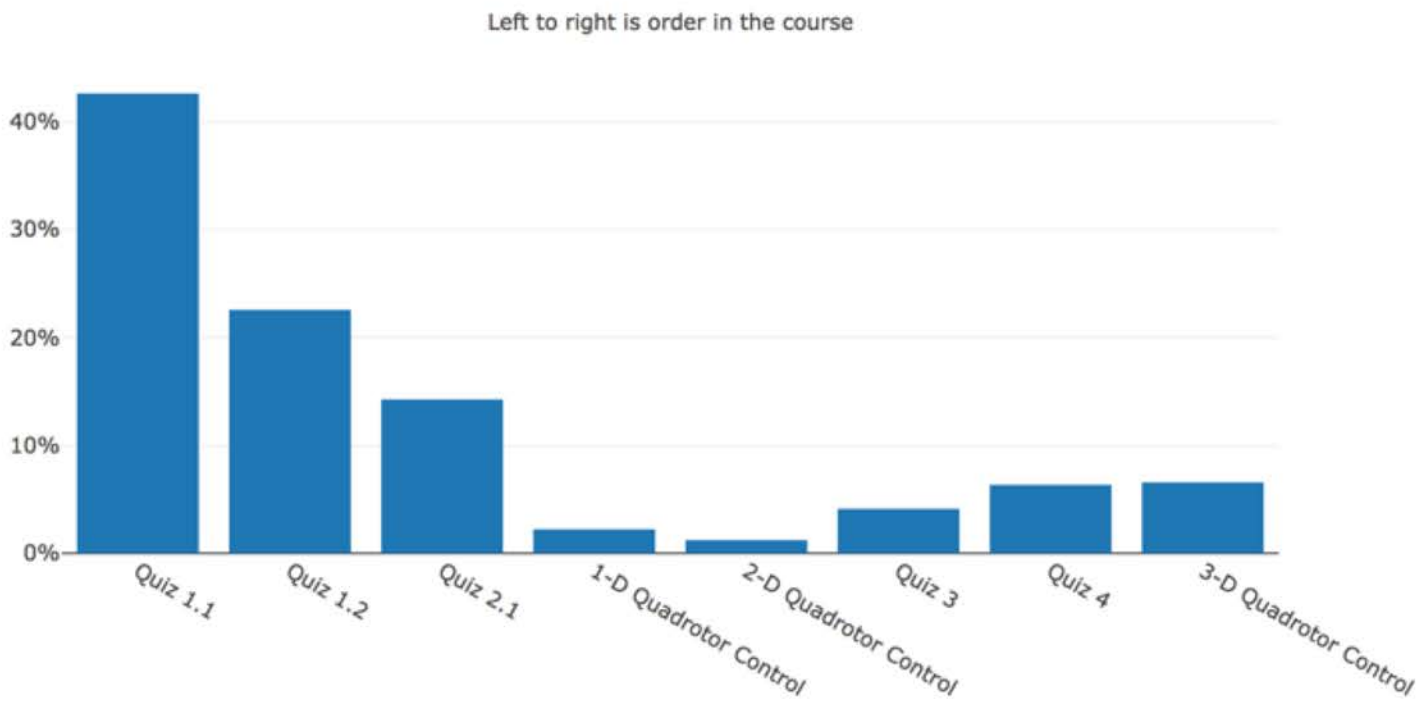
Reports on Disengagement to Instructors (UPenn OLI/PCLA)

- Study what content is associated with learner ceasing participation in a MOOC

% of ever-active users for whom video was the last video seen before dropping out of or completing the course.



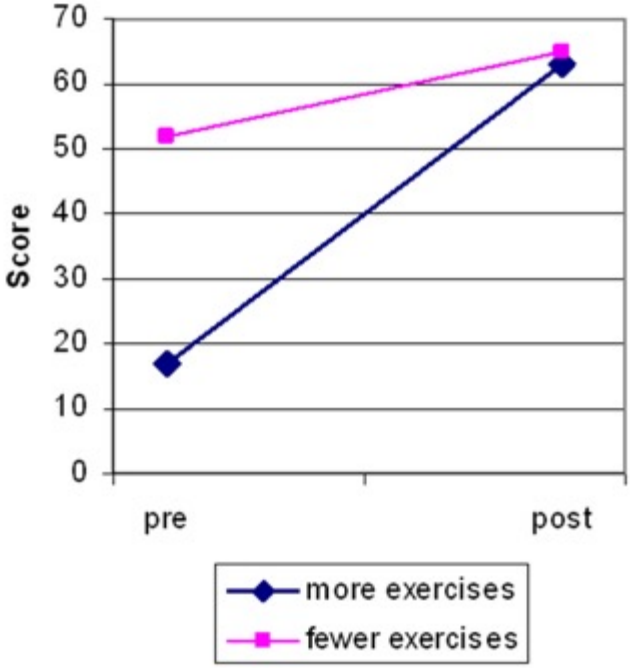
% of time each assignment was last seen before dropping out of or completing the course.

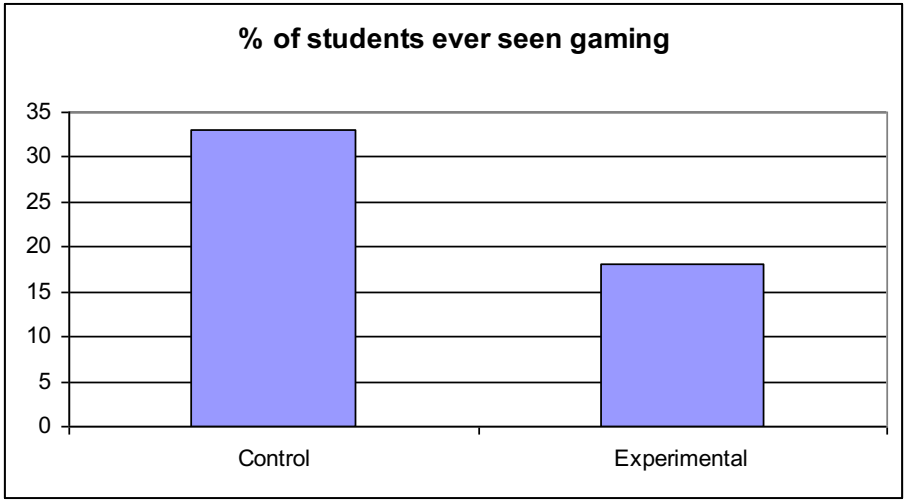


Course order	Assignment	Last seen (%)
1	Quiz 1.1	42.63
2	Quiz 1.2	22.54
3	Quiz 2.1	14.20

Automated Intervention/Individualization

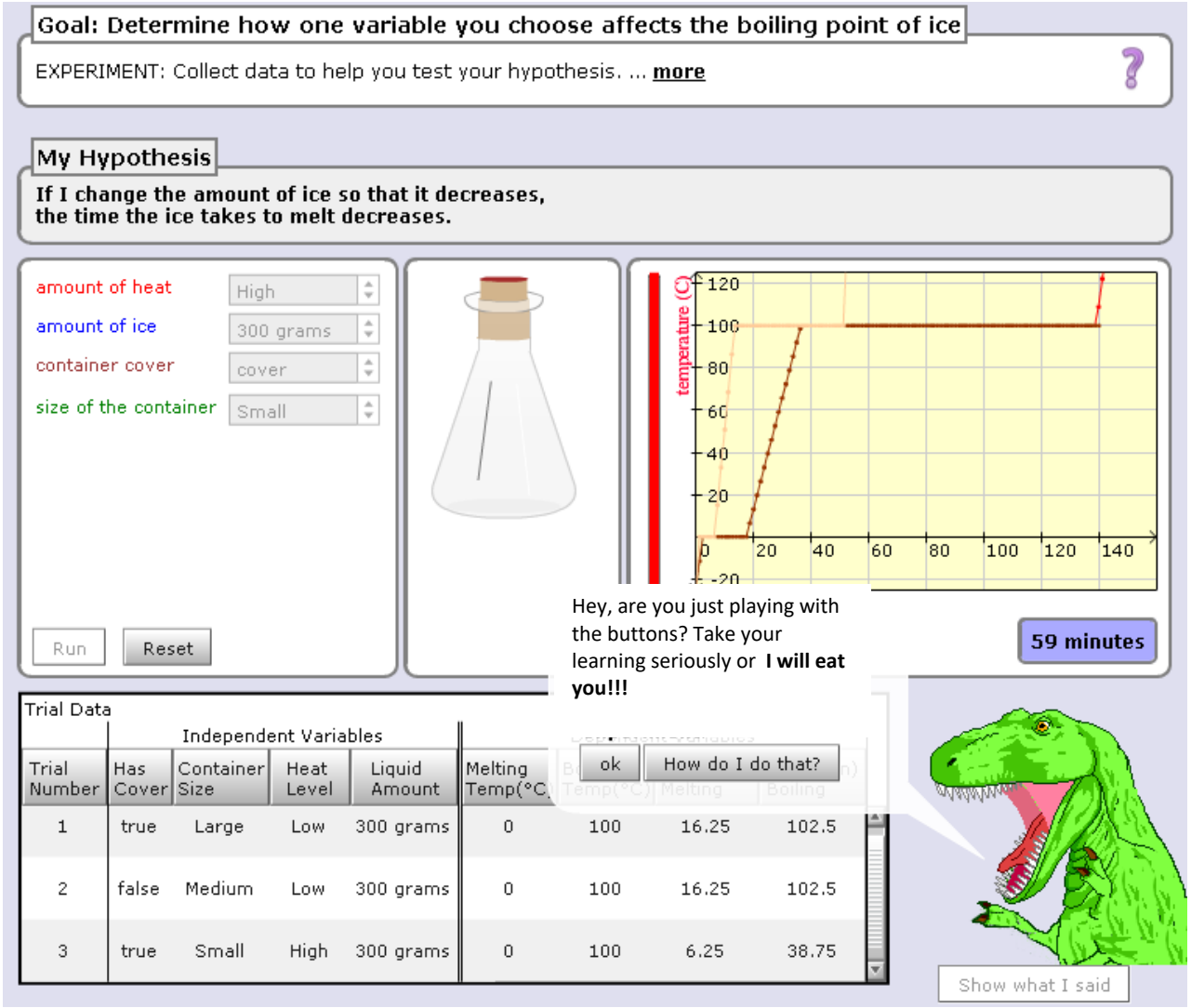
Scooter the Tutor (Baker et al., 2006)





Did students like Scooter?
Only if he didn't help them.

Inq-ITS


Goal: Determine how one variable you choose affects the boiling point of ice
EXPERIMENT: Collect data to help you test your hypothesis. ... more
My Hypothesis
If I change the amount of ice so that it decreases, the time the ice takes to melt decreases.
amount of heat
High
amount of ice
300 grams
container cover
cover
size of the container
Small
Run
Reset
temperature (C)
120
100
80
60
40
20
0
20
40
60
80
100
120
140
59 minutes
Hey, are you just playing with the buttons? Take your learning seriously or I will eat you!!!
ok
How do I do that?
Show what I said
Trial Data
Independent Variables
Trial Number
Has Cover
Container Size
Heat Level
Liquid Amount
Melting Temp(°C)
1 true Large Low 300 grams 0 100 16.25 102.5
2 false Medium Low 300 grams 0 100 16.25 102.5
3 true Small High 300 grams 0 100 6.25 38.75

TC3Sim (DeFalco et al., 2018)

- Frustration detector used to trigger multiple interventions
- Social identity intervention led to better learning outcomes

Conclusion

- Basic research on affect and engagement is ongoing
- The goal: more engaging and positive affective experiences for learners
- And ultimately, better learning outcomes and long-term participation

Learn More

Penn Center for Learning Analytics
University of Pennsylvania

twitter.com/BakerEDMLab

Baker EDM Lab

EdX MOOC “Big Data and Education”
All lab publications available online – Google “Ryan Baker”

Bill & Melinda Gates foundation

LearnLab
Pittsburgh Science of Learning Center

Penn
University of Pennsylvania

Obtaining Ground Truth: BROMP Field Observations

- BROMP 2.0 protocol (Ocumpaugh et al., 2015a)
- Conducted through Android app HART (Ocumpaugh et al., 2015b)

- Protocol designed to reduce disruption to student
  - Some features of protocol: observe with peripheral vision or side glances, hover over student ***not*** being observed, 20-second "round-robin" observations of several students, bored-looking people are boring
- Inter-rater reliability around 0.8 for behavior, 0.65 for affect
  - Only two other published approaches similar in reliability ☺
- Over 150 coders now trained in USA, Philippines, India, UK

Algorithms

- Try small number of algorithms that
  - Fit different kinds of patterns
  - All tend to under-fit (we don't have huge data sets during detector development)

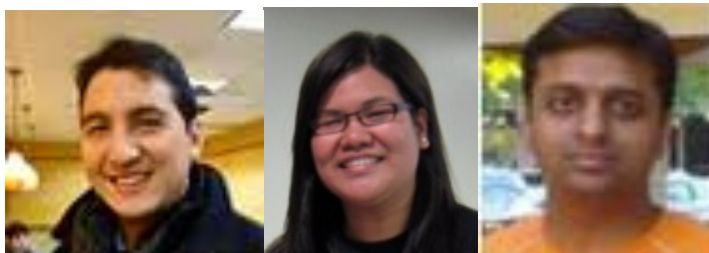
- A few I like
  - Decision Trees (J48)
  - Decision Rules (JRip)
  - Functional Classification (Step Regression)
  - Instance-Based Classification (K*)

Model Assessment

- Models assessed using
- A'/ AUC ROC
  - The model's ability to distinguish when an affective state is present (e.g. is student bored or not)
  - Chance = 0.5, Perfect = 1.0,
    First-level medical diagnostics = 0.75-0.80
- Cohen's Kappa
- Precision-Recall Curve
  - Increasingly often but not in this talk

Model Goodness (Pardos et al., 2013)

Construct	Algo	A'	Kappa
Boredom	JRip	0.632	0.229
Frustration	Naïve Bayes	0.681	0.301
Engaged Concentration	K*	0.678	0.358
Confusion	J48	0.736	0.274
Off-Task	REPTree	0.819	0.506
Gaming	K*	0.802	0.370



Other environments

- Not always boredom that's worst
  - For example, in vMedic, boredom detection was best, with A'=0.85 (Paquette et al., 2015b)
  - Varies by environment

Technical Detail (Ocumpaugh et al., 2014)

- Models trained only on students from a single population (urban, suburban, rural):
  - work well on that population
  - are inappropriate for different populations, where they perform just barely better than chance

- Models trained on the students on all three populations work just as well as single-population models for urban and suburban students
  - Still don't work very well for rural students

Efficacy

- Leads to better learning than traditional homework (Mendicino et al., 2009; Singh et al., 2011)
- Leads to better learning than traditional classroom practice (Koedinger, McLaughlin, & Heffernan, 2011)
- Recent large-scale RCT showing substantial effect (Roschelle et al., 2016)