

**AN INDUCTIVE METHOD OF MEASURING STUDENTS'
COGNITIVE AND AFFECTIVE PROCESSES VIA SELF-
REPORTS IN DIGITAL LEARNING ENVIRONMENTS**

A Dissertation

Submitted to the Faculty

Of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

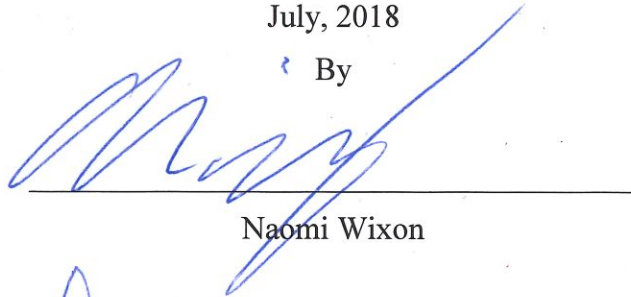
Degree of Doctor of Philosophy

in

Learning Sciences & Technologies

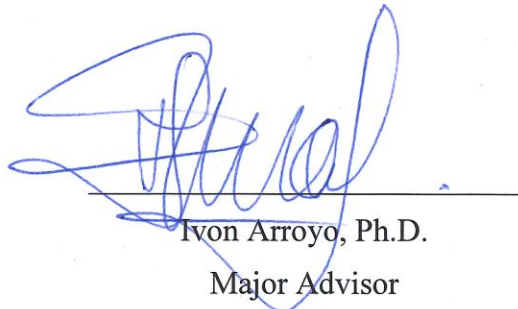
July, 2018

By



Naomi Wixon

APPROVED:



Ivon Arroyo, Ph.D.

Major Advisor

Department of Social Science & Policy
Studies

Worcester Polytechnic Institute

Ryan S. Baker, Ph.D.
Graduate School of Education
University of Pennsylvania

Joseph E. Beck, Ph.D.

Department of Computer Science

Worcester Polytechnic Institute

Kaska Porayska-Pomsta, Ph.D.

Institute of Education

University College London

Abstract

Student affect can play a profoundly important role in students' post-school lives. Understanding students' affective states within online learning environments in particular has become an important matter of research, as digital tutoring systems have the potential to intervene at the moment that students are struggling and becoming frustrated, bored or disengaged. However, despite the importance of assessing students' affective states, there is no clear consensus about what emotions are most important to assess, nor how these emotions can be best measured.

This dissertation investigates students' self-reports of their emotions and causal attributions of those emotions collected while they are solving math problems within a mathematics tutoring system. These self-reports are collected in two conditions: through limited choice Likert response and through open response text boxes. The conditions are combined with students' cognitive attributions to describe epistemic (neither purely affective nor purely cognitive) emotions in order to explain the relationship between observable student behaviors in the MathSpring.org tutoring system and student affect. These factors include beliefs, expectations, motivations, and perceptions of ability and control. A special emphasis of this dissertation is on analyzing the role of causal attributions for the events and appraisals of the learning environment, as possible causes of student behaviors, performance, and affect.

Acknowledgements

I'd like to thank Nuance and Shaughn for giving me a place to stay after my house caught fire. I was living with them when I started work on my proposal and beyond the physical need of shelter I definitely leaned pretty heavy on them in terms of emotional needs as well. Between my having just started HRT and my mom's breast cancer (thankfully in remission) I can't overstate how staying with them kept me from falling apart. Nuance and Shaughn will always be family.

I'd like to thank my mentors here in academia. Ryan Baker has always offered timely and rigorous feedback however busy his schedule is. Moreover his work on quantitative field coding of students' behaviors provided much of inspiration for this work. Janice Gobert taught me how to write to a particular audience, which is a skill I still struggle with. Erin Ottmar gave me a formative lesson on the daunting work of grant application, which gave me a lot more appreciation for Janice's considerable skill. I have benefitted greatly from Erin's commitment to her students' professional growth. Also to my committee members Kaska Porayska-Pomsta and Joe Beck. While Kaska and I haven't spoken much, her review of this dissertation has been perhaps the most thorough after Ivon's. She's given me a lot to consider and many avenues of potential work. Joe's insights are always incisive and pragmatic; he has a way of laying complex concepts bare. While she's not on my committee I'd like to thank Jaclyn Ocumpaugh for her insights and advice, I would have really liked having her on my committee, but I'm pretty confident she'd have told me to keep my committee as small as possible. Also Michael Sao Pedro has always taken an interest and offered words of support and invaluable career advice. Thank you Brittany for coffee, and reminding me to care for myself like a precision machine rather than a delicate flower. Finally, I'd like to thank my advisor Ivon for her copious notes,

meetings, and most of all encouragement throughout this dissertation. There were a lot of late nights, and a lot of revisions and throughout it all Ivon has been a great support.

I will always be grateful to the open-response coders: Danielle Alessio, Rashid Chatani, Taylyn Hulse, Colleen McShea, Tamisha Thompson, and Sarah Schultz. Coding these responses took several hours and then I had folks come back to discuss their own individual coding schemes. It was a very long and involved process and without their commitment this entire dissertation would have been impossible. Thank you all so much.

Specifically, I'd like to give even more thanks to my colleagues Danielle and Sarah though. Danielle is probably the hardest working member of the MathSpring team and I hope that translates into a successful defense. Sarah has worked most closely with me in my open-response coding work including our prior published paper "Blinded by Science". She left for Carnegie Mellon, near the start of my work I often miss her as a labmate to discuss ideas with.

I'd like to thank my parents. If you guys hadn't both gotten your PhDs in psychology at Clark I don't know if I would have viewed a doctorate as something within my reach. Also for the late night vent sessions and advice on the best wording for a self-report prompt. It really does mean a lot to be able to have a parent understand a struggle you're going through. While I'm at it I would also like to thank my aunt Sarah and uncle John who were there at Clark with my parents back in the day. Visiting them, and also Ellen and Thomaz, helped give me perspective on my work and academia in general.

Finally, I'd like to thank Holly for all the support: for the long walks, the flowers, the dinners, the home improvement projects, and the sharing of what's stressing us out at the moment. It's not your job to keep me stable, but I'd be a lot less stable without walking Tigger.

Table of Contents

Table of Tables	8
Table of Figures	10
1 Introduction.....	11
2 Background Literature	15
2.1 Student Beliefs & Volition: The Cognition of Appraisal and Attribution	15
2.2 Control-Value Theory of Emotion.....	16
2.3 Tutoring and Learning Environments that Model Student Appraisal.....	17
2.4 MathSpring	18
3 Motivation.....	21
3.1 Memory is an Unreliable Account of Emotion	22
3.2 Students may not Understand Terminology as Researchers Do	22
3.3 “Forced Choice” Self-Reports may “Induce” instead of “Elicit”	23
3.4 Survey Measures Interrupt Workflow.....	23
3.5 Pilot Study.....	24
4 Research Goals.....	26
4.1 Improving Upon Existing Measures	29
4.2 Consequential Validity.....	30
4.3 Coarse Grained (Student) Level.....	31
4.4 Fine Grained (Action) Level.....	32
4.5 Contributions.....	32
5 Methods	35
5.1 Participants.....	35
5.2 Procedure	35
5.3 Measures	36
5.3.1 Student Level Learning and Performance Measures.....	36
5.3.2 Fine Grained Learning & Performance Measures	36
5.3.3 Aggregation.....	37
5.4 Student Level Affective & Disposition Measures	37
5.4.1 Affective Measures	37
5.4.2 Performance/Learning Measures	38
5.4.3 Fine Grained Affective & Disposition Measures.....	41
5.5 Open-Response Coding Protocol	44

5.6	Open Response Coding Details.....	46
5.7	Multi-Coder Inter-rater Agreement Program.....	48
6	Results.....	51
6.1	Finalized Coding Scheme	51
6.1.1	Tag Descriptions	52
6.2	Inter-rater Reliability for the Finalized Scheme.....	64
6.3	Initial Student Level Analyses	65
6.3.1	Descriptives for Students' Pretest, Dispositional, & Behavioral Measures	67
6.3.2	RQ1: What are the emotions that students report in an open-response assessment, and do they match student emotions from the literature?.....	70
6.3.3	RQ2: Are the ways students feel in a learning environment predetermined by their general attitudes and goals and abilities that students bring to the learning environment?	72
6.3.4	RQ3: How do students express their emotions in an online tutor, and how are these emotions associated with students' behaviors in a digital learning environment?	75
6.3.5	RQ4: Why do students believe they feel a particular way?	78
6.4	Discussion of Initial Student-Level Results.....	86
6.5	Summary of New Research Questions from Student Level Analyses	89
3.	A high degree of “annoyed” reports of emotion seem linked to negative website attributions. What are the events that precede/follow these reports? Is this a case of students externalizing blame due to temporary lapses in performance (i.e. “sour grapes”), or are these students voicing concerns with bugs or errors in the system itself.....	90
6.6	Initial Action Level Analyses: Methods	90
6.6.1	New Hypotheses Building on Prior Work	92
6.6.2	Performance Measure: Probability of Performing Better than Random Guessing	96
6.6.3	Comparison by Student & Dependencies	99
6.7	Initial Action Level Analyses: Results.....	100
6.7.1	Increased challenge leads to confusion	100
6.7.2	Increased confusion leads to positive emotions if resolved/ frustration otherwise.....	101
6.7.3	Frustration leads to boredom through sustained poor performance.....	102
6.7.4	Consistent success may imply lack of challenge leading to boredom, or gratification at success	102
6.7.5	Boredom follows (and precedes) boredom	104
6.7.6	Complaints about the domain or learning environment are likely preceded by poor performance	105
6.8	Discussion of Initial Action Level Results.....	105

6.9	Extensions to Initial Student & Action Level Results	109
6.9.1	What Annoys Students?.....	110
6.9.2	“Bored”, “IDK”, “DTG”, Blank Responses: A closer look at potential Disengagement .	113
6.9.3	What Precedes/Follows Boredom? Is it a mood? Does it increase over a session?.....	113
6.9.4	What leads students to leave self-report prompts blank?.....	116
7	Discussion.....	121
7.1	Limitations	121
7.2	Students’ Perspectives in Self-Report.....	121
7.2.1	Positive Valence.....	123
7.2.2	Neutral Valence.....	123
7.2.3	Negative Valence	123
7.3	Priming Effects: Differences between Open Response and Forced Choice	124
7.4	Improved Forced-Choice Self-Report.....	126
7.5	Extending this Work	127
7.5.1	Collecting Additional Data	127
7.5.2	Additional Analyses: Detector of Affect.....	128
7.5.3	Additional Analyses: Structural Equation Modeling	129
8	Conclusion and Future Work	131
9	References.....	136

Table of Tables

Table 1 Participant Summary.....	35
Table 2 Student Actions & System Events	36
Table 3 Breakdown of How Features are aggregated at the Student Level	37
Table 4 Student Level Affective Measures Gathered via Pre & Posttest.....	38
Table 5 Student Level Learning & Performance Measures Gathered via Pre & Posttest.....	38
Table 6 Student Level Affective and Disposition Self-Reports Aggregated from within the MathSpring tutoring environment	39
Table 7 Student Level Behaviors Aggregated from Actions within the MathSpring tutoring environment	39
Table 8 Results of t-tests and Descriptive Statistics for Math Pre to Posttest Gain, Pre to Posttest Change in Learning Orientation (LO), Pretest Emotion Survey Items, Pretest Learning Orientation (LO), and Math Pretest Scores by Sample Group (if both pre & post complete)	41
Table 9 Example of Kappa Program Confusion Matrix	49
Table 10 Tags used in Finalized Coding Scheme	51
Table 11 Coder Tags Related to the Code finalized as ‘Bored’, for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts.....	53
Table 12 Tags Related to the Code finalized as “IDK”, for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts	54
Table 13 Coder Tags Related to the Code finalized as “DTG”, for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts.....	55
Table 14 Coder Tags Related to the Code finalized as Positive for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts.....	56
Table 15 Coder Tags Related to the Code finalized as Negative for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts.....	57
Table 16 Coder Tags Related to the Code finalized as Easy for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts	57
Table 17 Coder Tags Related to the Code finalized as ‘Hard/Confusing’ for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts	58
Table 18 Coder Tags Related to the Code finalized as ‘Material’, for Forced Choice Attribution, Open Response Attribution, and Open Response Agency prompts.....	59
Table 19 Coder Tags Related to the Code finalized as ‘Success’, for Forced Choice Attribution, and Open Response Attribution prompts	59
Table 20 Coder Tags Related to the Code finalized as Growth, for Open Response Attribution, and Open Response Agency prompts.....	60
Table 21 Coder Tags Related to the Code finalized as Website, for Forced Choice Attribution and Open Response Attribution prompts	60
Table 22 Coder Tags Related to the Code finalized as Failure, for Open Response Attribution prompts..	61
Table 23 Coder Tags Related to the Code finalized as annoyance, for Open Response Feeling prompt...	62
Table 24 Coder Tags Related to the Code finalized as neutral, for Open Response Feeling prompt.....	62
Table 25 Coder Tags Related to the Code finalized as bugs, for Open Response Agency prompt.....	62

Table 26 Coder Tags Related to the Code finalized as design, for Open Response Agency prompt	63
Table 27 Coder Tags Related to the Code finalized as fun, for Open Response Agency prompt	63
Table 28 Coder Tags Related to the Code finalized as quit, for Open Response Agency prompt	64
Table 29 List of Tags, Total Instances, & Kappas for re-coding by final coders: Coder N and Coder S.....	65
Table 30 Aggregate Behavior Measures Considered for Analyses	67
Table 31 Summary of Emotion & Attribution Tags Used, Total Instances (N), & Cohen’s Kappa of Interrater Reliability for Open-Response and Forced Conditions. Values in bold if N or Kappa are unacceptably low.	68
Table 32 Likert Scale Self-Reports for Closed Response Condition	69
Table 33 Descriptives for Emotion Tags for Open Response Condition	70
Table 34 Emotions vs Pre/Posttest Measures: Bivariate Correlations	74
Table 35 Emotions vs Behaviors: Bivariate Correlations	76
Table 36 Emotions vs Attributions: Bivariate Correlations	78
Table 37 Emotions vs. Attributions Count of Instances	81
Table 38 Infrequent & Partly Invalid Emotions vs Attributions: Bivariate Correlations	82
Table 39 Infrequent & Partly Invalid EMOTIONS vs ATTRIBUTIONS Count of Instances.....	83
Table 40 Correlations between Pretest measures & Frequent Emotion/Attribution Pairings.....	84
Table 41 Correlations Between Emotion/Attribution Pairings & Actions	85
Table 42 Emotional and Causal Attribution Measures for Open and Closed Response Conditions.....	95
Table 43 Example Cases Leading to 95% Better than Chance Likelihood.....	98
Table 44 Initial Fine Grain Results Summary: Results with marginal significance ($p < 0.1$) in accordance with hypotheses labeled “CON”, non-significance “INC”, significance counter to hypothesis “DIS”	104
Table 45 Attributions for Annoyance/Frustration	111
Table 46 Problems prior to Report of Frustration (N = 23).....	112
Table 47 Problems prior to Report of Annoyance (N = 15)	112
Table 48 Reports which precede/follow reports of Boredom	114
Table 49 Interest/Boredom SOF per Problem	115
Table 50 Interest/Boredom Wrong per Problem.....	115
Table 51 Bored vs Not Bored Over Time.....	116
Table 52 Forced-Choice Reports which precede/follow reports of Blank Reports	117
Table 53 Open Response Reports which precede/follow reports of Blank Reports	117
Table 54 Blank vs Not Bland Over Time	118
Table 55 Open Response Reports which precede/follow reports of Neutral Emotion and IDK attribution	119
Table 56 Reports which precede/follow reports of “DTG”	119
Table 57 Mean difference between behaviors coincident with, prior to, and following reports of “DTG” with paired samples t-test significance (p)	120
Table 58 Boredom & Low Excitement/Interest vs Attributions as Percentage of total	125

Table of Figures

Figure 1 MathSpring Intelligent Tutoring System	18
Figure 2 "Jake" one of MathSpring's Learning Companions fostering a Growth Mindset	19
Figure 3 MathSpring's Student Progress Page offering Feedback on Students' Skill Development.....	20
Figure 4 Forced-Choice Self-Report Prompt that students encountered in MathSpring every 5 minutes or 8 problems	42
Figure 5 Open-response Self-Report Prompt that students encountered in MathSpring every 5 minutes or 8 problems.....	43
Figure 6 Disequilibrium Hypothesis: Increased challenge leads to confusion.....	93
Figure 7 Productive Confusion Hypothesis: Increased confusion leads to positive emotions if resolved or frustration if not.....	93
Figure 8 Hopeless Confusion Hypothesis: Frustration leads to boredom through sustained poor performance	94
Figure 9 The Disengagement Hypothesis: Consistent success may imply lack of challenge leading to boredom, or gratification at success	94
Figure 10 Persistent Boredom Hypothesis: Boredom precedes & follows boredom.....	95
Figure 11 "Sour Grapes" Hypothesis: Complaints about the domain or learning environment are likely preceded by poor performance	95
Figure 12 T-Test Comparing Performance Observed to Performance Expected due to Random Guessing	97
Figure 13 Proposed Future Self-Report Prompt	126

1 Introduction

Student affect—the attitudes, interests, and values that students exhibit and acquire in school—can play a profoundly important role in students' post-school lives, possibly an even more significant role than cognitive achievements (Popham, 2009). Affect is recognized as a key indicator of student engagement and a variety of assessments of affect have shown affective constructs to be important predictors of learning (Linnenbrink-Garcia & Pekrun, 2011; Pardos et al., 2013; San Pedro et al. 2013) which raises the question: why isn't it assessed more often? In part, the answer is that evaluation of students' affective states remains a difficult challenge. No clear gold standard exists for identifying affective states, which has driven researchers to re-examine the intersection of general theories and concrete measurement methodologies (Graesser & D'Mello; 2011).

Many affective states in learning environments, such as boredom, confusion, frustration, and engaged concentration, are characterized as having an epistemic nature (Pekrun, 2010; D'Mello & Graesser, 2012). Epistemic states may be described as emotional (Silvia, 2009), or cognitive (Clare & Huntsinger, 2007), because they are often operationalized as partly dependent on particular events or cognition (Baker et al, 2010). Confusion is operationalized as an internal experience where the student is being confronted with an impasse and being uncertain what to do next (D'Mello & Graesser, 2012) or as the student experiencing challenge while attempting to understand a specific situation (Ocumpaugh, 2015). Boredom has been characterized as a state of disengagement from a learning task, or as a state where the student decides not to pursue a learning goal (D'Mello & Graesser, 2012). This makes it different from a student being “on task”, but simply disengaged (Ocumpaugh, 2015). Both D'Mello (2012) and the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) (Ocumpaugh, 2015) distinguish the same six

emotional states: Boredom, Confusion, Delight, Engaged Concentration, Frustration, and Surprise. Pekrun (2010, 2016) recognizes the same constructs with the exception of Engaged Concentration, and the addition of Anxiety and Curiosity as possible emotions.

In addition to the ambiguous cognitive/emotional nature of many of these epistemic emotions, there is also uncertainty regarding which constructs to consider. BROMP (Ocumpaugh, 2015) emphasizes boredom, confusion, engaged concentration, and frustration as typically being more prevalent. This is in part based on prior work which found lower incidence of delight or surprise (D'Mello & Graesser, 2012).

This following work investigates whether factors that are neither purely affective nor purely cognitive might moderate and explain the relationship between observable events (e.g. student log data from digital learning environments) and student affect. These factors include beliefs, expectations, motivations, and perceptions of ability and control. The goal of this dissertation is to analyze whether a link between behavior and affect exist, with a high emphasis on analyzing the role of causal attributions for the events and appraisals of the learning environment, as possible causes of student affect. A definition of each of these terms follows.

Student beliefs can be summarized as a student's assessment of a learning environment as it relates to that student's motivation. Students may believe a learning task to be valuable or not valuable, but even if they believe the task is valuable, they might believe the task's value is as a means to gain recognition (performance oriented learning), or valuable because of being a valuable task to further their own growth (mastery oriented learning). Given successive similar learning events students internalize an expected series of interactions and for beliefs based on these prior experiences. An example of this would be Carol Dweck's (2006) growth mindset,

wherein individuals may believe their abilities are fixed and immutable or a product of effort; whether or not this belief is true the narrative of events tends to follow students' expectations.

At the same time, motivation (Lepper & Henderlong, 2000) can be defined as students' drive (or lack thereof) to pursue means to achieve goals, the approach or avoidance of a task, which in the case of learning environments may include: *learning goals* to improve one's expertise, *performance goals* for external recognition of one's performance, or *work avoidance goals* to minimize required effort (Harackiewicz et al, 2002). Meanwhile, attributions, or causal attributions, are specific causal beliefs students may hold as to why particular events may have occurred. While the definitions of these terms are important to understanding this document, they are meant to inform our interpretation of students' open-ended self-reports, rather than rigidly adhering to specifically operationalized constructs. .

Given the complexity and overlap of the aforementioned constructs, this dissertation proposes an open analysis of students' self-reports of their emotions, and the causal attributions of them. This approach is based on four motivations. First, open-ended self-reports may highlight constructs that we had not previously considered. Second, if students volunteer a particular emotion, cause of an emotion, or motivation without specific prompting then we may trust that the construct exists within their own conceptualization of their learning environment, and that it is not a product of leading questions. While some responses may be due to social desirability (i.e. telling educators what we want to hear) that glimpse at students' understanding of our own goals may be a useful measure as these goals have already been internalized by students from a source other than self-report prompts. Third, as I am addressing emotions in their relation to cognition (Clore & Ortony, 2000) the most direct way to measure these internal cognitive processes is via open-response self-reports. Fourth, I hypothesize that these cognitive attributions may act as a

proxy between students' emotions and behaviors as they are in themselves a cognitive link between events and feelings as articulated the student in vivo. It is my hope that the inclusion of these cognitive attributions will lead to more accurate predictions of both self-reported emotions and behaviors by including a cognitive component that has previously been missing from many computational models of students' interactions within an intelligent tutoring system environment. Please see section 4 "Research Goals" for a more complete description of the intended contributions of this work and central philosophical motivation. I shall begin by discussing prior work that has analyzed links between cognitive attributions and students' emotional states and behaviors.

2 Background Literature

2.1 Student Beliefs & Volition: The Cognition of Appraisal and Attribution

Prior work by Rotter, (1966), Weiner (2010), and Elig & Frieze (1979) explored emotional states as a product of students' causal attributions of academic success or failure. Weiner (2010) provides a good initial summary of this work by articulating an expectancy-value model of behavior, affect, and motivation. The expectancy-value model explains students' behaviors by their expectation of success (or performance in general) and the value they place in the learning task. Weiner (1979) found that some emotions (happiness and disappointment) were independent of attributions, but were not independent of outcomes: whether students attributed their success or failure to internal or external causes was shown to be predictive of students' emotional states. Students who believed they were responsible for their own success reported feelings of *pride*, *competence*, and *confidence*. However, students who attributed their success to external causes were more likely to express *gratitude*, *thankfulness*, *surprise*, or even *guilt*.

The emotions experienced by students with attributions of failure similarly depended on whether those attributions were internally or externally directed. Students who reported feeling responsible for their own poor performance were more likely to report *guilt* or *resignation*, while students who attributed their failure outward more likely to report *anger* or *surprise*. As an extension of these findings, it's possible that students who harbor a sense of *guilt* for their poor performance may behave differently than those who express *anger*. Weiner's work modeled students' expectation and valuation of the outcome of a learning task as deterministic of their emotional state in addition to empirically observable events.

This background research inspired my methodology for this dissertation: the inclusion of self-report data of expectations, attributions, and valuation towards the construction of models of

student affect for emotion detection. By encouraging students to report their thoughts and feelings in an open ended way I hope to uncover how students' actions are influenced by their perceptions of their interaction with the learning environment.

2.2 Control-Value Theory of Emotion

Reinhard Pekrun (2007) extended Weiner's contributions by creating the control-value theory of achievement emotions. This theory provided a framework to describe causes and effects of emotions students experience in academic contexts. The control-value theory proposed that emotions experienced vary depending students' could focus on academic performance and tasks: prospective focus on future tasks (e.g. anxiety, hope, hopelessness), retrospective focus on past tasks (e.g. pride, sadness, shame, joy), and activity focus on current or ongoing tasks (e.g. frustration, boredom, enjoyment). Each focus allows for different emotions to arise.

Additionally, Pekrun proposed relationships between three continuous variables: perceived success/failure of a task, perceived value of a task, and perceived degree and locus of control a student has while performing a task (Pekrun et al 2007), which affect which emotions arise. The control-value framework allowed for further examination of interactions between terms. For example, both expectancy of success or failure and valuing of success or failure were hypothesized to combine in multiplicative ways. The more value students assign to a task the greater pleasure they will experience with success and displeasure with failure, while a task perceived to have little value would likely result in boredom regardless of the outcome (Pekrun et al 2007).

This work builds on Pekrun's, in that it focuses on the combination of students' cognitive attributions, their reported emotions experienced, and the fine grained expression of student

behaviors in a digital learning environment, which are tracked through log files of student behavior events (mouse clicks in specific situations, timing of events, entered answers).

I hypothesize that these internal cognitive processes can help to explain the relationship between affect and behavior, as explained further in section 3.

2.3 Tutoring and Learning Environments that Model Student Appraisal

Recent work has considered student appraisals within digital tutoring and learning environments, called Intelligent Tutoring Systems (ITS). Typically, work of this nature adopts the OCC model of cognitive appraisal of emotions (Clore & Ortony, 1988). The OCC model provides an organizational framework regarding the interactions between cognitive appraisals of particular emotional states and the emotional states themselves. Incorporating OCC into an ITS learning environment comes with the challenge of practical implementation and application. However, existing research has accomplished this feat through means such as direct survey measures of specific factors included in appraisal theory. Firstly, Sabourin's work (2011) within the Crystal Island ITS incorporated students' achievement emotions (e.g., anxiety, boredom, frustration, etc.) and goal orientation (whether focused on performance or learning). Secondly, Conati's (2009) extensive work in emotion and appraisal using the OCC model accounts for students' goals, personality traits, emotional states, and perceptions of the environment. Conati's (2009) work astutely avoids the problem of cognitive load inherent in asking students' to self-report on each of these dimensions. It does this by limiting self-reports to two simple and brief likert scale forced-choice prompts: "How do you feel about your game playing?", and "How do you feel about the agent?" (Conati & Maclaren, 2009).

Research on open learner models (OLMs) also approach students' self-appraisal of their learning (Dimitrova, 2003; Bull & Kay, 2007). While these models typically focus on students'

cognitive state and their mastery of educational materials, they also often require students' investment in learning goals. Although the accuracy of students' self-assessments may be debatable (Kruger & Dunning, 1999), the mere act of self-assessment may lead students to take greater responsibility in their learning (Boud et al. 1996; Bull et al. 1995). It remains to be seen if expanding these OLMs to include students' emotional states could shift students' relationships with their own emotional states. For example, perhaps through reporting on negative valence emotional states students might experience some form of relief similar to Sabourin's (2011) work showing that disengaged behaviors may lead students to re-engage with a learning task.

2.4 MathSpring

MathSpring is an intelligent tutoring system (ITS) which addresses middle school (6th through 9th grade) math content including number sense, pre-algebra, algebra, geometry (see Figure 1).

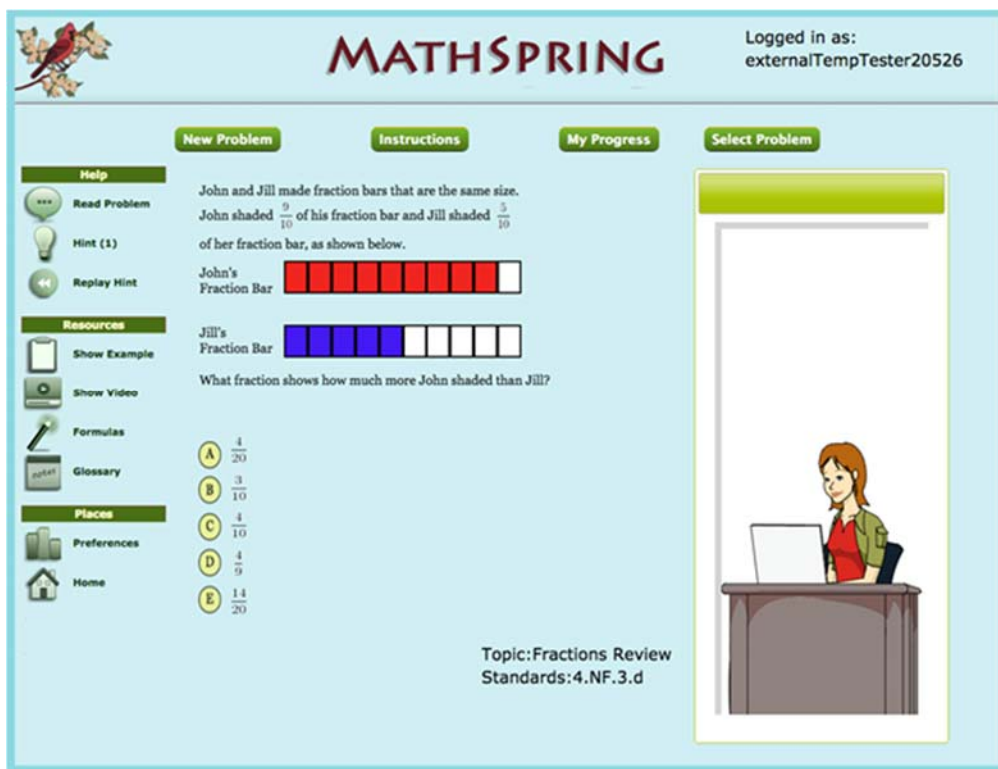


Figure 1 MathSpring Intelligent Tutoring System

MathSpring adapts difficulty level based on students' ability and scaffolds students with multimedia hints and pedagogical agents known as "Learning Companions" provide cognitive support (Woolf et al., 2010) in the form of problem solving strategies as well as motivational support by fostering a growth mindset (Dweck, 2006) in students (see Figure 2).

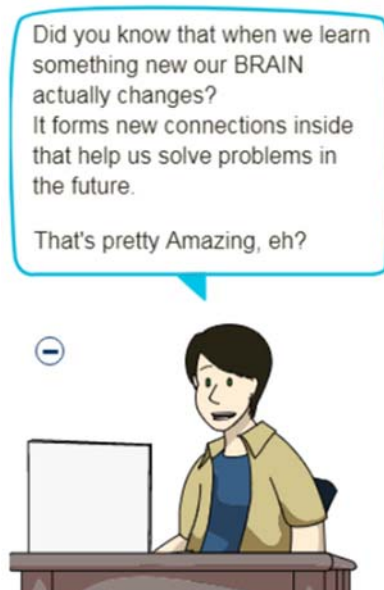


Figure 2 "Jake" one of MathSpring's Learning Companions fostering a Growth Mindset

MathSpring is based partially on an approach of cognitive apprenticeship (Collins et al., 1989) meant to bring cognitive processes out into the open so that students' may metacognitively build on their skills. MathSpring does this by emphasizing the three steps of cognitive apprenticeship through: modeling with example problems, offering coached feedback and hints to support students' attempts, and finally through reflection using the student progress page (see Figure 3). Furthermore, MathSpring involves instructors in the ITS learning environment through live updates on student progress in MathSpring's "Teacher Tools" which highlight who in class may be struggling and what content may they may find excessively challenging.




Topic	Remark	Performance	Effort	Action
Fractions Review	As you put more effort on solving the problems, the baby pepper plant grows to give pepper fruits. Comment	Mastery Level <div style="display: flex; align-items: center;"><div style="width: 10px; height: 10px; background-color: green; margin-right: 5px;"></div><div style="border: 1px solid gray; padding: 2px 10px; background-color: #e0e0e0;">10</div></div> Problems Done : 5/26 Learn More		Continue Review Challenge

Figure 3 MathSpring's Student Progress Page offering Feedback on Students' Skill Development

3 Motivation

As mentioned before, there has been other prior work emphasizing students' emotion as the result of the appraisal of a situation in terms of value and control exerted over the task/domain. This appraisal implied a cognitive component over the judgment of the situation in which the emotion arises. While emotion might be overt, the cognitive appraisal is covert and not visible, thus asking students to self-report the reasons for their emotions is one the few means to tap into such appraisals. There are two major reasons this work is limited to self-report of emotional states. First, the simple argument of limiting cost and scope of this work: a design goal of this work was to explore and justify improvements to the currently existing Likert style forced choice self-reports used within the MathSpring learning environment. Secondly, and more importantly, the constructs to be examined herein are not pure emotional states in and of themselves, but rather students' articulation and subjective understanding of their own emotional states. While this construct of students' self-described affect may be closely linked to students' internal emotional states the fact the nature of examining students' own understanding of their emotional states necessitates gathering these data via self-report measures.

However, it is not well understood what is a proper and accurate way to collect students' appraisals of a situation. Students' self-reported emotion data comes with potential risks and practical concerns, as explained next. These concerns have motivated this design, which is why understanding them is important to critiquing and advising how methodology might be improved upon. First I shall list supporting evidence for each concern followed by the problem statement, goals, and proposed methods.

3.1 Memory is an Unreliable Account of Emotion

There is reason to doubt the accuracy of our recollection of emotions, in retrospect. For instance one study found that students reported consistently stronger affect regarding their schoolwork in a posttest survey outside of the learning environment than they did during a learning task (Bieg et al, 2014). At the same time, it is possible that the cognitive appraisals themselves may change during an extended learning task. For example, students may believe a task is initially challenging but later quite easy or vice versa; alternatively students may believe a learning environment is fairly or unfairly designed depending on their experience. A post hoc summary would be less likely to capture dynamically shifting beliefs and perceptions. As a result, a method to measure such perceptions in the moment, and at a fine-grained level of detail, would be preferable.

3.2 Students may not Understand Terminology as Researchers Do

Often when researchers use affective terminology, terms come loaded with additional connotative meanings that students may not share. As a specific example of this phenomenon: a researcher's operationalization of "bored" may not match a particular student's operationalization of "bored" (Porayska-Pomsta et. al., 2013; Ocumpaugh et. al., 2015; Bieg et al., 2014). This may be due in part to the fact that communities of research strive to reach consensus on terminology. In a prior pilot study, students sorted affective terms and facial expressions with regard to valence and activation; the main result of the work was that students' responses varied widely (Wixon et al. 2015). As a result, "forced-choice" measures of self-report (which require students to select from a given set of responses) may provide inaccurate responses as students parse and interpret forced-choice measures differently than researchers' might intend (Porayska-Pomsta et. al., 2013; Bieg et al., 2014).

3.3 “Forced Choice” Self-Reports may “Induce” instead of “Elicit”

In response to the previously mentioned difficulty of students’ inconsistent understanding of terminology, a common solution is to initially explain the meaning of survey measures to students to ensure their understanding of terminology matches researchers’ understanding (Porayska-Pomsta et. al., 2013). However, providing an explanation to students before they report their emotions or affective predispositions, may induce students to answer in a particular way, rather than eliciting a genuine response. Another difficulty is that students’ internal affective experiences might not be listed among the choices we provide in self-report prompts within the tutoring environment or pre and post survey measures. Further, other studies have shown that the act of reporting can alter students’ affective state (Kassam & Mendes, 2013; Ocumpaugh et. al. 2015). In conclusion, accounting for possible bias which traditional self-report prompts may introduce would be an important contribution of this dissertation. By removing the subtle suggestion of asking about particular emotions we can account for this possible bias.

3.4 Survey Measures Interrupt Workflow

Given the initial point of memory being unreliable, it seems preferable to ask students’ to report their feelings and associated thoughts the moment that those perceptions occur. However, one danger is that self-reports may disrupt students’ work flow (Ocumpaugh et. al. 2015). Self-reports which require additional cognitive load, e.g. recalling instructions regarding the meaning of new terminology and how it relates to work, completing a lengthy or highly detailed listing of items or options, would likely exacerbate this risk. It is clear that self-reports, if used as a methodology to assess students’ affect, should be as noninvasive as possible. For this work, the point of remaining non-invasive is less to guarantee accuracy of reports as the constructs in question are ones of students’ self-reflection of their current emotional state and perceived

causes rather than student emotions in and of themselves; the more concerning aspect of self-reports here is the fact that taking time out to consider one's emotional state may distract from attention paid to learning tasks.

Finally, students may opt to apply minimal effort responding to survey measures. Surveys which require students to select from among a series of choices may be filled out without care or reflection (e.g. responding "Very Much So" to every Likert scale item). In this specific example, students' genuine desire to be left alone could be misread as the student having very strong feelings, when in reality the student was upset about the assessment itself. It is clear that forcing students to report on their emotions might not be an ideal way to assess emotion.

3.5 Pilot Study

In anticipation of this dissertation a study was conducted using self-report prompts already available within the MathSpring digital learning environment (Schultz et al 2016). Students from two sample groups collected in 2015 (N = 449) and 2011 (N = 464) were asked to make open ended causal attributions of their self-reported emotions. The process is described in greater detail in sections 5.5 and 5.6 which address how the coding scheme for open-response self-reports was determined.

Most students' self-reports of cognitive attributions were described as "positive" or "negative" in terms of valence and were directed either "internally" or "externally" being attributed either to themselves or to external factors like the digital learning environment, or the domain of mathematics itself. However, all of these attributions were in response to "forced-choice" self-report prompts which were regarding one of four pre-determined emotional constructs (confidence, excitement, frustration, and interest). It was unclear how dependent the open-response cognitive attributions collected were on the prior forced-choice emotion prompts:

while certain prompts were more likely to result in a given attribution this could be a result of the specific emotion students were asked about, or the internal emotions students were experiencing (if those emotions did indeed differ from the prompt).

Further, in many cases students responded to the cognitive attribution prompt with constructs typically considered to be within the purview of emotion: describing their experiences of boredom or their like or dislike of the material and learning environment. These students may have been using the cognitive attributional prompt to address emotional experiences they were having as the forced-choice prompt did not include the most apt descriptions of what they were feeling. To address this concern and the prior concern of priming students with forced-choice emotion reports I decided to test a prompt which included only open-ended prompts. This way I could test to see if completely open-response prompts yielded different results from those found using open-response cognitive attribution prompts which followed forced-choice emotion reports.

4 Research Goals

The main goal of this dissertation is to understand how tracking students' appraisals of a situation may help explain emotions and behaviors within a digital learning environment. Attribution and appraisal data including students' motivation and volition may allow us to predict students' future behavior more accurately than a combination of pure affective observations and behavior alone. Tracking students' causal attributions of their emotional states may allow for two possible improvements on current affect detection. First, it should be possible to see if students' attributions for why they feel a particular way are reflected in logged data of students' actions. If, for example, students claim they are bored due to easy material or frustrated due to challenging material it should be possible to look back to find log data which support these claims, the absence of supporting data may also yield important information regarding how students view their experiences as compared to how researchers' views. Second, regardless of students actual emotional state or the actual cause thereof, by inviting students to construct a causal relationship between events and their current emotional state we can more readily alter the learning environment to address their perceived needs. It is my intention that by including attributional and volitional components to students emotional self-reports we may form a closer link between the events which occur in an online tutoring environment and students' perceived well-being, taking us one step closer to closing the loop of affect adaptive online tutors.

This combination of cognitive attributions with emotional components are already present within epistemic emotional states. As described in the introduction, epistemic emotions may be described as emotional (Silvia, 2009), or cognitive (Clore & Huntsinger, 2007), because they are often operationalized as partly dependent on particular events or cognition (Baker et al, 2010). A good example of an epistemic emotion would be confusion which is often

operationalized as both a state of cognitive disequilibrium as well as the feeling of being uncertain how to proceed (D’Mello & Graesser, 2012).

The decision to analyze students’ cognitive attributions in relation to their emotions led to two more specific research questions:

- a) What constructs ought to be considered?
- b) To what extent should these states be described as cognitive, affective, or epistemic, a combination of cognitive and affective?

An association between students’ emotions and cognitive appraisals has been identified and researched within the control-value theory (CV) of emotions already (Pekrun, 2006). The control-value theory identifies students’ emotions in terms of students valuing of a task’s outcome and their perception of their degree of control in achieving a desirable outcome or avoiding an undesirable outcome. For example students who believe they risk failure in a particular task might anticipate feeling relieved if they believe themselves to have a high degree of control, or feel hopeless if they have a low degree of control. However, these associations have not been investigated within digital learning environments, nor in the moment they occur, which might be very important as expressed earlier. In addition, they have not been explored via a “bottom-up” approach either, starting from student data, but using a top-down approach, starting from prior theory which may originate from especially different student groups or learning environments. Rather than addressing these questions using a top-down approach, by selecting a set of theoretical constructs and then operationalizing them accordingly, I choose instead to approach these research questions from an empirical bottom-up approach, by providing students with free text open-response prompts, to then find common themes within the responses.

The affective and attribution constructs resulting from this research will be compared to the constructs that have been considered so far by other researchers in this field of emotions in education and learning technologies, to analyze differences and potentially novel contributions.

Besides the relationships between affect and cognitive attributions, a further goal of this research is to find associations between student affect/attribution with student behavior, by tracking students' written cognitive attributions alongside their students' actions within the tutoring environment. This leads to the third research question:

c) What are associations between student affect and attributions with student behavior?

Factors that are neither purely affective nor purely cognitive (but instead, something in between) may moderate and explain the relationship between student affect and observable events (e.g., log data). For example, students may attribute feelings of boredom to material that is unchallenging, or frustration to material that is excessively difficult. Alternatively, some students may enjoy particularly easy material and report feelings of confidence. These examples of emotion due to ease/difficulty relate to emotions largely in terms of students' perceived degree of control (i.e. dominance) in the learning environment based on academic ability (Broekens & Brinkman, 2013; Fontaine et al., 2007). However, boredom may be modeled in terms of whether or not students perceive a task has value: if students don't see a learning task as important they may experience boredom regardless their degree of control (Pekrun et al. 2007). Further, students may have prior mistrust of learning environments due to general student disengagement (Henry, 2007; Henry et al., 2012). I have cited a few possible factors which may influence students' emotional states within learning environments; each of these factors may contribute to students' interactions with a learning environment far more complex ways than I have described and further this list of factors is by no means exhaustive. Yet rather than presupposing a variety of

possible culprits for students' emotions and running the risk of overwhelming students with exhaustive survey measures designed to test for specific things (likely resulting in several null effects) I suggest that it might be better to first survey students on their perceptions of how their current emotions relate to possible causes within their learning environments. Let us use the model of physician and patient as an analogy for the relationship between educator and student: a physician begins by openly inquiring as to a patient's symptoms. Despite a patient's considerable lack of medical expertise as compared to a physician, a patient's description of their subjective experience of symptoms provides a physician with a starting point for further diagnosing the patient with specific tests, or treating the patient if diagnosis appears immediately evident based on symptoms reported.

This is why I propose self-report data as an assessment mechanic, a relatively direct and simple means to collect information about students' causal attributions for their feelings as well as chosen strategies to interact with a tutor environment. I intend on analyzing how these reported emotions and cognitive attributions relate to aggregates of student behavior, such as mistakes, help requests and other behaviors that are expressions of engagement and disengagement.

4.1 Improving Upon Existing Measures

Our research group has been gathering data through forced-choice self-report measures for several years, alongside open-response measures of causal attribution (see section 5 "Proposed Measures", particularly Figure 4 for a summary). Those affective self-reports asked students to report one of four experienced emotions (confidence, excitement, frustration, interest) via a forced choice scale going from "very [emotion]" to "not at all [emotion]", and an open-ended

text box for the student to attribute their emotion to reasons ('And why is that?'). This also constitutes one of the conditions used in this analysis, as described later.

However, based on the prior concerns mentioned, this dissertation explores the benefit of moving to an entirely open-response model of emotion reporting, where the student reports emotions via an open-text box.

One of the main goals of this dissertation is to evaluate and determine whether moving from forced-choice to open-response measures would make students' emotional self-reports more closely related to their cognitive attributions of their emotional states and further that this combination of emotional self-report and cognitive attribution might be predictive of and predicted by student behaviors within an online learning environment.

4.2 Consequential Validity

In having students describe their expectations, feelings, and values regarding their work we hope to accurately predict these appraisals from a combination of prior events and students prior appraisals. Additionally, I would like to predict students' future actions from students' prior reported emotions and cognitive attributions. Again, this work is meant to more closely link students' behaviors to their reported emotions by way of cognitive attributions. By getting a more complete model of students' perceptions of their emotions and the causes thereof we hope to tailor students' interactions with the learning environment to address elements students believe to have negative impacts on their experience.

Rather than placing emphasis on the veracity of reports, the primary goal I have is to determine which commonly reported symptoms can be used to predict students' actions and be predicted by system events. Predictions will be made at two levels of granularity. The first level,

“coarse grained”, wherein students’ self-reports may be predicted from pre and posttest survey and assessment items. The second level, “fine grained”, wherein particular sequences of actions may be predicted from students’ self-reports and vice versa (i.e. self-reports predicted from actions).

4.3 Coarse Grained (Student) Level

The pre and posttest survey measures are meant to determine student trait variables and an aggregate measure of students’ appraisals during the use of an ITS learning environment. A simple first sub-goal here is to determine how similar pre and posttest measures are to self-reports made within the tutor (Bieg et al, 2014). Secondly, I plan to examine how students who respond in a particular way to student trait survey measures later view the learning environment. For example, students may claim to be motivated by learning goals on a pretest survey and later within the learning environment report performance goals; if the environment is inducing performance goal orientation we might see a change in student performance as shown in prior work (Butler, 1993; Block et al., 1995). Asking a specific question may prompt a particular response from a student. However, asking them to simply report their predominant concerns may illuminate perspectives on the learning environment that are wholly orthogonal to the presumptions of a forced-choice measure. Yet, this approach would still allow for students to report impressions that align with forced-choice measures. Finally, pre to posttest learning gains and average performance may be compared against aggregate measures of students’ self-reports within tutor: students’ emotions and attributions during a learning tasks at discrete points of time may be related to overall learning gains and performance over the course of a six month semester. These analyses will be performed with simple correlational studies.

4.4 Fine Grained (Action) Level

As coarse grained analyses dealt with changes over an entire session within an ITS, fine grained analyses focus on changes from one action to the next. Students' judgment of their feelings and associated causes/attribution may explain the strategies they employ and the degree of enthusiasm/commitment they apply to these strategies. We may see these judgments reflected in strategies and styles of use of the tutoring system, engagement behaviors or disengagement behaviors. Likewise, tutorial actions and specific pedagogical actions (e.g. offering examples, suggesting hints, making students reflect on their performance) may cause different reactions on students of varying judgments of emotions and reasons/causes.

4.5 Contributions

This work will act as a foundation for investigating the role of open response self-reports within online tutoring environments. While there is an extensive body of work from several decades ago examining student's causal attributions in terms of expectancy value within learning environments (Frieze, 1976; Frieze & Snyder, 1980; Weiner, 1985; Weiner et al., 1979), these self-reports are likely subject to cultural shifts over the years, between populations sampled (Rodrigo et al., 2010; Ocumpaugh et al., 2014), and finally due to the rather large move from pencil and paper assessment to modern intelligent tutoring systems which adapt based on students' needs.

The potential for changes in what students feel to be salient emotions or the causes thereof lead to the next contribution of this work: an adaptive coding scheme meant to capture meaning from students open response self-reports. There is a tension between designing an emotion self-reporting tool that is applicable to a particular group of students in a particular learning environment and having that tool generalize to new populations which may report

different emotions or causes entirely. This work serves as an initial case of creating and testing a final measurement tool and documenting the process of the tool's creation that it may be entirely recreated to properly address the needs of distinct populations working in distinct learning environments.

The primary mechanism used to test measurement tools is through inter-rater agreement measures of Cohen's kappa. Cohen's kappa has been used in the past as a means to measure agreement between separate coders looking for a particular construct by comparing the number of agreements and disagreements about observing this construct in question (Cohen, 1960; Ocumpaugh et al., 2014; Gobert et al., 2015; Henrie et al., 2015). Cohen's kappa as applied here measures the agreement between coders who have not been instructed to identify a particular construct or list of constructs, but rather coders who have been instructed to try and identify summary tags to classify open response self-reports. As such, the Cohen's kappa here provides not only a matter of reliability of a defined test measurement, but whether multiple parties will independently create the same sorts of constructs without instruction. Using Cohen's kappa as a means to develop measurement tools and determine what constructs ought to be measured rather than simply as a test of two individuals agreeing when applying a particular set of instructions required performing several inter-rater agreement tests between coders. Further, the coders might not use the same wording (i.e. tag) for a given construct. This required the creation of a program to measure the high points of coincidence in separate coders classification of a data set, and then to calculate Cohen's kappa for agreement between those coders. This process is described in greater detail in section 5.7.

Finally, the last major contribution of this work is that it will make more accurate affect detection possible. By incorporating cognitive attributions, hopefully we will be able to more

easily link students' self-reported emotions to their actions using these attributions as a proxy.

Knowing why a student believes they feel a particular way should have some association to events we can observe in the learning environment, perhaps more so than that student's emotional self-reports on their own.

5 Methods

5.1 Participants

The first study involved 85 eighth grade students from a central Massachusetts middle school. In order to protect the anonymity of this particular school, school demographic data below (Table 1) was rounded to the nearest quartile. Students at this school outperformed the average schools within the same municipality, but performed below the state average on the Composite Performance Indices for English Language Arts, Mathematics, and Science.

Table 1 Participant Summary

Variable	Hispanic	White	First Language Not English	English Language Learner	High Needs	Economically Disadvantaged
% Student Population	25%	50%	50%	25%	50%	50%

These studies were conducted with a single teacher, who taught 3 separate periods of mathematics.

5.2 Procedure

As MathSpring covers a variety of topics aligned to the Massachusetts state standards for eighth grade mathematics, the study was performed throughout the school year in tandem with units students were working on in class for a total of 7 days within a 6-month period. On each of these days students spent their entire period of Math class working with MathSpring. During day 1, students completed a brief pretest that included both mathematics content as well as affective and goal orientation survey measures (see section 5.4) and began working on MathSpring immediately afterward. The following days were spent working within MathSpring. On the final 1-2 days MathSpring experienced technical difficulties, thus we decided to provide an identical posttest to the pretest, this time on pencil and paper.

Each class of students worked with MathSpring as a class in a computer lab within their middle school. Both their teacher and the author were present during students' work. Students were discouraged from conversing with one another or using a calculator to solve problems and instead encouraged to do their work out on with pencil and paper when necessary.

5.3 Measures

5.3.1 Student Level Learning and Performance Measures

Students' achievement and learning (i.e. learning gains) were assessed at pretest and posttest time with items extracted from the Massachusetts Comprehensive Assessment System Standardized Test (MCAS) practice exams (see Appendix C). Additionally, measures of students' behavior within the tutoring environment (as described in the following section) were aggregated to the student level to provide an overall student level measurement of students' behavior and performance.

5.3.2 Fine Grained Learning & Performance Measures

As students worked within MathSpring, their performance and behaviors as they attempted each problem were logged. Students' interactions with the tutoring environment were tracked in the central relational database on the server where MathSpring runs, at UMASS Amherst (Table 2). Log data including each action and the time at which each action occurs (see table 2) were recorded and used as predictive measures in the fine-grained analyses covered later in the Results section.

Table 2 Student Actions & System Events

Action	Description
Hint	Student receives a hint for a given problem
Right	When a student solves a problem correctly
Wrong	When a student attempts to solve a problem but makes an incorrect attempt
Quit	When a student chooses to quit a particular problem and proceed to a new problem

Besides detailed per problem measures of performance and behavior (e.g. measures of engagement of a student with each individual math problem), selected problems were aggregated to form performance measures at the beginning and end of each problem set (i.e. topic, or knowledge unit) to form a within-tutor estimation of pretest performance and a within-tutor estimation of posttest problem solving performance. As a result students' improvement or growth could be tracked in addition to change as measured by external pre and posttest.

5.3.3 Aggregation

Aggregate variables at the student level were calculated in two ways: first, by a simple average of all of a particular student's responses for each measure (see table 3). Second, by the change in responses from the beginning of their work (e.g. pretest), and the end of their work (e.g. posttest).

Table 3 Breakdown of How Features are aggregated at the Student Level

Average vs Change	Pre & Posttest Measures	Within Tutor Measures
Student Level Average	Mean of Pre & Posttest Scores	Mean Performance within Tutor
Differential Across Time	Pre to Posttest Gain	Change in Performance throughout tutor use

5.4 Student Level Affective & Disposition Measures

5.4.1 Affective Measures

Students' affective predispositions were measured at pre and posttest time using previously validated items in table 4 (Arroyo et al, 2012). The validation process showed the items were tightly statistically related to Reinhard Pekrun's measures in the control value theory of emotion (Pekrun et al., 2016).

Table 4 Student Level Affective Measures Gathered via Pre & Posttest

Affective Measure	Appendix of Measure and Cited Work
Interest	Pre & Posttest Forced Choice of Interest (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Confidence	Pre & Posttest Forced Choice of Confidence (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Frustration	Pre & Posttest Forced Choice of Frustration (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Excitement	Pre & Posttest Forced Choice of Excitement (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Anger	Pre & Posttest Forced Choice of Anger (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Anxiety	Pre & Posttest Forced Choice of Anxiety (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Boredom	Pre & Posttest Forced Choice of Boredom (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Enjoyment	Pre & Posttest Forced Choice of Enjoyment (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Hopelessness	Pre & Posttest Forced Choice of Hopelessness (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)
Pride	Pre & Posttest Forced Choice of Pride (Appendix B) from (Arroyo et al, 2012; Pekrun et al., 2016)

5.4.2 Performance/Learning Measures

Learning, Performance, and Work Avoidance goals were measured through the 18 item GOALS-S survey (Dowson & McInerney, 2004). The difference between these affective survey measures at posttest time and pretest time were also computed.

Table 5 Student Level Learning & Performance Measures Gathered via Pre & Posttest

Performance or Goal Orientation	Appendix of Measure and Cited Work
Math Score	Students' scores on a Math content pre and/or posttest (Appendix C) Items from MCAS practice exams.
Mastery LO	Mastery Learning Orientation (Appendix A) from GOALS-S Survey (Dowson & McInerney, 2004)
Performance LO	Performance Learning Orientation (Appendix A) from GOALS-S Survey (Dowson & McInerney, 2004)
WorkAvoidance LO	Work Avoidance Learning Orientation (Appendix A) from GOALS-S Survey (Dowson & McInerney, 2004)

Initial analyses were conducted at the student level, involving many of the measures gathered during the pre and posttest, as described in tables 4 and 5 above. However, data was also collected within the tutoring environment, both through self-report (Table 6) and directly in the form of individual actions performed by students (Table 7). These fine-grained measures were aggregated by student to generate overall student measures for our initial student level analyses.

Table 6 Student Level Affective and Disposition Self-Reports Aggregated from within the MathSpring tutoring environment

Emotion, Attribution, or Agency Prompt	Description and Relevant Figure of Prompt
Forced Choice Confidence	Likert Scale Response on a Scale of 1-5 (5 being most) to the question of “Tell us about your level of Confidence in Solving math problems” as in figure 4
Forced Choice Excitement	Likert Scale Response on a Scale of 1-5 (5 being most) to the question of “Tell us about your level of Excitement in Solving math problems” as in figure 4
Forced Choice Frustration	Likert Scale Response on a Scale of 1-5 (5 being most) to the question of “Tell us about your level of Frustration in Solving math problems” as in figure 4
Forced Choice Interest	Likert Scale Response on a Scale of 1-5 (5 being most) to the question of “Tell us about your level of Interest in Solving math problems” as in figure 4
Forced Choice Attribution	Human Coded tags for the attribution question “Why is that?” regarding an emotion self-report as in figure 4
Open Response Emotion	Human Coded tags for the emotion question “How would you describe your emotions now (as opposed to the last time you were asked)? in figure 5
Open Response Attribution	Human Coded tags for the attribution question “Why do you feel that way?” regarding an Open Response Emotion self-report as in figure 5
Open Response Agency	Human Coded tags for the control/agency centered question “What do you wish you could do to improve this class right now?” as in figure 5

Table 7 Student Level Behaviors Aggregated from Actions within the MathSpring tutoring environment

Action	Description
Hints Per Problem	Number of times a student requested hints per problems attempted
Errors Per Problem	Number of incorrect attempts a student made per problems attempted
Quits per Problem	Number of times a student chose to end a problem without solving per problems attempted
Seconds per Problem	Seconds a student spent per problem, for all encountered problems

The measures of students' behaviors were selected to be as simple as possible, only measuring the frequency of a particular action against problems completed: hints, errors, and quits. These events encompass the most frequent types of actions students may take as they work within the software with the exclusion of a correct attempt or getting a problem right. This is because MathSpring is designed to allow students multiple attempts until they get a problem correct, as well as simply choosing to stop working on a problem and quit out to a new problem instead. So by controlling for both errors and quits we can see how thorough students' are as they progress through their work, as well as examine the possible causes of errors and quits. Students might quit problems because they are not challenging, or students may make errors as a part of their learning process. Finally, the measure of "seconds per problem" gives an overall idea of how quickly students are working, in addition to the other features we may be able to discern when extra time taken indicates a student being deliberate and thoughtful or alternately disengaged and off-task. Or whether a student is working quickly and competently or simply racing through problems with a combination of rapid guessing, hint abuse, or skipping.

Finally, students' responses to all pre test measures and pre to post change in terms of math skills, and goal orientation were compared across conditions forced-choice and open-response. Not every student completed all pre test and post test items due to leaving responses blank or transferring between schools or teachers. There were no significant differences between students in each condition according to these measures (see table 8).

Table 8 Results of t-tests and Descriptive Statistics for Math Pre to Posttest Gain, Pre to Posttest Change in Learning Orientation (LO), Pretest Emotion Survey Items, Pretest Learning Orientation (LO), and Math Pretest Scores by Sample Group (if both pre & post complete)

Outcome	Group						95% CI for Mean Difference	t	df
	Forced-Choice Self-Report			Open-Response Self-Report					
	M	SD	n	M	SD	n			
Math Pre to Post	0.18	0.23	37	0.13	0.18	37	-0.04, 0.15	1.14	72
Pre to Post	-0.23	0.78	40	-0.06	0.52	38	-0.47, 0.13	-1.14	76
Pre to Post	-0.2	0.67	39	-0.13	0.86	40	-0.41, 0.28	-0.4	77
Pre to Post Work	-0.06	0.74	40	0.01	0.53	39	-0.36, 0.22	-0.49	77
Pre Interest	2.93	1.09	42	2.76	0.94	41	-0.27, 0.62	0.77	81
Pre Confidence	4.05	0.97	41	3.93	1.18	42	-0.35, 0.59	0.51	81
Pre Frustration	2.86	1.16	42	2.88	1.23	42	-0.54, 0.5	-0.09	82
Pre Excitement	2.1	1.11	41	2.08	0.92	40	-0.43, 0.47	0.1	79
Pre Anger	2.24	1.21	42	2.63	1.22	41	-0.93, 0.13	-1.49	81
Pre Anxiety	2.21	1.09	42	2.26	1.06	42	-0.52, 0.42	-0.2	82
Pre Shame	1.9	1.1	42	2.1	1.09	41	-0.67, 0.29	-0.8	81
Pre Boredom	3.19	1.25	42	2.95	1.22	40	-0.3, 0.78	0.88	80
Pre Enjoyment	2.21	1.18	42	2.38	0.94	42	-0.63, 0.3	-0.72	82
Pre Hopelessness	2.07	1.2	42	2.02	1.14	42	-0.46, 0.55	0.19	82
Pre Pride	3.48	1.19	42	3.49	1.36	41	-0.57, 0.55	-0.04	81
Pre Mastery LO	4.18	0.58	41	3.98	0.62	40	-0.07, 0.46	1.49	79
Pre Performance	2.95	1.02	40	2.74	0.8	42	-0.2, 0.61	1.02	80
Pre Work	2.25	0.78	41	2.26	0.81	41	-0.36, 0.34	-0.06	80
Math Pretest	0.21	0.15	38	0.21	0.17	38	-0.08, 0.06	-0.24	74

* $p < .05$.

5.4.3 Fine Grained Affective & Disposition Measures

Students were randomly assigned to the “open-response emotion self-report” (Figure 5 below) condition or the “forced-choice emotion self-report condition” (Figure 4 below). Self-reports were requested roughly every 5 minutes without interrupting students during their work in a particular problem.

Please tell us how you are feeling.
Based on the last few problems tell us about your level of
Confidence in solving math problems

- Not at all Confident
- A little Confident
- Somewhat Confident
- Quite a bit Confident
- Extremely Confident

Why is that?

OK

Figure 4 Forced-Choice Self-Report Prompt that students encountered in MathSpring every 5 minutes or 8 problems

Students reported their emotional state given a set of possible choices (see Figure 5). These measures served as a main control to determine whether open-response self-report measures might reveal something more informative than our traditional closed-response measures. If students are asked to report on their emotions with minimal instruction (i.e. without being asked about a particular emotion) perhaps they will report different emotions than the ones we might initially expect. By associating this emotional report with cognitive attributions we may get a clearer idea as to why students believe they feel particular way, hopefully this information can help inform pedagogy in addressing students' emotional needs during a learning task. It's important to acknowledge here that open response self-report measures may bias in favor of students who are willing and able to recognize and articulate their current emotional state and attributions of their emotion's cause. Self-report measures in general face issues of validity in terms of students' understanding of their own emotions (Porayska-Pomsta et al., 2013; Ocumpaugh et al., 2015). However, I am primarily concerned with how closely they different types of reports may be associated with student behaviors and student experiences within a

tutoring environment. While this may not speak directly to questions of validity of these reports as emotions in and of themselves, it does address how these reports relate to empirically observable events during tutoring. Although it should be noted that this relationship with student action may be explained in part by students' act of reporting: the act of measuring students perceived emotions and attributions in this way likely influences the emotions and attributions being measured. While the results of these were examined at a fine-grained level of analysis, they were also aggregated at the student level.

We will ask these questions a **few** times, so its **OK** to change your mind. Please be as **honest** as possible in answering these questions.

1. How would you describe your emotions right now (as compared to the last time you were asked)?
2. Why do you feel that way?
3. What do you wish you could do to improve this class right now?

OK

Figure 5 Open-response Self-Report Prompt that students encountered in MathSpring every 5 minutes or 8 problems

Asking open-responses to students for the reasons attributed to their emotions ('And Why is that?') had already been investigated in our prior work, used in the past by the author to measure those students judgments, yielding a publication (Schultz et al., 2016). However, Figure 5 extends that idea to include an open-ended response measure of their emotion ('How would you

describe your emotions right now?’), followed by the same sort of causal attribution as before (‘Why do you feel that way?’) asking about the reason for the emotion, and finally ending with an open-ended question meant to allow input regarding students’ prospective beliefs of control and value (“What do you wish you could do to improve this class right now?”). While these measures are applied here at the fine-grained level of analysis they are also aggregated to the student level after human coders have summarized them with simple tags.

5.5 Open-Response Coding Protocol

Human coders carried out the process of coding the open-responses in Figure 5, as well as the attribution open-response question in Figure 4, in the three-step coding process described next.

Initially, coders were given students’ responses to the above prompts with associated contextual information. The contextual information included: a) responses sequenced by individual student and time of day; b) time of day was supplied to the coder to illustrate the time between responses; and c) students’ responses to all questions students were asked in the prompt. This was important as, in many cases, students’ responses in one particular prompt were given in reference to their responses in another previous prompt. Coders were given minimal instruction regarding how to code students’ responses, specifically no intended labels for how to code each response were given to coders. Instead coders were instructed only to “take students descriptions of their feelings and beliefs, and assign sets of one word tags (like hash tags) to those descriptions”, and were provided an example of the behavior using descriptive tags of popular films rather than students.

In order to avoid the risk of coders creating a very large and highly specific set of tags targeted at each individual student response guidelines were given in relation to the total number of tags coders should create for each given prompt. Specifically, they were told that their “total

list of tags for each question should be in the single digits 1-9 with an absolute maximum of 12, if necessary.” Coders were instructed to come up with a distinct tag list for each prompt. A complete “Coder Release Form” including protocol instructions as well as some brief survey questions for each coder is included in Appendix D.

In the second step, coders’ responses were examined and compared in order to find commonalities between coders’ individually devised tags. An open discussion was held in order to reach a consensus between coders, so as to reach agreement on which codes to use and whether certain researchers’ codes were actually equivalent (e.g. “irritated” vs “annoyed”). The goal of this conversation was to reach an agreed upon set of codes that would bias in favor of students’ intended meaning over any particular coder’s personal interpretation. Given that several coders were enlisted from different backgrounds, it was possible to come up with a set of codes that were agreed upon by coders who independently arrived at a similar coding scheme. In other words, if it had been that teachers viewed students’ intended meaning differently from educational psychologists, the method would allow to maintain the different interpretations as different tags that would preserve any potentially different meanings. Details of the meanings of the finalized tags are covered in sections 6.1. The coders ($N = 7$), were predominantly women with the exception of a single man. Coders were recruited from colleagues and classmates on a volunteer basis using a coder release form (Appendix D). A full transcript of discussions between coders is included as well (Appendix G). In addition to the challenges in determining a coding scheme described in section 5.6, there was also the challenge of acting as both a coder and facilitator of discussion and negotiation between coders. Practical concerns like limiting the total number of tags and ensuring that tags were sufficiently common (i.e. applicable in more than a handful of cases) were helpful in negotiations. Despite the instructions included with the

coder release form (Appendix D), some coders disregarded the guidelines of limiting themselves to approximately 10 tags per prompt; some coders used as many as 26 possible tags for a given prompt, many of which were excluded for either being infrequent or redundant with a very similar tag.

Finally, in the third step, coders were asked to apply this new consensus set of tags to students' responses. It was at this point that inter-rater reliability metrics between coders were calculated, as described in the next sections.

5.6 Open Response Coding Details

The second step of discussing and reviewing coders' responses was a highly involved and partly qualitative process. The task of assigning meaning to students' self-reports required attention to two main design goals:

- 1) **Specificity** – A “tag” should be as specific and semantically similar to a given student's self-report as possible. Further, semantic similarity was meant to include minimal “editorializing” on the part of the coder. For example, while “boredom” can be described as a combination of negative valence and low activation (Russell, 2003; Baker et al, 2010); students are unlikely to describe it as such. For example, rather than addressing activation directly a student might simply describe their feelings as “annoyed, bored”. The tags are meant to describe students' responses as semantically close and as specific to the responses themselves as possible.
- 2) **Generalizability** – Only tags that may be broadly applied to many students' self-reports should be used. This design goal may be in opposition to the first goal of Specificity. However, as my goal is also to find general patterns in students' self-reports and behaviors, the most common student sentiments are most likely to meet both goals.

The Coder Release Form (Appendix D) was distributed to several individuals, six of which responded and developed their own distinct coding schemes. In addition to the author, this made a total of seven. While the Coder Release Form included several instructions including that coders' "total list of tags for each question should be in the single digits 1-9 with an absolute maximum of 12." Some coders disregarded these instructions. Some coding lexicons exceeded 15 or even 20, with the largest lexicon of tags containing 46 distinct codes. Each coder's full lexicon of all tags used as well as the total instances of tags used for each self-report prompt are compiled in Appendix F.

The corpus of coders' tags for all students' responses of each prompt was processed into an initial "Coder Discussion Document" (Appendix G). This document meant to find highly correlated (coincident) tags between coders, as well as search out for codes that had similar semantic meaning. This process was done via manually searching through the corpus of coders' responses: for example one coder might apply a tag "boring" whereas another coder might use a tag "disinterested". Sets of tags with a high number of agreements that also shared some semantic meaning were paired together.

We should note that the "Coder Discussion Document" was not produced through an exact procedure due to the qualitative nature of finding similar semantic meaning, and somewhat error prone due to manually searching for coincident tags. There were several factors which might have affected the possible errors in compiling that document: spelling errors in individuals' tags, the fact that multiple tags could be applied to a single student response complicating the matrix, and the question of whether or not to pair coders' tags with replacement ("with replacement" is a term from combinatorics which means that after an item is selected from a pool it is available for selection again rather than being removed from the pool). For

instance, imagine an instance where coder A might have two tags: “frustration” and “negative valence”, whereas coder B might have “frustration” and “boredom”. It is possible that coder A’s “negative valence” tag could be more correlated with coder B’s “frustration” than coder A’s tag of “frustration”. The exact weight of semantic similarity between coder A’s “frustration” and coder B’s “frustration” must be weighed against the slightly larger R in correlation between coder A’s “negative valence” and coder B’s “frustration”. Further, coder A’s “negative valence” might be highly correlated with several of coder B’s tags, such as the aforementioned “boredom”.

Because both the qualitative human judgment aspect of this work and the quantitative process of determining coder agreement were both given high priority in this coding scheme, it was decided that a Python program should be authored to determine the degree of agreement between all coders.

5.7 Multi-Coder Inter-rater Agreement Program

A computer program (Appendix H) was designed to calculate the largest Cohen’s kappa values between all pairs of coders. Essentially, this program created a two-dimensional confusion matrix using each pair of coders’ tag lexicon of available tags for each of the axes. Then the program searched and identified the maximum value of agreements in this matrix, calculated Cohen’s kappa for those values, updated each coders’ tag lexicon and confusion matrix by removing the tags identified in the prior step, and searched the newly produced confusion matrix for the new maximum value. The process was repeated for as long as the shorter lexicon still had entries. An example is given in Table 9, and described next.

Table 9 Example of Kappa Program Confusion Matrix

	Amused	Angry	Annoyed	Confused	Disinterested	Optimistic	Sad	Satisfactory
annoyed	2	22	30		25	2	2	10
anxious			1			1	5	
bored	2		3	1	80	3	2	28
confused			1	8	4		1	2
depressed			1		5		4	1
good				1	7	60		6
null					8			
ok				1	7	5		72

For example, in the confusion matrix shown in Table 9, Coder N’s tags are on each row, while Coder C’s tags are on each column. The largest number of agreements ($N = 80$) happened between Coder N’s tag “bored” and Coder C’s tag “Disinterested”. Then Cohen’s Kappa is calculated for this value using the number of agreements (80), the sum of all elements in Table 9 ($N = 413$), the number of times Coder N used the “bored” tag ($N = 119$), and a value for Coder C using the “Disinterested” tag ($N = 136$). It is important to distinguish this may exceed the number of times Coder C used the “Disinterested” tag --as Coder N may have applied multiple tags to a self-report. In those cases “Disinterested” is counted for each tag applied, so a single tag instance that might have been “bored” + “depressed” would be applied to each of those tag categories.

After Kappa for “bored” vs “Disinterested” is calculated, the tags of “bored” and “Disinterested” are removed from the possible selections for agreement, and the cycle iterates again, this time selecting “ok” vs “Satisfactory”. It is important to note that this algorithm

calculates particular Cohen's Kappas: it searches for the highest number of agreements between two coder's tags and then excludes those tags from further selections.

This method does not guarantee the highest Kappas for each selection, as it is possible that something with relatively few agreements, but even fewer disagreements and therefore a more "pure" tag, would generate higher Kappas. However, I stand by this particular method as it weights selections based on high prevalence in both coders' data sets, biasing against infrequently applied tags.

Further, it may be possible that one coder's single tag might be best applied to two of another coder's tags rather than excluded after a single use. For example, one coder's tag of "material" could encompass another coder's two tags of "math" and "fractions". There are several cases where the meaning of a particular tag may straddle across the meaning of two of another coder's tags. Yet the author again stands by this method, as it is generating the **closest** agreement between each pairing of tags.

Unfortunately, this program was not authored until after discussions between coders regarding the overall coding scheme. This was due in part to the amount of time required to author this program, as well as a desire to have coders discuss a unified coding scheme soon after coding while their tags were still fresh in their minds. After the program was completed, recordings of the coder discussions were revisited, transcribed, and compared against the program's output.

The Finalized Coding Scheme is presented at the beginning of the next Results Section.

6 Results

6.1 Finalized Coding Scheme

The final tags were determined through a combination of the final tables (see Appendix F). Each table in the Appendix (F) corresponds to a different question and prompt asked to the student.

Not every tag was applicable to every type of prompt: note in Table 10 that some tags were applicable for the feelings prompts, some were applicable for attribution prompts, and some were applicable for “agency” prompts where students were asked how the system could be improved (see Figure 5). After this coding scheme was finalized the two final coders were given a set of instructions (see Appendix E).

Table 10 Tags used in Finalized Coding Scheme

Feelings Tags	Attribution Tags	“Agency” Tags
bored	bored	
DTG	DTG	DTG
	easy	easy
	growth	growth
confused	hard	hard
IDK	IDK	IDK
	material	material
	needs	needs
negative	negative	
positive	positive	
	success	
	website	
annoyed		
neutral		
	failure	
		bugs
		design
		fun
		quit

We use the term *Forced Choice Attribution* to indicate the answers to the question “Why is that?” regarding a student’s self-report of their feelings using Likert scale (see Figure 4).

For open response, the following names are used from now on:

We use *Open Response Feeling* to refer to answers to the question “How would you describe your emotions right now (as compared to the last time you were asked)?” in the open response prompt (see Figure 5). We use the term *Open Response Attribution*, to refer to answers to the question “Why do you feel that way?” in the open response prompt, also in Figure 5). Last, we use the term *Open Response Agency* to refer to the question “What do you wish you could do to improve this class right now?”, also in Figure 5).

In cases where Cohen’s Kappa between two coders was greater than or equal to 0.4, the value was highlighted in bold. Then, tags that were already bolded and seen as semantically similar were highlighted with the same color. Some tags had the same semantic meaning across multiple prompts, for example “bored” came up very frequently in multiple prompts. In those cases, the same color highlighting scheme was preserved across prompt/table in Appendix F. For cases where a tag was unique to a single prompt, the tag was left unhighlighted.

What follows next is a discussion of the rationale behind each tag in the finalized tag list, and it heavily references Appendix F as well as the transcripts of discussions between coders (Appendix G).

6.1.1 Tag Descriptions

The following is a list of the tags used, and a description of how the author came about with the decision of each of them.

bored – “bored” was a fairly common and self-explanatory tag used by coders. In those cases, students would discuss feeling bored, or refer to states of being that could be described as boredom or disinterest. Every coder had a tag that roughly reflected boredom with the exception of Coder T who only coded a portion of the data set. Table 11 shows the level of agreement between coders in **“bored” is highlighted in dark red in Appendix F.**

Table 11 Coder Tags Related to the Code finalized as ‘Bored’, for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	Disinterested	boring	bored	#I’m bored	boring	bored	negative engagement
Forced N	57	19	31	27	29	31	5
Feeling Tag	Disinterested	bored	bored	#bored	Deactivating		boredtiredmeh
Feeling N	116	35	119	30	39		12
Attribution Tag		Boring		#I’m	boring	boredom	Bored
Attribution N		8		17	11	18	1

Apart from Coder T’s and Coder C’s codes, all coders above tags achieved a kappa of 0.5 or higher with one another. One possible reason for disagreement with Coder C’s codes was that “Disinterested” was used for situations that could have been tagged as “bored” as well as “IDK” or “DTG” (Gobert et al., 2015). While it’s plausible that students may respond with “I don’t know” or self-report disengagement as a means of expressing boredom, this last inference assumes there are deeper and further reasons and causes behind students’ responses than answering the question as posed. That is not necessarily true and goes beyond the scope of this dissertation, which already is pursuing deeper into students’ causes for their answers than usual in the scientific community of affect detection and assessment.

IDK – “IDK” is an abbreviation of “I don’t know”. It is meant to identify cases where students claimed they didn’t know why they felt a particular way; sometimes students would simply

answer “nothing” or “because I do” when asked why they feel a particular way (see Table 12).

“IDK” is highlighted in orange in Appendix F.

Table 12 Tags Related to the Code finalized as “IDK”, for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	disinterested	IDK	IDK	#IDK	idk	Does not know	idk
Forced N	57	59	46	11	10	5	3
Feeling Tag	disinterested	IDK		#IDK	idk	blank	idksilly
Feeling N	116	68		9	10	44	16
Attribution Tag	avoidance	IDK		#IDK	idk		Idk
Attribution N	146	37		27	45		11
Agency Tag	unsure		Idk	#idk	Idk		Idk
Agency N	37		153	35	37		7

IDK is one of the many tags that were used in prior work (Schultz et al. 2016). However, it is important to note that, in that prior work, “IDK” was used as a catch-all tag that could also include responses that would currently be tagged as “Disengaged from Task Goal”, which we refer as DTG from now on.

DTG vs. needs – “DTG” or “Disengaged from Task Goal”. In prior work (Gobert et al., 2015; Schultz et al. 2016), this construct typically means that students are engaging in a task in a way that is not related to the goal of the intended goal of the task. In this context the students’ reports illustrate a focus on something unrelated to working within MathSpring. These responses often seem absurd, for example responding with “eating chicken”, “ya”, “swagger”, or “cats”.

Distinguishing between “DTG” and “IDK” was a common element in the group discussions about coding (as found in Appendix G):

“I used to code IDK being like “Nonsense” or “Uninterpretable” but it’s a little bit different. IDK can mean “I don’t know why I feel that way” you can also have a student typing like “bbbbbbbbbbbbbbbbbb” or just a nonsense set of text that doesn’t seem to be made to communicate something”

“Coder D: I just, I sort of indicated it in one place but my “IDK” straight is “I don’t know”, IDK with a question mark which is like random text which is off task.”

“DTG” is highlighted in dark brown in Appendix F

Table 13 Coder Tags Related to the Code finalized as “DTG”, for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	disinterested	IDK	IDK	#notrelevant	nonsense	blank	xxx
Forced N	57	59	46	46	43	44	162
Feeling Tag	disinterested	idk		#notrelevant	nonsense	blank	Idksilly
Feeling N	116	68		73	22	62	16
Attribution Tag	environment	idk?	Out	#notrelevant	nonsense	unrelated to system	outside influence
Attribution N	4	14	14	83	13	10	3
Agency Tag	personal	idk?	basic needs	#notrelevant	nonsense	sustenance	idk
Agency N	72	13	5	55	16	4	7

Another that should be distinguished from “DTG” or “IDK” is “needs”. The “needs” tag refers to students asking for accommodations to allow them to perform a task. For example, a student might attribute their emotions to: “it’s first period” or “because i haven't eaten anything”. In these cases, students can complain about the amount of heat in the classroom, the fact that they feel tired or thirsty or hungry. Students’ basic needs such as hunger (Kleinman et al., 2002; Adolphus et al., 2013) have been shown to negatively impact student performance when not met/ This includes elements of home life such as laundry and an emotional support structure (Carney-

Crompton & Tan, 2002). These responses are only tangentially related to the learning task, however they are not the absurd self-reports found in DTG (see Table 13). “needs” is

highlighted in light brown in Appendix F.

pos – Stands for “positive”. This tag was used to indicate a positive valence. While the tag was relatively simple, it could be used as a modifier in conjunction with other tags (see Table 14).

“pos” or “positive” is highlighted in light green in Appendix F.

Table 14 Coder Tags Related to the Code finalized as Positive for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	Good	positive	fun	#MathisFun	positive	fun	positive engagement
Forced N	28	57	11	12	15	17	2
Feeling Tag	optimistic	positive	good	#happy	positive	positive	goodawakebetter
Feeling N	66	171	74	21	100	73	31
Attribution Tag	good	positive				experience is positive	
Attribution N	34	69				42	

neg – Stands for “negative”. This tag was used to indicate a negative valence much like the previously mentioned positive tag. Again while this tag could simply mean that a student was “unhappy”, it could also be used as a modifier in conjunction with other tags (see Table 15).

“neg” or “negative” is highlighted in yellow in Appendix F.

Table 15 Coder Tags Related to the Code finalized as Negative for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag		negative	hate		negative	frustration	
Forced N		70	25		38	23	
Feeling Tag		negative			negative	frustration	annoyed confused not ok
Feeling N		158			113	77	26
Attribution Tag	frustration	negative			negative		doesn't like task
Attribution N	40	91			27		8

easy – Easy referred to instances where students described a low difficulty level. This label can be used in combination with “pos” or “neg” to in cases where students either like or dislike the fact that they find the material to be “easy” (see Table 16). For example, one student explained why they felt bored with the single word response “Unchallenged”... this would be a case of “easy” + “neg”. A case of “easy” + “pos” would be when a student reports feeling calm “because i know how to do this”. “easy” is highlighted in dark blue in Appendix F.

Table 16 Coder Tags Related to the Code finalized as Easy for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	Not challenging	Easy	easy	#tooeasy	easy	Needs challenge	understanding
Forced N	59	28	15	30	36	29	15
Attribution Tag	Not challenging	easy	easy	#tooeasy	easy		Too easy
Attribution N	45	18	11	22	19		2
Agency Tag			hints	#hints	hints		easier
Agency N			5	4	6		2

hard/confused – Hard referred to instances where students described a high difficulty level, a high level of challenge. Because some coders tended to refer to this as the student being confused, while others referred to the level of challenge expressed in the response, the decision

was made to call this tag ‘hard/confused’ (see Table 17). This tag operates the same way as “easy”. “hard/confused” can also be used in combination with “pos” or “neg”. For example “hard/confused” + “neg” could be used to describe an instance where a student described feeling annoyed “because i am not able to understand the problem”.

“hard” and “confused” are highlighted in gray blue in Appendix F.

Table 17 Coder Tags Related to the Code finalized as ‘Hard/Confusing’ for Forced Choice Attribution, Open Response Feeling, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	confusion	hard	confused	I don’t understand	hard	math	content
Forced N	3	7	15	27	10	34	3
Feeling Tag	confused	confused	confused	#confused	confused		
Feeling N	8	13	16	8	7		
Attribution Tag	confusion	unsure	too hard	#I don’t understand	lack of proficiency	math	stuck
Attribution N	9	14	15	15	8	41	10
Agency Tag				#morechallenges		more challenges	
Agency N				9		17	

While “hard” and “confused” could be used interchangeably, it bears mentioning that “confused” shows up when students are asked about how they feel, while “hard” occurs when students are discussing their attributions for why they feel a particular way. In this sense it makes sense to have a distinct “confused” tag for feelings, even if challenging “hard” material causes students to feel “confused”.

material – This tag refers to mathematics content. It could refer to “mathematics” in general, or a specific unit such as “fractions & decimals” (see Table 180). It’s distinct from “easy” or “hard” because some student may claim to dislike math regardless of difficulty or the way it’s presented

in MathSpring, this would be represented by “material” + “neg”. An instance of this would be a case where a student said they felt “terrrrrrrrrrribleeeeeee” and then explained that the reason was “cause of the inventor of fractions”.

“material” is highlighted in light blue in Appendix F.

Table 18 Coder Tags Related to the Code finalized as ‘Material’, for Forced Choice Attribution, Open Response Attribution, and Open Response Agency prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag			dislike math	#I don’t like math	domain	math	content
Forced N			10	18	18	34	3
Attribution Tag			math	#I don’t like math	domain	math	doesn't like math
Attribution N			35	25	24	41	12
Agency Tag	content	content	content	#difficultylevel	questions	more challenges	content
Agency N	40	7	34	12	36	17	3

success – This tag corresponds to students describing they are doing well or answering several questions correctly (see Table 19). This can be related to “easy”, but not necessarily. Sometimes students don't mention difficulty in their responses (e.g. “i got the problem right”), or report even feeling pride at being successful despite adversity.

“success” is highlighted in dark green in Appendix F.

Table 19 Coder Tags Related to the Code finalized as ‘Success’, for Forced Choice Attribution, and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	Good	Positive	success	#I’m good at math	successful	success	successful
Forced N	28	57	26	23	16	33	2
Feeling Tag	good	positive	success	#I’m good at math	successful	confidence	successful
Feeling N	34	69	33	9	22	10	14

Growth – This tag was related to students attributing their feelings to personal improvement or learning (e.g. “I feel like i’m learning new stuff”). This can also be used in the context of the agency prompt when students take responsibility for improving their interactions with MathSpring themselves by learning more or working harder (when given the “Agency” prompt students might respond with “study” or “work harder”) as displayed in Table 20.

“growth” is highlighted in pink in Appendix F.

Table 20 Coder Tags Related to the Code finalized as Growth, for Open Response Attribution, and Open Response Agency prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Attribution Tag		learning	learning	#i'mlearning	improvement		
Attribution N		2	14	10	7		
Agency Tag		study		#improve myself		change self	
Agency N		2		15		8	

website – This tag was coded when a student referenced the MathSpring website itself. Again these references can be positive or negative. For example an instance of “website” + “neg” would be “these problems are hard to read and i keep getting the same problem over and over” (see Table 21). “website” is highlighted in purple in Appendix F.

Table 21 Coder Tags Related to the Code finalized as Website, for Forced Choice Attribution and Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Forced Tag	Website		software	#Website	problem with system	system is repetitive	tech
Forced N	16		35	30	3	5	3
Attribution Tag	System issue	unsupportive	bugs	#website problems	problem with system	frustration with system	Tech issues
Attribution N	4	14	16	27	16	32	9

failure – When students are doing poorly and or feel they are failing. This can be related to “hard”, but not necessarily. Rather than assessing item difficulty, students may focus instead on their own ability level (see Table 22). For example “I’m not good at math”. Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way.

Table 22 Coder Tags Related to the Code finalized as Failure, for Open Response Attribution prompts

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Attributions Tag			failure	#low achievement	unsuccessful		stuck
Attributions N			12	6	8		10

annoyed – refers to when students describe a feeling of annoyance. Many coders specifically used variations of the phrase “annoyed” as opposed to frustration (see Table 23). Annoyance may also imply a distinction in where the student places themselves in import in relation to the learning environment. Frustration implies negative affect due to a lack of one’s own ability to affect a change in one’s environment, while annoyance implies a negative affect due to an unimportant or trivial element of one’s environment. Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way.

Table 23 Coder Tags Related to the Code finalized as annoyance, for Open Response Feeling prompt

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Feelings Tag	annoyed	annoyed	annoyed	#annoyed	negative	frustration	Annoyed confused Not ok
Feelings N	31	21	93	20	113	77	26

neutral – when students describe their feelings as neither positive nor negative, simply “fine” or “ok” (see Table 24). Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way.

Table 24 Coder Tags Related to the Code finalized as neutral, for Open Response Feeling prompt

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
Feelings Tag	satisfactory	calm	ok	#content	neutral	neutral	Calm fine ok
Feelings N	91	5	85	113	73	123	57

bugs – refers to when students identify some sort of error in the system. These are not intentional design features, but rather issues like receiving the same problem repeatedly. Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way (see Table 25).

Table 25 Coder Tags Related to the Code finalized as bugs, for Open Response Agency prompt

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
“Agency” Tag	error		bugs	#debugit	system		
“Agency” N	5		7	10	13		

design – refers to criticisms or suggestions about design improvements. These design elements are largely aesthetic about layout, color, sound, or the way the learning companions talk to students. Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way (see Table 26).

Table 26 Coder Tags Related to the Code finalized as design, for Open Response Agency prompt

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
“Agency” Tag	structure	color	design	#aesthetics	display	aesthetics	more engaging
“Agency” N	33	4	23	8	10	20	3

fun – refers to requests that MathSpring be more fun or include more game-like elements. Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way (see Table 27).

Table 27 Coder Tags Related to the Code finalized as fun, for Open Response Agency prompt

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
“Agency” Tag		fun	more fun	#more fun	more fun	fun	more fun
“Agency” N		3	17	11	15	12	1

quit – refers to requesting to quit or leave the learning task, may refer to quitting work within MathSpring, quitting math class, or leaving school entirely. Notice that the tags hereafter including “failure”, “annoyed”, “neutral”, “bugs”, “design”, “fun”, and “quit” are not highlighted in color as these tags are only applicable to a single prompt, e.g. only in response to how a student is feeling, or only in response to why they feel that way (see Table 28).

Table 28 Coder Tags Related to the Code finalized as quit, for Open Response Agency prompt

Coder	Coder C	Coder D	Coder N	Coder R	Coder S	Coder SH	Coder T
“Agency” Tag		leave	disengage	#not relevant	leave	quit	
“Agency” N		7	25	55	13	14	

6.2 Inter-rater Reliability for the Finalized Scheme

The finalized coding scheme was applied by both Coder N and Coder S. I discovered that the previously described inter-rater agreement program actually biases against applying multiple tags for a given self-report. If the codes “easy” and “negative” are applied to a report by both coders, there will be an agreement added for “easy” and “negative”. However, there will also be a disagreement added for “easy” and “negative” as the program will consider the case where one coder used the tag “easy” while the other coder used the tag “negative”. Despite this bias that generates conservatively low Kappas as applied in the prior section, I maintain that the primary purpose of the inter-rater agreement program was to illustrate that these codes are viable as evidenced by relatively high Kappas (>0.4), when coders are given no specifically defined tags. The fact that the program penalizes instances where multiple tags are applied means that in many cases the actual Kappas in the prior sections would be higher than reported. Unfortunately, it does also mean that Kappas previously reported are lower for cases where multiple tags are applied, so these tags are less likely to be included in the final scheme. Table 29 lists the final Kappa values for the final pair of coders that re-coded according to the finalized list of codes presented in the previous section.

Table 29 List of Tags, Total Instances, & Kappas for re-coding by final coders: Coder N and Coder S

Forced Choice											
Attributions			Open Feelings			Open Attributions			Open Agency		
Tags	N	Kappa	Tags	N	Kappa	Tags	N	Kappa	Tags	N	Kappa
bored	23	0.87	annoyed	39	0.70	bored	21	0.82	bugs	6	0.92
DTG	29	0.88	bored	28	0.98	DTG	36	0.60	design	34	0.77
easy	36	0.95	confused	11	0.84	easy	32	0.76	DTG	26	0.76
failure	1	1.00	DTG	27	0.74	failure	9	0.61	easy	9	0.87
growth	3	0.99	IDK	15	0.72	growth	11	0.79	fun	15	0.84
hard	16	0.72	neg	47	0.70	hard	14	0.69	growth	13	0.93
IDK	18	0.89	neutral	109	0.86	IDK	86	0.89	hard	8	1.00
material	26	0.80	pos	78	0.88	material	29	0.93	IDK	133	0.38
needs	2	0.66				needs	11	0.69	material	18	0.77
neg	77	0.57				neg	65	0.60	needs	7	0.83
pos	26	0.92				pos	7	0.35	quit	17	0.91
success	23	0.87				success	15	0.66			
website	36	0.76				website	33	0.77			

When the finalized coding scheme was applied only three tags fell short of the cut-off of Kappa = 0.6: for the forced choice attributions “neg” kappa = 0.57, for the open attributions “pos” kappa = 0.35, for the open “agency” “IDK” kappa = 0.38. The negative “neg” comes very close to satisfying the 0.6 cut-off threshold, so these disagreements were reviewed. In many cases, Coder N tagged a response as “neg” while Coder S did not. Often these were instances of “easy” & “neg” where students would say that problem were “too easy” or request harder work. Additionally, one student in particular described MathSpring as “Walmart ixl” and criticized MathSpring’s originality referring to us the designers as “stupid monkeys”. Given this, the “neg” tag was kept despite the few instances of disagreement.

6.3 Initial Student Level Analyses

As proposed, the analyses begin at the student level. As a reminder to the reader, the goal was to illustrate how tracking students’ appraisals of a situation may help explain students’ emotions as well as behaviors within a tutor-learning environment. Attribution and appraisal data, including students’ motivation and volition, may allow us to understand (and a machine-based tutor to

predict in the future as described in section 4 Research Goals) students' behaviors more accurately than a combination of pure affective assessments and behavior alone. First, we analyze the relationships between attributions and pre/posttest variables, an analysis at a high level using aggregate measures that vary student by student, to later proceed to analyze finer grain action sequences, in future sections. We thus started by framing these high level analyses with three further research questions:

Research Question #1: What are the emotions that students report in an open-response assessment, and do they match student emotions from the literature, and/or the ones that we asked in the forced-response condition? This is essentially a re-phrasing of the initial question from section 4 "Research Goals" which asks "What constructs ought to be considered?"

Research Question #2: Are the ways students feel in a learning environment predetermined by their general attitudes and goals and abilities that students bring to the learning environment? This addresses part of the general question in section 4 "Research Goals": "To what extent should these states be described as cognitive, affective, or epistemic, a combination of cognitive and affective?" as it relates cognition and attitudes to in tutor emotional states.

Research Question #3: How do students express their emotions in an online tutor and how are these emotions associated with students' behaviors in a digital learning environment? This repeats the question from section 4 "Research Goals" in greater detail "What are associations between student affect and attributions with student behavior.

Research Question #4: Why do students believe they feel a particular way? We investigate the causal attributions students assign to their emotional states. This is really quite similar to RQ 2

above except it addresses situational rather than trait factors. As a result it also relates more directly with RQ 3 which in how these attributions relate to behavior.

6.3.1 Descriptives for Students' Pretest, Dispositional, & Behavioral Measures

Students' behaviors were aggregated to an average per student; see Table 30.

Table 30 Aggregate Behavior Measures Considered for Analyses

Measure	Description
SOF	Solved on First Attempt per Problem
Wrong/Problem	Incorrect Attempts per Problem
Hints/Problem	Hints Requested per Problem
Time/Problem	Average Seconds spent per Problem

Some attributional tags were excluded from analyses due to either an insufficiently small sample size in the Forced-Choice condition (e.g., “Failure”, “Growth”, & “Needs”) or due to an insufficiently small Cohen’s Kappa in the Open-Response condition (e.g., “Positive”). These tags are separately analyzed later in Table 38: and Table 39.

For these student-level analyses the open-response and forced-choice self-reports were averaged for each student. For example, if a particular student gave 3 responses to the question of “how confident are you?” as their time in the tutoring session progressed, consisting of values 2, 3, & 5 (remember that 1=not at all confident and 5=extremely confident), then the average for that student would be $10/3 = 3.33$. In the case of open-response prompts the total number of times a student responded to an open response prompt was used. So if a student referred to the ‘material’ (math content) twice out of 50 responses then their average would be 0.04 for the ‘material’ coded tag.

Table 31 Summary of Emotion & Attribution Tags Used, Total Instances (N), & Cohen's Kappa of Interrater Reliability for Open-Response and Forced Conditions. Values in bold if N or Kappa are unacceptably low.

Tag	Example student response	N Open/ Forced	Kappa Open/ Forced
Emotions			
Annoyed	“this is getting annoying”	39/NA	0.70/NA
Bored	“my emotions are bored and tired”	28/NA	0.98/NA
Confused	“kind of confused but still happy”	11/NA	0.84/NA
DTG	“swagger”	27/NA	0.74/NA
IDK	“IDK”	15/NA	0.72/NA
Negative	“I'm still very stressed.”	47/NA	0.70/NA
Neutral	“ok i guess”	109/NA	0.86/NA
Positive	“good i get the hang of it and it gets easier.”	78/NA	0.88/NA
Confident	N/A	NA/40	NA/NA
Interested	N/A	NA/38	NA/NA
Frustrated	N/A	NA/39	NA/NA
Excited	N/A	NA/37	NA/NA
Attributions ('Why is that?')			
Boring	“very boring same problem”	21/23	0.82/0.87
DTG	“because im batman and awesome”	36/29	0.60/0.88
Easy	“They are not very hard”	32/36	0.76/0.95
Failure	“cause i got some questions wrong”	9/1	0.61/1
Growth	“because these questions are helping”	11/3	0.79/0.99
Hard	“ITS HARDDDD!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!”	14/16	0.69/0.72
IDK	“Because I do”	86/18	0.89/0.89
Material	“cause its math”	29/26	0.93/0.80
Needs	“i dont really know, im just hungry”	11/2	0.69/0.66
Negative	“not fun”	65/77	0.60/0.57
Positive	“THEY ARE FUN”	7/26	0.35/0.92
Success	“because, im getting the questions right .”	15/23	0.66/0.87
Website	“doing this on the computer is kind of weird”	33/36	0.77/0.76

As is apparent in Table 31, the two most prevalent emotional states reported were neutral (N=109) and positive (N=78), followed by the negative valence in general (N=47), annoyance (N=39), and boredom (N=28). Finally, confusion was the least common emotional state (N=11), This may be in part due to the fact that more students likely found their work easy as compared to difficult as evidenced by the fact that roughly twice as many students reported “easy” as compared to “hard” attributions (see table 32). Please note: tags of simple valence are more common than epistemic emotions. Specific contextual elements such as cognitive difficulties or

disengaging activities apply an added requirement beyond basic valence. Perhaps the increased specificity just makes them less likely to occur: these emotional states require particular valence as well as specific cognition or engagement/disengagement in the task. The control-value theory would similarly hold these states as dependent on students' assessment of their control over each specific situation, the degree they value an outcome, and whether the student's focus is on a prospective outcome or a retrospective outcome (Pekrun, 2006). These additional requirements could possibly render epistemic emotions less common than broad descriptions of valence. However, assessing these experiences may require increased metacognition or self-awareness (Bieg et al., 2014; Porayska-Pomsta et al; 2013), which students' may not possess as discussed in section 3.2.

A question of utmost importance then becomes whether these emotion tags are associated to student behaviors or other outcomes, and whether they might be better (or worse) predictors of behaviors than what MathSpring had in place before, as specified in Table 32, which had a very even number of responses for each emotion report. The first row which addresses "Total Self-Reports" tallies the number of self-reports each student made and then averages that total across the cohort. So students made an average of 13 self-reports, with a standard deviation of about 6 reports, so roughly 68% of students made between 7 and 19 self-reports.

Table 32 Likert Scale Self-Reports for Closed Response Condition

Measure	Mean	SD	Low Mean	High Mean	N (Students)
Total Self-Reports	13.21	5.82			42
Confident	3.31	1.2	0.33	0.52	40
Excited	1.98	1.08	0.73	0.16	37
Frustrated	2.73	1.5	0.57	0.41	39
Interested	1.91	1.08	0.70	0.12	38

6.3.2 RQ1: What are the emotions that students report in an open-response assessment, and do they match student emotions from the literature?

In the closed response condition the four pre-determined affect states were measured via a self-report Likert scale ranging from lowest (1) to highest (5). Generally, students reported a relatively high level of confidence, but low degrees of excitement or interest. Once again, this may be due to students' finding the material less challenging as acknowledged in in section 6.3.1. The Low and High mean values (Table 32) convert the continuous Likert scale measure to discrete present/absent measures comparable with open-response measures in Table 33. Out of the total responses of each type in Table 32 (e.g. Confidence), the average total responses that were either low (<3 on a Likert scale) or high (>3 on a Likert scale). Again, the top row "Total Self-Reports" is the average number of self-reports each student made. Notice that students were less likely to make open-response self-reports as opposed to forced-choice. Possibly, because simply clicking a multiple-choice option requires less effort than typing in a response in a text box. The following rows are the average number of reports of each type given by each student. If a hypothetical student reported "Bored" 5 times out of a hypothetical 9 self-reports their score would be 0.56. That score was then averaged with every other students' score to generate the Table 33.

Table 33 Descriptives for Emotion Tags for Open Response Condition

Measure	Mean	SD	N (Students)
Total Self-Reports	8.5	5.38	42
Annoyed	0.09	0.17	39
Bored	0.08	0.14	39
Confused	0.03	0.05	39
DTG	0.09	0.15	39
IDK	0.03	0.08	39
Negative	0.11	0.16	39
Neutral	0.31	0.33	39
Positive	0.23	0.28	39

The positive valence in open-response seems consistent with the higher reports of confidence in the forced-choice condition. With regard to negative emotions, students tended to use the term “annoyed” to describe their feelings, however this term may be used interchangeably with “frustrated” or dissatisfied (Baker et al., 2010). Finally, the fact that simple tags were so common suggests that students in the forced-choice condition may simply be using the Likert scale as a means of communicating valence rather than more subtle emotional states. Three of the students in the open response didn’t respond to any self-report prompts (see N in Table 33).

As a general response to the research question, the most frequent emotions that students report are feeling positive or negative in general (without much specificity), or simply feeling neutral. The emotions in the literature that were reported were only bored and confused, though they occurred less than 10% of the time. Also, the emotion of being ‘annoyed’ was reported 9% of the time, and may overlap with the construct of ‘frustration’; however, it is not necessarily the same. Finally, two other emotions were reported: ‘not knowing how they are feeling’ and being simply ‘disengaged from task goals’; each of these were again reported less than 10% of the time. Thus, the conclusion is that students only report a few emotions that have been investigated in the literature when asked openly.

This leads to the question of why students’ responses differ from the emotions which researchers would ordinarily ask about. There certainly has been doubt cast upon students’ self-reporting competency or self-awareness of their emotional states. For example, when asked to report on their emotional states at the end of a lesson had one of the lowest inter rater reliability scores as compared to other collection methods (D’Mello et al., 2008). Furthermore, when asked

to report on their emotional state and the end of a day students' reports differed significantly from their reports within a learning environment (Bieg et al. 2014). However, neither of these findings undermine students' self-reports within a learning environment. Further, neither compare students' ability to report their emotional states against adults' ability at the same task. However, there are findings that show that much younger students (i.e. ages 2 through 9 years) tend to use broader categories to identify emotions in facial expressions which narrow as they grow older (Widen & Russell 2003, 2008). For example a preschooler might use a term like "anger" for facial expressions meant to illustrate all negative valence emotions including "fear", "sadness" and "disgust" (Bullock & Russell 1984, 1985, 1986). While these prior works address the abilities of much younger students to identify emotions from facial expressions, it is notable that they too are applying emotional categories broadly on the basis of valence or activation. While the simplicity and breadth of middle school students' responses may just be accurate reports of their own experienced emotions, the similarity with reports of younger students here suggests these reports deserve a greater degree of scrutiny to determine if this is due self-reporting competency rather than an intentional choice to respond in terms of simple valence.

6.3.3 RQ2: Are the ways students feel in a learning environment predetermined by their general attitudes and goals and abilities that students bring to the learning environment?

Students' self-reported emotional state within the tutor was significantly correlated with the pre and posttest scores and survey measures in a few instances. The summary of these analyses can be found in Table 34. The following results were found:

- a) Work avoidance goals were negatively correlated with confidence ($R=-0.47$). This means that students who tend to report avoiding academic work also report feeling less confident while solving math problems within the MathSpring Tutoring environment, or

that students who report higher degrees of confidence within the tutor are more likely not to identify as having work avoidance goals. This seems consistent to other findings in the literature (Dowson & McInnery, 2004).

- b) Work avoidance goals were positively correlated with the student NOT answering the forced-choice response condition question about how they feel ($R=0.47$), and also with IDK in the open response condition ($R=0.34$). This means that students who tend to report work avoidance goals also avoid answering the forced-choice emotion questions. Interestingly, the pattern reverses when students are asked openly about how they feel ($R=-0.26$), but students with work avoidance goals apparently also tend to answer more “I don’t know” in the open-response condition ($R=0.34$). This may indicate that several students who left the forced-choice prompt blank may have wished to communicate that they simply did not know how they were feeling at the time rather than being unwilling to report their feelings.
- c) Work avoidance goals were marginally correlated with frustration ($R=0.27$). This means that students who tend to report work avoidance goals also tend to report feeling more frustrated.
- d) Interestingly, whether a student pursued ‘Mastery Goals’ was found to be negatively correlated with students’ open-response reports of feeling neutral ($R=-0.33$) and positive ($R=-0.40$). One possible explanation could be that students with high mastery goals might not have felt as though they were meeting their goals of personal growth, or given enough challenge, but this would have to be determined with further examination.

Students who responded to the emotion prompt with some variation of “I don’t know” tended to also reported ‘Work Avoidance’ goal orientation. This is in contrast to students who

responded to the question with a disengaged or unrelated utterance “DTG” who appeared less likely to report Work Avoidance goals. These two tags were initially described as similar so it’s notable that they appear to have opposite relationships with Work Avoidance goals.

Table 34 Emotions vs Pre/Posttest Measures: Bivariate Correlations

	Correlation Coefficients					N (Students)
	Math Pre/Post Gain	Mean Pre&Post Score	Mean Pre&Post Mastery Goals	Mean Pre&Post Performance Goals	Mean Pre&Post Work Avoid Goals	
Forced-choice Measures						
Confidence	0.16	0.40*	0.19	0.17	-0.47**	40
Excitement	-0.06	0.10	0.11	-0.18	-0.13	37
Frustration	0	-0.02	-0.11	0.06	0.27 [†]	39
Interest	0.04	0.19	-0.01	-0.11	-0.11	38
Blank	-0.03	0.08	-0.13	0.08	0.47**	42
Open-Response Measures						
Annoyed	0.04	-0.25	-0.18	0.07	0.19	39
Bored	-0.22	0.09	-0.12	-0.04	0	39
Confused	0.03	0	-0.11	0.04	0.15	39
DTG	0.18	0.26	-0.03	0.30 [†]	-0.35*	39
IDK	-0.03	-0.18	0.01	-0.06	0.34*	39
Negative	0.18	-0.05	0.09	-0.1	-0.1	39
Neutral	0.13	0.14	-0.33*	-0.08	0.09	39
Positive	-0.2	-0.16	-0.40*	-0.1	-0.09	39
Blank	-0.23	0.12	0.01	0.04	-0.26 [†]	42

†=p≤0.1, *=p≤0.05, **=p≤0.01 Italicization for findings significant under Benjaminni-Hochberg 25% false discovery rate

Interestingly, students’ emotions were unrelated to simple raw gains in a mathematics pre to posttest. Because the MathSpring sessions extended over several months in parallel to math class, it is unclear what this learning measure is actually measuring --clearly not only learning within the digital learning environment but also learning within the traditional math class led by the teacher. Another measure of learning solely inside of the tutor should be considered for a better analysis of how the emotions reported within the tutor might relate to the learning gained inside of the tutor.

6.3.4 RQ3: How do students express their emotions in an online tutor, and how are these emotions associated with students' behaviors in a digital learning environment?

A variety of behaviors inside of the tutor were considered, including: solving problems correctly on the first attempt, time spent in problems, incorrect answers per problem, help/hint requests in a problem, and giving up on problems after starting ('problems quit). Table 35 contains results that help to answer this research question.

In comparing open and closed measures of emotion we find that the forced question about students' confidence captures the most behaviors: performance measures of solving problems correctly on the first attempt, making few errors per problem, and increased use of hints. In many ways, confidence appears to be capturing a metacognitive awareness of doing well, as much as a positive predisposition of seeking for help/hints. A few marginally significant correlations highlight some trends: students who report being highly 'interested' marginally solve more problems correctly in the first attempt, and students who report higher levels of 'excitement' also tend to request more hints.

Table 35 Emotions vs Behaviors: Bivariate Correlations

	Correlation Coefficients						N (Students)
	SOF per Prob	Time per Prob	Wrong per Prob	Hint per Prob	Problems Quit	Avg Prob Difficulty	
Forced-choice Measures							
Confidence	<i>0.41</i> **	0.07	<i>-0.54</i> **	<i>0.33</i> *	0.13	0.08	40
Excitement	0.13	0.16	-0.22	<i>0.31</i> †	0.14	-0.18	37
Frustration	-0.04	-0.14	0.15	-0.01	0.04	0.21	39
Interest	<i>0.31</i> †	0.06	-0.14	0.01	-0.18	0.09	38
Blanks	0.00	-0.18	<i>0.33</i> *	0.05	<i>-0.30</i> *	0.21	42
Open-Response Measures							
Annoyed	-0.07	0.17	-0.11	0.21	0.19	0.00	39
Bored	0.00	-0.26	<i>0.27</i> †	-0.18	-0.12	0.15	39
Confused	-0.02	0.25	-0.14	0.15	0.16	-0.17	39
DTG	<i>0.33</i> *	-0.13	-0.19	-0.14	-0.09	0.19	39
IDK	-0.07	-0.05	0.10	-0.02	-0.06	0.13	39
Negative	-0.18	<i>-0.35</i> *	0.07	-0.09	0.15	0.15	39
Neutral	0.00	0.1	-0.12	0.12	0.01	-0.23	39
Positive	-0.09	0.12	0.18	-0.11	-0.02	0.07	39
Blanks	-0.16	-0.04	<i>0.31</i> *	-0.06	-0.19	0.13	42

†=p≤0.1, *=p≤0.05, **=p≤0.01, Italicization for findings significant under Benjaminni-Hochberg 25% false discovery rate. Note: N varies for closed response questions based on how many students received each type of exclusive question. While most students received at least 1 of each of the 4 emotions, some may not have gotten them all.

Meanwhile, for the open-response tags, students who reported emotions that were coded as ‘negative’ valence spent less time per problem, perhaps rushing through an unpleasant task. Further, we see “DTG” (disengaged from task) tags associated with more problems solved correctly on the first attempt, perhaps consistent with the higher reported ‘Performance Goals’ in Table 34. This is addressed in greater detail at the end of section 6.3.5, but briefly: some students referenced material unrelated to the learning environment when they felt they were succeeding at learning tasks.

Last, when students who chose to ignore and not answer the emotion question (Blank), also received more incorrect attempts per problem, both in the closed and open emotion

assessment. This is a useful new metric, suggesting that students might be simply disengaged, and incorrect answers may be more due to carelessness as opposed to low mastery.

In general, the conclusion to research question RQ3 is that only some emotions are associated to behaviors inside of MathSpring. Confidence (closed), interest (closed) and DTG (open) are good predictors of student math performance (solved correctly on first attempt); Negative valence (open) is associated to rushing through problems (Time Per Prob); giving incorrect answers to problems is associated to low confidence (closed), being bored (open), and to students refusing to report their emotion (blank, in both closed and open); hint requests are associated to high confidence (closed) and excitement (closed); quitting problems after started is associated with completing the emotion question (at least in the closed prompt condition) which is a surprising association. Students' choice to leave self-report prompts blank is covered in greater detail in an upcoming subsection "What leads students to leave self-report prompts blank?". Students appear more likely to leave self-reports blank due to fatigue from using MathSpring for a prolonged period of time. This might not align with students quitting problems, which may occur due to bugs within MathSpring.

Being Frustrated (closed), annoyed (open), confused (open), neutral (open), or having general positive valence (open) was not associated to any one of the behaviors we considered. This is interesting given that those are emotions typically considered in the emotions in education and affect detection. It is always possible that these emotions might be further associated to other behaviors, or to these same behaviors when looking at the data at a finer-grained level.

In general, the results reveal that students feeling well or positive (not in general but in ways that may relate to cognitive processes such as being interested in the material or confident

in one's performance) also tend to perform better, or at least behave more productively than students who experience negative emotions, or ignore reporting their emotion altogether.

6.3.5 RQ4: Why do students believe they feel a particular way?

To examine how students' causal attributions related to their emotional states we ran simple bivariate Pearson correlations between each attribution and each emotional state at the student level. Table 36 shows the results of this analysis.

Table 36 Emotions vs Attributions: Bivariate Correlations

	Attributions								
	Boring	DTG	Easy	Hard	IDK	Material	Negative	Success	Website
	Forced-choice Emotions								
Confidence	-0.21	-0.20	0.43*	-0.14	0.17	0.19	0.18	0	-0.18
Excitement	-0.29	-0.29	0.02	0.35 [†]	0.19	-0.1	<i>-0.18</i>	-0.03	<i>-0.20</i>
Frustration	-0.16	0.09	-0.34	0.13	0.05	-0.29	<i>0.05</i>	-0.15	0.12
Interest	-0.13	-0.23	0.09	-0.03	0.14	-0.06	-0.38*	0.13	-0.26
	Open-Response Emotions								
Annoyed	-0.20	-0.12	-0.11	0.27 [†]	-0.21	-0.01	0.70**	-0.08	0.50**
Bored	0.67**	-0.08	0.43**	0.04	-0.23	0.08	0.19	-0.09	0.07
Confused	-0.08	-0.23	-0.03	0.62**	-0.06	-0.21	0.16	-0.02	0.25
DTG	-0.07	0.64**	-0.22	-0.23	-0.23	0.19	-0.26	0.02	0.14
IDK	0.39**	0.09	-0.15	0.23	-0.01	0.46**	0.28 [†]	-0.12	-0.09
Negative	0	0.15	-0.15	-0.15	-0.22	0.2	<i>0.11</i>	0.22	-0.05
Neutral	-0.15	-0.16	-0.13	-0.15	0.69**	-0.1	<i>-0.16</i>	-0.15	-0.18
Positive	0.05	<i>0.02</i>	0.39*	0.22	<i>-0.24</i>	-0.16	-0.24	<i>0.18</i>	-0.22

N=39, †=p≤0.1, *=p≤0.05, **=p≤0.01 Italicization for findings significant under Benjaminni-Hochberg 25% false discovery rate

As an example for the forced-choice prompts, students tended to report confidence and then that attribute this confidence to the material being particularly easy, rather than attributing their confidence to their ability to successfully solve problems. This may be an artifact of looking for correlations at the student level rather than at finer grained levels where success tended to be more associated with confidence. Disinterest was found to be correlated with negative causal attributions.

Regarding the open-response prompts, annoyance was found to be highly correlated with negative attributions/appraisals of the website; these two attributions (website and negative) happened to be highly correlated as well ($R=0.626$, $p<0.001$). Boredom however had two distinct significant correlations with attributions. The first was redundant: students who described feeling bored, were also likely to attribute this feeling to boring material/experiences. In the second case, students attributed their feelings of boredom to 'easy' material. However, easy material was also associated with positive feelings. Pekrun found that boredom was present in cases where students had both low control and low value of a learning situation (2006, 2010). Perhaps students with a high degree of control (e.g. confidence or positive feelings) who also find their work easy, are less likely to value their work; thus resulting in boredom. This hypothesis is later tested in sections 6.6 and 6.7 which address how students' emotional states may change over time in relation to their interactions with the learning environment. Disengaged from task reports of emotion were significantly likely to be followed by attributions that were also 'disengaged from the task'.

Finally, there was a stark distinction between students who did not know what they felt ("IDK" emotion) and students who didn't know why they felt a way ("IDK attribution). "IDK" attributions were significantly correlated with a 'neutral' emotional state (Table 36) and often follow (table 37) students reporting a "neutral" emotional state. This means that being neutral (a neutral emotional state) has no attributable cause. However, when students responded "IDK" when asked how they were feeling, they were significantly likely to point to both 'boredom' sometimes and the 'material' as reasons at other times. However, they did not seem to give these causal attributions as causal for "IDK" (table 37).

To further address RQ4 we combined the statistically significant emotion/attribution pairings and examined how these new measures correlated with survey measures and students' behaviors. We achieved lower degrees of significance, perhaps due to the increased specificity of our combined emotion/attribution.

There are two important notes about the above Table 36: attributions that could not be validated in as described in prior table are not present, there are instances where pairings of reported emotions and attributions are associated at the student level but not at the action level (or vice versa). The first note is relatively self-explanatory: if an attribution fails an inter-rater reliability check (due to too low of a kappa or very few instances) it is not included on the above table, but instead is included in Table 38 and Table 39 below. The second note is a bit more complex: while the above Table 36 shows a correlation between certain emotions and attributions at the student level, those emotions do not necessarily pair with their attributions. For example, while students who reported more excitement were more likely to give attributions of challenge, there were no instances where a student attributed high excitement (Likert >3) to a cause of item difficulty (see table 37 below). It is simply true that students who tended to report higher excitement were likely to attribute item difficulty as a cause for other emotions they felt at different points of time. A counter example is Frustration: in Table 36 we see no significant relationship between frustration and negative attributions, however we note in Table 37 below that negative attributions were frequently (N=12) given for reports of high frustration (Likert >3). So we can say that while negative attributions are often given for feelings of frustration, there is no significant correlation between reports of frustration and negative attributions at the student level (see Table 36). The association exists at the level of individual reports. Students

who report more frustration are not significantly more likely to attribute their feelings to negative causes.

Table 37 Emotions vs. Attributions Count of Instances

	Attributions								
	Boring	DTG	Easy	Hard	IDK	Matl.	Negative	Success	Website
Forced-choice Emotions									
High Confidence	0	8	12	0	1	1	6	9	2
High Excitement	0	0	0	0	3	2	0	2	1
High Frustration	1	3	0	3	0	1	12	1	6
High Interest	0	0	0	0	0	3	0	0	2
Low Confidence	1	4	1	3	1	1	5	2	3
Low Excitement	9	6	6	3	3	7	22	0	12
Low Frustration	4	3	7	1	4	2	7	1	2
Low Interest	6	4	6	2	2	5	17	0	8
Open-Response Emotions									
Annoyed	1	2	0	1	1	5	22	0	16
Bored	9	0	5	1	6	6	8	0	3
Confused	0	0	0	7	0	1	2	0	2
DTG	0	16	0	0	1	0	0	0	3
IDK	0	0	0	0	9	1	4	0	1
Negative	2	6	1	2	7	6	11	0	3
Neutral	9	1	7	3	48	7	13	4	5
Positive	0	11	18	2	14	3	6	11	1

Totals of >9 instances where reported feeling and attribution are coincident are in bold.

As previously discussed, some attributions were either very rare (less than 4 instances) or achieved a very low Cohen's kappa (<0.4 for example). However, no attribution was found to be invalid for both the open and forced conditions. The attributions of "failure", "growth", and "needs" were exceedingly rare in the forced choice condition, while the attribution of "positive" achieved a very low kappa in the open-response condition. The correlations between these attributions and various self-reported emotional states are displayed in Table 38.

Table 38 Infrequent & Partly Invalid Emotions vs Attributions: Bivariate Correlations

	Attributions			
	Failure	Growth	Needs	Positive
Forced-choice Emotions				
Confidence	N<4	N<4	N<4	0.17
Excitement	N<4	N<4	N<4	0.54**
Frustration	N<4	N<4	N<4	0.17
Interest	N<4	N<4	N<4	0.66**
Open-Response Emotions				
Annoyed	-0.13	-0.17	0.12	Kappa < 0.4
Bored	-0.15	-0.03	0.14	Kappa < 0.4
Confused	-0.13	-0.08	-0.01	Kappa < 0.4
DTG	-0.08	-0.11	0.01	Kappa < 0.4
IDK	-0.10	-0.15	0.02	Kappa < 0.4
Negative	0.04	-0.15	-0.02	Kappa < 0.4
Neutral	-0.15	-0.12	0.06	Kappa < 0.4
Positive	-0.13	0.51**	0.06	Kappa < 0.4

N=39, †=p≤0.1, *=p≤0.05, **=p≤0.01

As with the other relationships between emotions and attributions in Tables 36 and 37 above, the reported feelings and attributions which are correlated at the student level were not always coincident by each student report: an attribution report could be correlated with an emotion report at the student level, while not immediately following that emotional report. Only instances of high interest (>3 per Likert scale) and positive attributions had at least 10 co-occurrences. However, reports of positive valence feelings and attributions of growth had a relatively large number of coincident reports as well (N=8).

Table 39 Infrequent & Partly Invalid EMOTIONS vs ATTRIBUTIONS Count of Instances

	Attributions			
	Failure	Growth	Needs	Positive
Forced-choice Emotions				
High Confidence	N<4	N<4	N<4	5
High Excitement	N<4	N<4	N<4	4
High Frustration	N<4	N<4	N<4	4
High Interest	N<4	N<4	N<4	10
Low Confidence	N<4	N<4	N<4	0
Low Excitement	N<4	N<4	N<4	0
Low Frustration	N<4	N<4	N<4	1
Low Interest	N<4	N<4	N<4	0
Open-Response Emotions				
Annoyed	1	0	0	Kappa < 0.4
Bored	0	0	2	Kappa < 0.4
Confused	0	0	0	Kappa < 0.4
DTG	0	0	1	Kappa < 0.4
IDK	0	0	0	Kappa < 0.4
Negative	7	0	2	Kappa < 0.4
Neutral	0	3	6	Kappa < 0.4
Positive	0	8	1	Kappa < 0.4

After identifying common pairings of emotions and attributions (based on which pairings had >9 instances where a reported emotion and attribution were both present), the new emotion/attribution pairings were tested for significant correlations with the same measures as seen in Table 36 and Table 37 previously. There are fewer significant correlations that previously found in Table 36 and Table 37, roughly on the level of what we might expect due to chanced based on the number of tests. Table 36 contains 64 statistical tests, and we would expect 6.4 to achieve marginal significance ($p < 0.1$) which is in fact what we find. Table 40 below contains 96 statistical tests, meaning we would expect to find 9.6 marginally significant results, and we find 12.

As a reminder to the reader, and specified in the Methods section, students' learning gains were assessed via pretest and posttest with items extracted from the Massachusetts Comprehensive Assessment System Standardized Test (MCAS) practice exams. Learning,

Performance, and Work Avoidance goals were measured through the 18 item GOALS-S survey (Dowson & McInerney, 2004) which provided a means of assessing students' values in terms of control-value theory (Pekrun, 2006). Finally, measures of students' behavior within the tutoring environment (as described in the following section) were aggregated to the student level to provide an overall student level measurement of students' behavior and performance.

Table 40 Correlations between Pretest measures & Frequent Emotion/Attribution Pairings

	Correlation Coefficients					N (Students)
	Math Pre/Post Gain	Mean Pre&Post Score	Mean Pre&Post Mastery Goals	Mean Pre&Post Performance Goals	Mean Pre&Post Work Avoid Goals	
Forced-choice Measures						
Confident/Easy	-0.04	0.10	0.19	-0.03	-0.32 [†]	36
Frustrated/Negative	-0.12	-0.32 [†]	-0.08	0.02	0.14	36
LowExcitement/Negative	-0.18	-0.09	-0.04	-0.05	-0.02	36
LowExcitement/Website	-0.19	0.02	-0.10	-0.06	-0.09	36
LowInterest/Negative	-0.15	0.05	0.05	0.09	0.16	36
Interested/Positive	0.09	0.31 [†]	-0.23	-0.30 [†]	-0.10	36
Open-Response Measures						
Annoyed/Negative	0.03	-0.28 [†]	-0.06	0.03	0.20	39
Annoyed/Website	-0.01	-0.25	-0.01	0.01	0.09	39
DTG/DTG	0.15	0.22	-0.22	0.33*	-0.16	39
Negative/Negative	0.03	-0.11	-0.11	-0.12	0.20	39
Neutral/IDK	-0.09	-0.12	-0.37*	0.08	0.17	39
Neutral/Negative	0.12	-0.06	-0.19	-0.18	0.23	39
Positive/DTG	0.22	0.05	0.07	-0.18	-0.23	39
Positive/Easy	-0.23	-0.12	0.28 [†]	0.11	-0.04	39
Positive/IDK	-0.10	0.03	0.25	0.07	0.00	39
Positive/Success	-0.02	-0.06	0.32*	0.00	0.00	39

†=p≤0.1, *=p≤0.05, **=p≤0.01 Italicization for findings significant under Benjaminni-Hochberg 25% false discovery rate

As a result, we ought to view these findings with a high degree of skepticism. Some of them may warrant deeper examination at the action level to see if possible hypotheses to explain these results are consistent with fine grain action-by-action data.

Table 41 Correlations Between Emotion/Attribution Pairings & Actions

	Correlation Coefficients						N (Students)
	SOF per Prob	Time per Prob	Wrong per Prob	Hint per Prob	Problems Quit	Avg Prob Difficulty	
Forced-choice Measures							
Confident/Easy	0.17	-0.18	-0.02	-0.37*	0.36*	-0.01	36
Frustrated/Negative	-0.30 [†]	0.15	-0.04	0.22	0.16	-0.19	36
LowExcitement/Negative	-0.16	-0.33 [†]	-0.24	0.16	0.00	0.03	36
LowExcitement/Website	-0.05	-0.13	-0.19	0.09	-0.02	0.03	36
LowInterest/Negative	-0.19	-0.23	0.16	0.19	0.04	0.04	36
Interested/Positive	0.20	0.03	-0.15	-0.10	0.06	0.13	36
Open-Response Measures							
Annoyed/Negative	0.03	0.14	-0.06	0.18	0.00	0.02	39
Annoyed/Website	0.02	0.14	-0.10	0.19	0.09	-0.07	39
DTG/DTG	-0.02	-0.30 [†]	0.07	-0.06	-0.02	0.28 [†]	39
Negative/Negative	-0.17	0.10	-0.04	0.19	0.23	-0.19	39
Neutral/IDK	-0.17	0.22	0.03	0.26	-0.01	-0.40*	39
Neutral/Negative	0.15	-0.01	-0.04	0.08	-0.20	0.11	39
Positive/DTG	0.36*	-0.20	0.00	-0.14	-0.36*	0.30 [†]	39
Positive/Easy	-0.27 [†]	-0.22	0.31 [†]	-0.14	-0.08	0.10	39
Positive/IDK	-0.16	-0.08	0.11	-0.22	0.07	0.00	39
Positive/Success	-0.14	0.14	-0.09	0.21	0.24	-0.01	39

†=p≤0.1, *=p≤0.05, **=p≤0.01 Italicization for findings significant under Benjaminni-Hochberg 25% false discovery rate

Among the more interesting findings are the fact that students who report positive valence emotions due to easy problems also seem to be less likely to solve a problem correctly on the first attempt, and make more attempts wrong per problem (see Table 41). This is in contrast to students who report feeling confident because problem are easy, who request fewer hints, but seem to quit more problems (see Table 41). Perhaps the positive/easy cohort are experiencing relief when MathSpring gives them less challenging problems in contrast to feeling like they are behind the rest of the time. As compared to students in the Confident/easy cohort who may be more likely to skip problems they feel are too easy. Both of these hypotheses can be examined in greater depth at the action level. If students who report Confident/easy are indeed skipping problems because they believe they are too easy, we ought to look more closely at their performance to see if their assessment is correct or if they are overconfident and skipping useful

learning opportunities. If their behavior is indeed motivated by challenge seeking it could benefit us to give these students a greater role as partners in their own learning. As for the former students who report positive affect due to easy problems, MathSpring is designed to adjust difficulty level to meet students' needs; however, if success provides an affective and/or motivational boost to a student who may be disengaged it may make sense to introduce artificially easy problems simply as a means to encourage students to re-engage with a learning task. Finally, the case of DTG is an interesting one: students who feel positive and report DTG attributions seem to be solving more problems correctly on the first attempt, quitting fewer problems, and also tackling higher difficulty level problems. If these students are prone to affirmations which appear disengaged from the learning environment (per Table 31: "because im batman and awesome"), it's difficult to determine the direction of causality. Sabourin (2011) found that students who are disengaged from tasks tend to re-engage and perform better. It's also possible that in this case students who are doing well express positive emotions through reference to fictional characters (e.g. Batman).

6.4 Discussion of Initial Student-Level Results

RQ1: To a large extent students' reports of emotion seemed to fall under simple valence (Table 31). Epistemic emotions were less common than others have found (D'Mello & Graesser, 2012), but it seems that students did indeed identify many common emotion constructs (i.e. boredom, confusion, frustration). The criticism of self-report that students are incapable of correctly articulating their emotions (Bieg et al., 2014) would seem to be supported by students' imprecise use of terms: for example "boredom" related tags occurred for both emotional and attributional prompts. However, given the aforementioned imprecision of whether such constructs are emotional or cognitive (Clore & Ortony, 2000) perhaps students' difficulties are a product of the

constructs themselves rather than their own self-awareness. D’Mello & Graesser (2012) identify that while some researchers believe confusion to be an emotion (Keltner & Shiota, 2003; Rozin & Cohen, 2003; Silvia, 2009), others identify it as a cognitive state (Clore & Huntsinger, 2007). If epistemic emotions indeed include both emotional and cognitive components (Pekrun, 2010), then ambiguity between whether they be identified as emotions or cognitive attributions might not be a misidentification of these constructs, but rather an accurate reflection of their dual nature.

RQ2: Student Goals (performance, mastery and work avoidance) are associated to a variety of student emotions inside MathSpring, with negative valence emotions being related to work avoidance and positive emotions being associated with mastery goals.

RQ3: There were relatively few significant findings between students’ reported emotional states and their behaviors in Table 35. Confident students seemed likely to perform well which is expected. Confident students’ increased use of hints is unexpected, it’s tempting to suspect hint use might lead to confidence rather than students with a greater degree of confidence in their abilities choosing to request help.

The negative valence emotions being negatively correlated with problem time suggests students may be rushing through their work due to discomfort. Boredom seems to follow the same trend here with higher incorrect attempts, it seems consistent that more cognitively engaged types of negative emotions (annoyed and confused) are less error prone (Baker et al 2010). Yet closed response frustration shows the opposite trend.

RQ4: Exploring the relationship between attributions and emotions (Table 38 through Table 39) revealed that students can identify multiple distinct causes for the same emotional state (e.g.

boredom), and further that the same cause can lead to multiple distinct emotional states (e.g. an easy task). However, looking at the correlations in this table we find that by looking for a more specific construct given a particular attribution in some cases we simply decrease the significance we might find if we simply looked at an emotion alone. For example, compare confidence in Tables 35 to confident/easy in Table 41. In comparing the closed vs. open response self-reports, we find that each set of prompts captures different associations with students' predispositions and behaviors. In terms of identifying these associations at the student level, neither method appears to be superior, although the open response approach is quite a bit more labor intensive.

Additionally, each method appears to have different strengths: closed response measures can focus better on identifying expected constructs within students while open response methods appear to provide a better link between emotions and attributions. A common theme throughout this work is a tension between generalizability and specificity. We want constructs that are particularly germane to students' experience yet we would also want these constructs to be common enough to warrant study, or at least to produce statistically significant results.

Students appear to recognize similar emotional states and attributions to those that researchers describe, however they seem more likely to report the simple valence of feelings than more subtle epistemic emotions as operationalized by Baker, D'Mello, and Pekrun. As the present study was conducted at the student level (i.e. aggregating behaviors/reports to an average for each student) we can't clearly see the sort of causal moves from confused to frustrated to bored as articulated in D'Mello & Graesser (2012). Evidence to support the control-value theory is modest as well: boredom is theorized to be attributable to low control or low value, therefore

as students who find the material easy would likely have high control we would expect to find low value among these students (Pekrun et al 2010).

Our work provides several leads to further analyses directed to the fine grained individual action or problem level, rather than limiting our analyses to the coarse grained student level. It is our hope that by digging deeper we may find additional support from the moves described by D’Mello (D’Mello & Graesser;2012), and the attributions identified by Pekrun (2010).

6.5 Summary of New Research Questions from Student Level Analyses

Several, possibly spurious, trends have been suggested due to the student level analyses. While there are already several questions to address in the action level analyses, these new items are presented here for the reader and author to view in a neatly summarized form.

1. Students who report “IDK” when asked how they are feeling are also likely to cite “boring” and “negative” “material” as attributions at other self-report opportunities (not “IDK”). Perhaps reported feelings of “bored” and “IDK” are describing the same state of boredom. If so, then it seems likely students might go from reporting “IDK” to reporting “bored” or vice versa.
2. Emotion reports of “DTG” appear to be correlated with solving problems correctly on the first attempt and performance goal orientation, while negatively correlated with work avoidance goals. Prior work (Sabourin et al., 2011) hypothesizes that students who engage in disengaged type behaviors may do so as a means of taking a break and becoming re-engaged. Examining the actions and reported emotions which directly precede and follow emotional reports of “DTG” may support this: for example students may make fewer errors after reporting “DTG”.

3. A high degree of “annoyed” reports of emotion seem linked to negative website attributions. What are the events that precede/follow these reports? Is this a case of students externalizing blame due to temporary lapses in performance (i.e. “sour grapes”), or are these students voicing concerns with bugs or errors in the system itself.

6.6 Initial Action Level Analyses: Methods

The finer grain action level analyses of this work began was inspired by prior work by D’Mello & Graesser (2012), which found empirical support for temporal transitions between emotional states within a tutoring session, and hypothesized that student experiences in the tutoring system might be the cause of these changes within short periods of time. As well as Pekrun’s control-value theory (Pekrun, 2006) which theorized that appraisals of students’ experiences within a learning task are the cause of a variety of different emotional states. We began by examining these hypotheses in an attempt to replicate past results in the MathSpring tutoring system (Karumbaiah et al, 2017).

First, we consider the work of D’Mello & Graesser (2012) that tested four hypotheses regarding engagement and emotional state:

1. The Disequilibrium Hypothesis: Students in a state of engaged flow may become blocked, enter a state of cognitive disequilibrium, and experience confusion.
2. The Productive Confusion Hypothesis: Cognitive disequilibrium is an opportunity for students to learn something new while they process information that is new or challenges previously held schemas (Graesser & Olde, 2003; Rodrigo, 2011; VanLehn et al, 2003).
3. The Hopeless Confusion Hypothesis: Students who experience sustained confusion will eventually experience frustration at their inability to resolve their disequilibrium

4. The Disengagement Hypothesis: Persistent failure associated with frustration will eventually lead to boredom.

D'Mello & Graesser (2012) tested these hypotheses by tracking students' emotional states across time and using one-sample t-tests to see if a particular transition from one emotional state to another occurred at a likelihood significantly greater than zero. Across two studies they found support for hypotheses 1, 2, & 3: students did indeed transition between emotional states as hypothesized significantly more often than chance. They found significant support for hypothesis 4 in their second study, possibly due to changes in experimental design. While they had not hypothesized it, they also found that transitions from Boredom to Frustration were significantly more likely than chance.

D'Mello & Graesser's (2012) work models emotions as a consequence arising from situational factors including preceding emotional state and cognitive processes (i.e., disequilibrium). These factors are proxy to events that occur within the learning environment: cognitive disequilibrium occurs when students are presented with information that does not fit within their existing schemas (Piaget, 1977). It is implied that resolved or sustained cognitive disequilibrium is a result of students' perceptions of their own performance within a learning environment. Pekrun's (2006) control-value theory approaches achievement emotions which occur during a learning task in a similar way. According to the control-value theory, students' degree of control (ability to affect the outcome) and value (whether or not the outcome is seen as positive, negative, or unimportant) drive students' emotions. Frustration, for example, is theorized to be the result of having a low degree of control over an outcome despite valuing that outcome, while boredom is due not valuing the learning outcome.

The control-value theory and the four aforementioned hypotheses may be two distinct ways of explaining how students experience cognitive disequilibrium. Students who are motivated to resolve their cognitive disequilibrium (Piaget, 1977) but are unable to do so could be described in terms of the control-value theory as having high value for a particular outcome but a low degree of control. While Piaget's work focuses cognitive function, an understated aspect of the theory is that a state of cognitive disequilibrium is described as uncomfortable and motivates students toward a resolution (Fosnot, 1996; Piaget, 1977). This less emphasized aspect of cognitive disequilibrium bears further exploration in terms of control-value theory. Perhaps subjective value of the learning task moderates the relationship between cognitive disequilibrium and discomfort; within control-value theory, the motivation to resolve disequilibrium could be explained simply with the student's task value (whether or not the learning task is seen as positive, negative, or unimportant).

The relationship between control-value theory and the 4th hypothesis can be explained as follows: sustained frustration would cause students to become resigned to expected failure and use the coping strategy of valuing the task less than before, become disengaged, and express that they are now 'bored'.

6.6.1 New Hypotheses Building on Prior Work

We set out to first replicate many of the temporal transitions between emotional states as found by D'Mello & Graesser (2012), and further to examine students' performance between those states. Many of D'Mello & Graesser's hypotheses attribute cognitive function to the transitions between emotional state, i.e. by accounting for cognitive performance we expect to better explain transitions between emotional states.

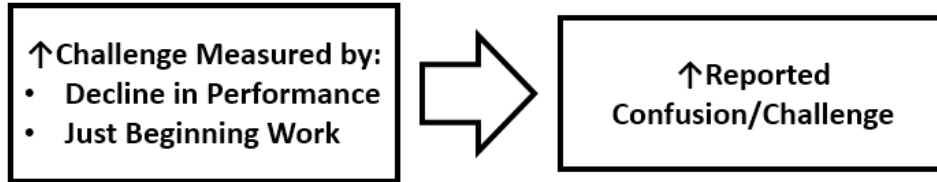


Figure 6 Disequilibrium Hypothesis: Increased challenge leads to confusion

1) Confusion occurs when students first encounter unfamiliar material (as opposed to frustration or boredom which may occur with sustained confusion), see Figure 6. As a result we hypothesize confusion will be more likely to occur:

- a. Shortly after a drop in performance
- b. At the start of a day's session as compared to the base incidence of confusion for a given student

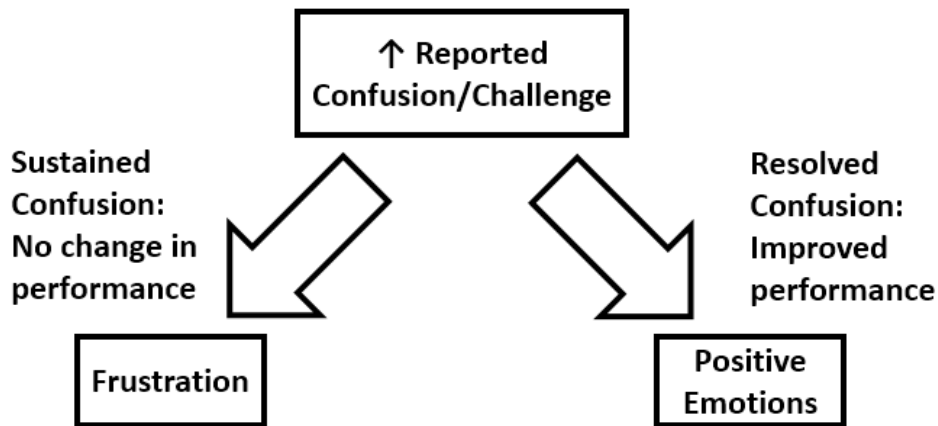


Figure 7 Productive Confusion Hypothesis: Increased confusion leads to positive emotions if resolved or frustration if not

2) Sustained confusion is theorized to lead to frustration. Do we replicate the findings of D'Mello where frustration is more likely to occur after confusion?

- a. If we specifically look for instances of confusion followed by poor performance is subsequent frustration even more probable (see figure 7)?

b. If confusion is followed by good performance, then are students likely to report positive affect more often for resolving their confusion than average (see figure 7)?

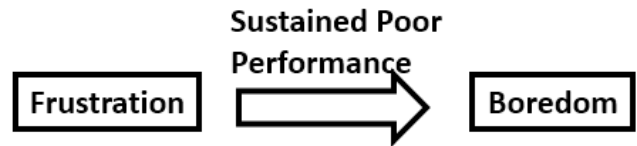


Figure 8 Hopeless Confusion Hypothesis: Frustration leads to boredom through sustained poor performance

3) Sustained frustration is theorized to lead to boredom (see Figure 8). Is the probability of boredom higher after reported frustration than for other cases? If we specifically look for cases where frustration is followed by poor performance is boredom even more likely there?

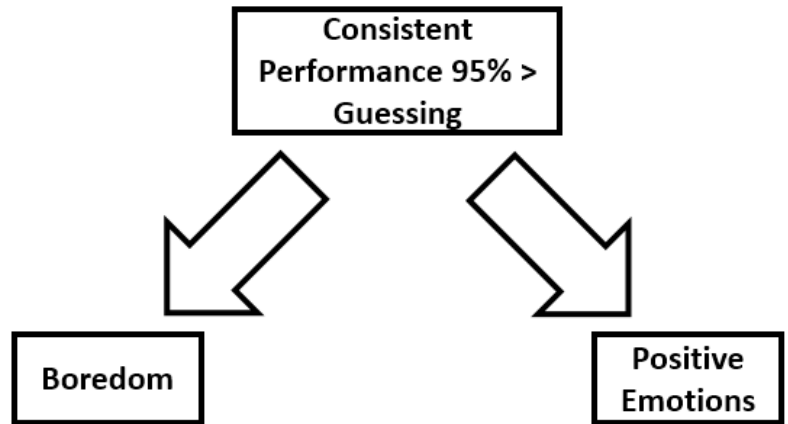


Figure 9 The Disengagement Hypothesis: Consistent success may imply lack of challenge leading to boredom, or gratification at success

4) Students who consistently perform well may become bored their work, or they may be more likely to experience positive valence emotions due to their consistent success. Given a set of consistent high performance we examine whether either of these affective states are more likely in this case than average (see Figure 9).

- a. Consistent Success precedes Boredom
- b. Consistent Success precedes Gratification (positive emotions)



Figure 10 Persistent Boredom Hypothesis: Boredom precedes & follows boredom

5) Prior work (Baker et al, 2010; D’Mello & Graesser, 2012; McQuiggan et al., 2010; Rodrigo, 2011) has shown boredom to be a persistent state for learners: students who become bored are likely to stay bored (see Figure 10). We attempt to replicate this result here.

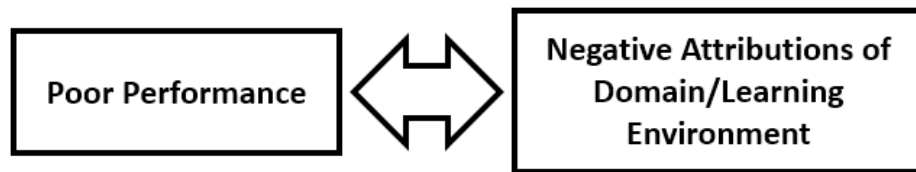


Figure 11 “Sour Grapes” Hypothesis: Complaints about the domain or learning environment are likely preceded by poor performance

6) Students may be more critical of the domain or learning environment after performing poorly than they would when performing normally. Perhaps reducing these students’ perception of task value as a means of avoiding discomfort due to poor performance (see Figure 11).

- a. Poor Performance precedes Negative Attributions of Domain or Learning Environment
- b. Negative Attributions of Domain or Learning Environment precede Poor Performance

Table 42 Emotional and Causal Attribution Measures for Open and Closed Response Conditions

Open Measure	Analogous Forced-Choice Measure
Confusion	Causal Attributions of Challenge
Annoyance	High (>3) Frustration Likert Report
Positive Valence	High (>3) Excitement Likert Report
Boredom	Low (<3) Interest Likert Report
Negative Causal Attributions Domain/Learning Environment	Negative Causal Attributions Domain/Learning Environment

As part of our work was to compare open response vs forced-choice responses, we devised analogous measurements between each system as shown in Table 42 above. Because there was no way for students to report the feeling of confusion (as confusion was not one of the four emotional states they were surveyed on in the forced-choice condition) we instead used students' attributions of "hard" or challenging work. The remaining analogous measures are close comparisons: annoyance is treated as analogous to high degrees of frustration, positive valence as analogous to high degrees of excitement, and boredom as analogous to low degrees of interest. It should be noted that we do not claim these analogous states to be equivalent. Excitement for example has been described as being a state of high activation in addition to being positive in valence. Further, disinterest and boredom may be conceptually distinct constructs even if they may be closely correlated.

6.6.2 Performance Measure: Probability of Performing Better than Random Guessing

Because MathSpring uses multiple choice prompts, on a problem with 4 choices there's at least a 25% chance of getting the problem right by simply guessing. Conversely, a student who makes a single mistake before solving a problem correctly is less likely to be guessing than a student who makes 2 or 3 errors before choosing the correct answer. Looking solely at a single problem gives very little idea of how a student may be doing in general at the time that a self-report is given, as a result the prior six problems to a self-report are considered. Further, rather than simply marking performance as percent correct out of a possible 100% we consider the likelihood that a student is outperforming guessing as our performance metric.

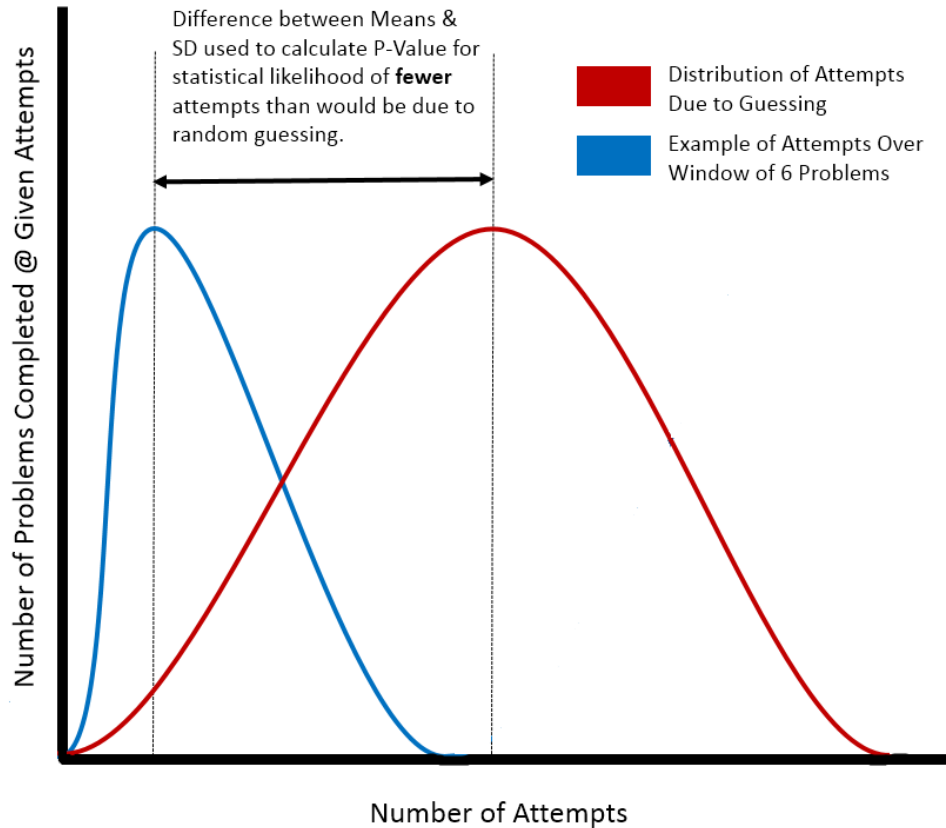


Figure 12 T-Test Comparing Performance Observed to Performance Expected due to Random Guessing

This metric incorporates partial credit by accounting for the number of attempts students make over a set of six problems, and using an independent samples t-test (see figure 12) to compare those attempts to the statistical average and standard deviation we would expect due to random guessing. A one-tailed test is used, because we are only considering that students may systematically outperform random guessing rather than intentionally making efforts to select incorrect answers. The t-test outputs a p-value wherein a lower value (i.e. $p < 0.05$) suggests the two samples only have a 5% likelihood of being drawn from the same population.

The minimum value or floor that this metric could have is $p=0.5$ or 50% likelihood of being performing better than chance. This is because a normal curve that shares a mean with the normal curve distribution of attempts due to guessing is 50% likely to be better than guessing,

and also 50% likely to be worse than guessing. We calculate discrete probabilities at 5% increments from 95% ($p < 0.05$) through 75% ($p < 0.25$), and assign them to windows containing 6 prior problems. In cases where the probability of outperforming guessing is less than 75% we face a challenge in that most t-statistic tables don't offer measures of significance above $p < 0.25$ for a one tailed test ($p < 0.5$ for a two tailed test). So to create a good measure for all the windows that were relatively close to performance due to random guessing we took all problems from windows which fell short of the 75% likely to be better than chance bar, and ran an independent samples t-test comparing their mean and standard deviation to the mean and standard deviation guessing would produce over a fictional sample of $N=6$ problems. The sample size N was decreased to six problems, because while this performance was close to chance at a sample size of $N=5458$ we achieved statistical significance which would not be present in a small window of 6 problems. We found that given this smaller window of $N=6$ these problems were 59% likely to be outperforming random guessing. This value was then assigned to windows which fell below the 75% threshold.

Table 43 Example Cases Leading to 95% Better than Chance Likelihood

Example	Problem 1	Problem 2	Problem 3	Problem 4	Problem 5	Problem 6	Performance > Chance
1	0 Errors	0 Errors	3 Errors	0 Errors	0 Errors	0 Errors	95%
2	0 Errors	1 Error	1 Error	0 Errors	1 Error	0 Errors	95%

See Table 43, on six multiple choice problems with 4 options (i.e. A,B,C,D) a student could achieve a 95% probability of performing better than guessing with either 5 problems solved correctly on the first attempt and 1 problem solved correctly on the third attempt, or with 3 problems solved correctly on the first attempt and 3 problems solved correctly on the second attempt. Any probabilities below 75% were rounded to 50%.

6.6.3 Comparison by Student & Dependencies

We tested the aforementioned hypotheses using a simple paired samples t-test comparing students' behaviors/reports in a particular case (e.g. performance in terms of % better than guessing right before reporting confusion) against measures of the same students averaged across the entire assignment (that student's average % better than guessing on all given problems). In this way we can see if the measures taken at a particular point in time differ significantly from the norm. It should be noted that this particular use of T-Tests is not meant to identify statistically significant results, but rather to provide a rough measure of students' performance accounting for the likelihood of that performance being due to random guessing. As a result, we are not claiming that each of the hundreds of problem windows are likely better than chance at the significance we calculate, rather this is meant to give us a probability that that particular window is better than random guessing and account for the ever present likelihood that a student's performance is in fact due to random guessing.

In this example if we hypothesize that a drop in performance precedes confusion we would only consider all instances where students report confusion. Then we would look at the events prior to these self-reports of confusion, aggregating all instances of a report of confusion and the performance prior to that report, and later aggregate this at the student level, obtaining a single value (a mean of the metric) per student. Finally, we run a paired sample t-test comparing each student's performance before reporting confusion against each student's performance in general, regardless of emotion. The result of this test shows how much greater or less than average our specific condition is (at the student level), and whether or not that effect is statistically significant.

One concern with this method is the set of pre-conditions for each hypothesis may limit our sample size (N). For example, while the total sample group of students who report ‘challenge’ in their work may be small for the forced-choice condition (N = 12), that number becomes even smaller when we are limited only to students who are randomly surveyed on their degree of ‘frustration’ immediately after reporting ‘confusion’ (N = 4). As a result, we set N = 8 as a minimum threshold to report results.

6.7 Initial Action Level Analyses: Results

6.7.1 Increased challenge leads to confusion

As there were no forced-choice prompts for confusion, instances in which students attributed their emotional state to feeling challenged by hard material were considered as confused (see Table 44).

Even so, reported instances of ‘confusion’ or attributions of challenge were relatively rare in both the closed (N=12 students) and open-response (N=8 students) conditions. The following step of identifying which student had completed at least 6 problems prior to reporting dropped our sample sizes even further (N=6 & N=4 respectively) so the two conditions were merged.

As previously described in methods, the probability that a student was performing better than guessing was calculated using a sequence of 6 problems. For this test, decreases in performance were measured based on comparing the averaged probability calculated for 3 problems before and 3 problems after the self-report. This further constrained the number of available students leaving only N=10 students who reported ‘confusion’ after completing at least 6 problems and before completing an additional 3 problems.

The performance slope prior to a self-report of ‘confusion’ or ‘challenge’ was compared against each of these students’ average performance slope prior to any and all self-reports. On average, students were slightly more likely to experience a negative performance slope on average than they were prior to reporting ‘confusion,’ but not significantly so. However, students reporting ‘confusion’ or challenging material were 10.5% marginally more likely ($p < 0.1$) to report it within 6 problems of starting their work than they were on an average self-report.

6.7.2 Increased confusion leads to positive emotions if resolved/ frustration otherwise

In testing Hypothesis 2 we encountered difficulties testing our forced-choice condition due to too small a sample size. Only 4 students were asked to report on their degree of ‘frustration’ after reporting challenging work. Fortunately, there were still 9 students in the open-response condition who reported feeling ‘confusion’ who could be tested for subsequent emotions (see Table 44).

Part A) If we specifically look for instances of ‘confusion’ followed by poor performance is subsequent ‘frustration’ even more probable?

Our findings were neither significant nor in the same direction as our hypothesis: students were 6% less likely to report ‘frustration’ after ‘confusion’ than they were to report ‘frustration’ on average.

Part B) If ‘confusion’ is followed by good performance, then are students more likely to report positive affect than average for resolving their confusion?

Perhaps our non-significant results were due to students resolving their cognitive disequilibrium. We selected students whose performance was >75% likely better than guessing. Again, unfortunately this sample group only included 6 students only 1 of whom reported

positive valence emotions. Our findings were again non-significant and counter to our stated hypothesis (students were 35% more likely to be positive on average than after reporting confusion).

6.7.3 Frustration leads to boredom through sustained poor performance

Once again, the forced-choice condition had too few instances (N=4) where particular reports of high 'frustration' (>3 on a Likert scale) were followed by self-reports requesting students to report on their degree of interest. However, the open-response approach allowed for these analyses to be performed. Out of the students who reported annoyance on a particular self-report (N = 12), none followed up with a subsequent report of 'boredom.' In fact, reports of 'boredom' were 10.6% significantly ($p < 0.025$) less likely to occur after a report of annoyance as compared to each student's average likelihood of reporting 'boredom' (see Table 44).

This finding obviated the need to proceed to the next test to see if 'boredom' after 'annoyance' ('frustration') would become even more likely given continued poor performance. However, it appeared that many students reported additional annoyance after an initial report of annoyance. This led us to test a new alternative hypothesis 3: 'frustration' precedes 'frustration.' We found students (N = 12) were indeed 17.6% significantly ($p < 0.05$) more likely to report feeling annoyed after a prior report of annoyance ('frustration') as compared to average. These findings are consistent with Rodrigo (2011) which found frustration to be a persistent state.

6.7.4 Consistent success may imply lack of challenge leading to boredom, or gratification at success

Our metric for consistent success was maintaining a probability of performing >95% better than guessing and solving a problem in <60 seconds on average for 8 consecutive problems (see Table 44).

Part A) Initially we tested to see if after consistently performing well, students were more likely to report 'boredom.' The test of the open-response students (N=20) found students were more likely to report 'boredom' after consistently performing well 3.6% greater than average, but with no significance.

For the Forced-Choice condition we found that students' reports of interest were actually 33% higher than average Likert ratings when sampled at times after meeting our consistent success criteria.

Part B) The open-response measure of emotions showed reported feelings of positively valenced emotions. We found that students (N = 20) were 13.9% significantly more likely to report positively valenced emotions after consistent success against positive emotions ($p < 0.05$).

For the Forced-Choice condition there was not a ready analog to express gratification. There weren't sufficient instances where students were surveyed in terms of "Excitement" or replied to the attribution prompt to achieve statistical significance, measures of "Confidence" were deemed too distant from "gratification".

Table 44 Initial Fine Grain Results Summary: Results with marginal significance ($p < 0.1$) in accordance with hypotheses labeled “CON”, non-significance “INC”, significance counter to hypothesis “DIS”

	Merged Data	Open Response	Closed Response	Result
Hyp 1A	A drop in performance precedes confusion			
Hyp 1A Results	2% > Avg Not Sig N=10	N < 8 : N/A	N < 8: N/A	INC
Hyp 1B	Starting an assignment precedes confusion			
Hyp 1B Results	11% > Avg ($p < 0.1$) N=21	14% > Avg ($p < 0.1$) N=13	6% > Avg Not Sig N=8	CON
Hyp 2A	Confusion precedes frustration (N insufficient to test performance)			
Hyp 2A Results	6% < Avg Not Sig N=9	N < 8 : N/A	N < 8: N/A	INC
Hyp 2B	Confusion precedes Positive Affect (N insufficient to test performance)			
Hyp 2B Results	1% < Avg Not Sig N=10	N < 8 : N/A	N < 8: N/A	INC
Hyp 3	Frustration precedes boredom			
Hyp 3 Results	N/A	11% < Avg ($p < 0.025$)	N < 8: N/A	DIS
Hyp 3 ALT	Frustration precedes Frustration			
Hyp 3 ALT	N/A	18% > Avg ($p < 0.05$)	N < 8: N/A	CON
Hyp 4A	Consistent Success precedes boredom			
Hyp 4A Results	N/A	4% > Avg Not Sig N=20	33% > Avg ($p < 0.05$)	DIS
Hyp 4B	Consistent Success precedes gratification			
Hyp 4B Results	N/A	14% > Avg ($p < 0.05$)	N < 8: N/A	CON
Hyp 5	Boredom precedes boredom			
Hyp 5 Results	N/A	1% < Avg Not Sig N=10	34% > Avg ($p < 0.01$) N=9	CON
Hyp 6A	Poor performance precedes complaints about domain or learning environment			
Hyp 6A Results	2% > Avg Not Sig N=26	N/A	N/A	INC
Hyp 6B	Complaints about domain or learning environment precede poor performance			
Hyp 6B Results	1% < Avg Not Sig N=19	N/A	N/A	INC
Hyp 6 ALT	Poor performance precedes frustration			
Hyp 6 ALT	2% > Avg Not Sig N=28	3% > Avg Not Sig N=9	1% > Avg Not Sig N=19	INC

6.7.5 Boredom follows (and precedes) boredom

We replicated prior findings in the forced-choice condition, after reporting instances of low interest (Likert < 3) we found that subsequent reports of interest were 34% lower in interest than students’ average reported interest (or 34% higher in boredom), this difference was highly significant ($p < 0.01$) although over a small portion of students (N=9), see Table 44.

The same was not true for instances in the open response condition where students could elect to report boredom, but were not being directly surveyed on their degree of interest. Perhaps this is due to boredom being ubiquitous, but under-reported when reporting one’s boredom

requires writing out text. Further analyses follow in section 6.9.4 examining students' tendency to leave prompts blank.

6.7.6 Complaints about the domain or learning environment are likely preceded by poor performance

Analyzing a relatively large sample of students ($N = 26$) with negative attributions of the domain of mathematics (or sub-topics) or of the MathSpring ITS learning environment showed no significant difference with performance before or after such attributions and how those same students performed on average. Students who complained about the domain or the learning environment were no more likely to perform poorly before (Part A) or after (Part B) such an attribution than normal (see Table 44).

This led to an alternate hypothesis: Was below average performance even related to reports of annoyance or 'frustration'? Surprisingly, we found that students were not significantly likely to have performed below their average performance prior to reporting annoyance or high 'frustration' (Likert >3). In fact, students appear to perform slightly (not significantly) better than their average before reporting 'frustration/annoyance'. It should be noted that these students may perform below their class average in general, but in terms of problem to problem these students did not perform worse than usual.

6.8 Discussion of Initial Action Level Results

Our findings are markedly different from prior research. Many of our initial hypotheses are found to be inconclusive with low statistical insignificance and still others ran significantly counter to our expectations, e.g., the case of 'frustration' preceding 'boredom'. We describe three possible sources for this discrepancy: small sample size, differences in data collection techniques, and relationship to the learning environment.

The small sample size is in some ways the result of running experiments that involve a series of dependent requirements. By first selecting students who report ‘confusion’, and then selecting students who go on to report ‘frustration’ we limit our total sample simply by finding a more specific series of events. A particular limitation of our work is testing both an open and closed Likert scale types design for reporting one’s emotions. Both self-report techniques have limitations. The closed response Likert report relies on asking students about one of four emotions at random; thus sequential orders of emotions are rarer because students don’t have the opportunity to report any common emotional state, only specific ones. However, the fully open self-report may also introduce a challenge in that students may be unable or unwilling to articulate their emotions at length via text (Conati & Maclaren, 2009; Nielsen, 1991; Tourangeau & Yan, 2007). Even if students report wholly accurately, the open-response offers so much choice that too many options are available to productively construct models. The process of tagging the responses to quantify the responses reach a tractable number of tags for emotion (N=8) or attributions (N=13) is also time intensive and potentially provides another area where coders may misinterpret students meaning. Perhaps providing students with a set of pre-existing tags to select their meaning would help. Alternatively, if Likert scales are to be used, it may make sense to simply offer a scale of valence from extremely positive to extremely negative in every case.

While a lack of viable data can explain the lack of statistical significance in many cases, the significant findings are consistent with what we would expect from an ITS learning environment which is adaptive and reduces difficulty level so that each problem is within easy reach of students’ abilities. Baker et al.’s (2010) work has shown that the persistence of various affect states varies by learning environment. Consistently, across three platforms (AutoTutor,

Aplusix, & The Incredible Machine) they find boredom to be the most persistent affective state. This comes despite engaged concentration being the most prevalent affective state. This makes our finding that 'boredom' is not significantly likely to be followed by 'boredom' anomalous. Likewise we find that 'frustration' precedes 'frustration.'

The notable differences between our results and those in Baker et al., (2010) may be attributable to differences in how affect/emotion is measured in each study. Baker et al.'s work asked students to classify 20 second video clips of themselves as they worked in terms of 7 affect states (boredom, confusion, delight, engaged concentration, frustration, surprise, and a neutral state). Several differences exist between these two approaches. First, our reports occurred in real-time while students were working in the learning environment; prior research (Baker et al., 2010; Graesser & Olde, 2003) asked students to report their emotion after the fact based on reviewing video of themselves. This is different in two ways: in one case, the report is made in situ and the other is not, and in one case the information available to the student is an immediate internal experience, and in the other a classification of externalized affective quantities.. Secondly, prior research (Baker et al., 2010; D'Mello & Graesser, 2012) provides students with a list of 7 possible affective states whereas we provide students with an open-response text box. It's possible that students are more likely to select boredom repeatedly when it's offered as an option as opposed to reporting boredom when given no initial prompt to suggest their emotional state. This distinction may be due to students' naivete in articulating their emotional experiences (Conati & Maclaren, 2009; Porayska-Pomsta et al., 2013) or perhaps due to one method influencing students emotions by providing a limited number of emotion prompts.

Ultimately, a major failing of our work is that in addressing/replicating prior work we could have more closely followed the prior methods used. If we had used both the in situ self-

report methods herein in addition to students reviewing video after the fact we might better be able to discern whether our different results were due to self-report methodology or due to possible differences between MathSpring and the 3 ITSs tested in prior work (Baker et al., 2010). This is particularly unfortunate, because the narrative that MathSpring provides to adapt problems to challenge students less consistently explains many of our results.

First, note that ‘confusion’ appears to occur not after an immediate drop in performance, but does occur after students have just begun work within MathSpring. Given that confusion occurs when students experience challenge, the environment may be limiting students’ degree of challenge to the Zone Of Proximal Development (Chaiklin, 2003). So while material may be challenging there is no strong link between failure and challenge: despite being challenging the work is within students’ grasp. We do however note confusion early on, when students have done five or fewer problems. In this case the system may not yet have calibrated difficulty level based on students’ ability.

The finding that ‘frustration’ does not significantly precede ‘boredom,’ but does significantly precede more ‘frustration’ doesn’t fit with the hypothesis that students eventually give up or become hopeless & bored due to perceived insurmountable challenge. Fortunately, our work asked students follow up self-reports after having them report on their emotional state. Specifically we asked students “Why do you feel that way?” and “What do you wish you could do to improve this class right now?” (Figure 10). If students reported frustration due to insurmountable difficulty we would expect their attributions to either/both of these questions to relate to the difficulty of the items. They didn’t. We examined the total open-response reports of annoyance for what the most prevalent tags for causal attributions for why students felt that way and what they thought the authors of MathSpring could do to improve their work experience.

After averaging attributions by student we found some of the most common causal attributions were “negative” at 46.5% and “website” at 34.5% by report by student for the total reports of feelings of “annoyance”. In contrast, attributions of challenge made up <1% of the attributions for reports of “annoyance”. When asked what MathSpring’s authors could do to improve students’ experience the most common responses were “I don’t know” at 16.7%, calls to improve MathSpring’s design at 14.8%, and critique of the material at 12%. These data seem to suggest that students report annoyance not due to a high degree of difficulty so much as an active dislike of the MathSpring system or the content. Students’ reported feelings of antipathy seem to have less to do with their performance than with their preferences. This is later consistent with the findings of Hypothesis 6: students who complain about mathematics or MathSpring are not externalizing blame due to poor performance. The data suggest that some students just dislike the domain (of mathematics) and/or the learning environment. Further, this is consistent with our extension to Hypothesis 6 applied after seeing our initial results: students who report annoyance or a high degree of frustration do not seem to be making these reports after poor performance.

This suggestion that the results described in this paper are due to fundamental differences between MathSpring and the results found using other ITS learning environments (Baker et al., 2010) does have some precedence. While Baker et al. (2010) find that boredom is consistently the most persistent emotion, there are differences between the three learning environments tested. MathSpring may simply be a more extreme example of variance in dynamics between affect states in ITS learning environments.

6.9 Extensions to Initial Student & Action Level Results

In sum, we found a lack of support for several prior findings in the literature, e.g., our results do not support the hypothesis that ‘confusion’ follows poor performance; that ‘frustration’ follows

‘confusion;’ or that ‘boredom’ follows ‘frustration’. We did find that students tend to feel good after a period of sustained success, but did not find that ‘frustration’ follows a period of failure or poor performance.

More importantly though, we found that students feelings of ‘frustration’ were largely attributable to critiques of the learning environment or the domain (mathematics). These concerns may be related to long term performance as positive feelings of affect have been related to learning later in life (San Pedro et al., 2013), but in the short term these negative feelings are unrelated to performance.

A critical weakness of our work is the fact that we did not employ prior techniques of affect detection (Baker et al., 2010; D’Mello & Graesser, 2012). If we had, we might be able to show that this particular cohort of 85 students using MathSpring showed no link between ‘annoyance/frustration’ and poor performance. Such a finding might simply be explained by different methods in data collection. If our work had incorporated these methods, and we had shown no significant link between performance and reported feelings of ‘annoyance/frustration’ the lack of relationship could be attributed to a quality of the MathSpring itself. Perhaps this result is due in part to the fact that MathSpring is designed to foster a Growth Mindset (Dweck, 2006; Karumbaiah et al., 2017) in students which encourages framing failures as an opportunity for further growth rather than as a cause for negative emotions.

6.9.1 What Annoys Students?

While our prior findings (Hypothesis 6) illustrate that frustration and annoyance are not significantly preceded by poor performance. This leads to the question of what annoyance might be attributed to? To focus on this question in depth, we present Table 45 below which depicts the

most common causal attributions students gave for reports of either high frustration (Likert >3) or annoyance. For a more detailed idea of the sorts of attributions students would offer for their reported feelings of frustration see Appendix I.

Table 45 Attributions for Annoyance/Frustration

	Total	Attribution Negative	Attribution Website
High Frustration	27	12	6
Annoyance	39	22	16

Most student responses explaining why they felt annoyed or frustrated seemed to be composed of negative attributions and attributions of the website. Perhaps these critiques of the website are due to bugs as it has been in other studies (Mentis, 2007). At the time that this study was conducted MathSpring was undergoing some technical difficulties which causes some problems to appear to simply as blank white space. In these instances students would often quit a problem rather than complete it. In order to test this, Tables 46 and 47 contain the first 4 sequential problems leading up to a self-report, as well as the average by student and overall average for the total sample group. For the forced choice (frustration > 3 per Likert report), the reported number of problems quit (completed without solving) did not appear to be much higher than the average by student.

While we did not find that performance is worse prior to a self-report of frustration when looking at performance better than chance given a running average of problems (see hypothesis 6A in Table 44 we do find that immediately prior to self-reports of frustration that students do seem to make more errors. The sample size here differs for similar reasons: earlier we were limited to problems which had a window of 6 problems attempted prior to a self-report, here we consider all instances and look at the 4 prior problems.

Table 46 Problems prior to Report of Frustration (N = 23)

TotalProb	4 Prior	3 Prior	2 Prior	1 Prior	AvgForStud	AvgOverall
Quit	0.305556	0.402778	0.305556	0.037037	0.140576	0.137844
Solved	0.916667	0.858333	0.916667	0.983333	0.79724	0.797784
Wrong	0.640152	0.575758	0.916667	1.7875	0.875367	0.84471
Hint	0.068182	0.151515	0.015152	0.5	0.08626	0.103063

In the open-response condition we find that if anything students who report annoyance appear to be more likely to quit on problems a few problems before a self-report which may indicate that students are dealing with bugs in the system. Again, we see a slightly higher proportion of incorrect attempts immediately before self-reports of frustration.

Table 47 Problems prior to Report of Annoyance (N = 15)

TotalProb	4 Prior	3 Prior	2 Prior	1 Prior	AvgForStud	AvgOverall
Quit	0.523737	0.270952	0.368333	0.151111	0.17878	0.163554
Solved	0.556838	0.774206	0.693056	0.874074	0.788842	0.804615
Wrong	0.653419	0.837698	0.790972	0.930556	0.848011	0.766501
Hint	0.109402	0.097024	0.302083	0.162037	0.112287	0.132616

Finally, the full text of students' own causal attributions for their reported feelings of annoyance/frustration are worth noting. In appendix X attributions which can be described by four common categories are listed: attributions to bugs/repeat questions (N=14), attributions to disliking MathSpring (N=8), attributions to disliking Math (N=5), and attributions to excessive difficulty (N=5). It may be that the attributions to the repeated questions refer to the self-reports themselves, and that students feel annoyed/frustrated due to being asked to self-report. While prior work (Wixon & Arroyo, 2014) has not shown a link between self-report frequency and negative emotions, it's possible that the open-response prompts may generate more negative responses.

6.9.2 “Bored”, “IDK”, “DTG”, Blank Responses: A closer look at potential Disengagement

Previously we found that for the forced choice condition students who reported boredom were likely to continue reporting boredom, however students were not significantly likely to go from annoyed to bored. Here we examine what states precede and follow boredom instance by instance, and look to see if boredom is a persistent mood which exists between a trait of a student, and a momentary state, instead looking at how linked boredom may be to behavior at the level of the mood of particular students on particular days.

We then examine reported boredom to see if it increases due to measures of fatigue such as after spending longer in the MathSpring tutoring environment or after solving larger numbers of problems. Students may also have a reporting bias regarding fatigue: we apply the same tests to determine if time in tutor or number of problems solved relates to whether or not students choose to answer self-report prompts for both the forced and open conditions.

Finally, we explore the reports of “IDK” and “DTG”: while coders saw them as being fairly closely related initially we have noted that they appear to be correlated with different performance and self-report measures of goal orientation. Students who report “IDK” when asked how they are feeling often attribute this lack of knowledge to boring material, does boredom lead to “IDK” or vice versa? As to the state of “DTG” we have noted that students who tend to report “DTG” responses when asked how they are feeling may solve more problems on the first attempt. Is this a case of disengaged students allowing themselves a break so that they may re-engage as hypothesized in prior work (Sabourin et al., 2011)?

6.9.3 What Precedes/Follows Boredom? Is it a mood? Does it increase over a session?

Boredom has already been examined earlier in section 6.3: we identified that for the forced-choice condition reports of boredom are most likely to lead to boredom. For the open response

condition findings were inconclusive: there was no statistically significantly more likely state to follow boredom. However, as we can see in Table 48 below tags of “bored” make up the largest minority of tags which precede or follow tags of “bored”, there are simply not enough cases to achieve statistical significance.

Table 48 Reports which precede/follow reports of Boredom

	Prior to Boredom	Following Boredom
Beginning/End of Day	8 (25%)	9 (27%)
annoyed	4 (13%)	1 (3%)
Bored	7 (22%)	7 (21%)
Confused	1 (3%)	0
DTG	1 (3%)	1 (3%)
IDK	1 (3%)	1 (3%)
Negative	1 (3%)	5 (15%)
Neutral	3 (9%)	2 (6%)
Positive	3 (9%)	3 (9%)
Blank	3 (9%)	4 (12%)

Out of 29 open response reports of boredom only 7 were preceded/followed by boredom. This makes them the largest minority aside from the 8 cases where there were no prior feelings reported due to the start of a session and 9 cases where there were no following feelings due to the end of a session. For the forced choice self-reports, reports of low interest (Likert <3) were likely to be followed by reports of significantly below average interest student by student, suggesting that boredom is likely a persisting emotional state as prior work has shown (D’Mello & Graesser, 2011).

At the student level these students who report boredom have more incorrect attempts per problem ($p < 0.1$) and interest seems to be correlated with solving problems correctly on the first attempt ($p < 0.1$). Perhaps the reason these results of the open-response condition differ from the

forced-choice condition and those of D’Mello and Graesser (2011) is that students are feeling consistently bored, however they’re not voicing that experience. Of the total student sample (N=85) only 30 students ever report boredom, at the student level there are marginally significant correlations between measures of poor performance and reported boredom and disinterest. However, if we control for a given class period, these effects may become more pronounced. Perhaps rather than boredom being a trait of a given student students who are bored tend to remain bored throughout a class period, but don’t report this experience of boredom.

The following tests were performed to see if the correlation between boredom (interest/disinterest in the case of forced choice) becomes more significant or has a larger effect size if rather than focusing on the student level (see section 6.3.4) for relationships between these features and performance measures solved on first attempt (SOF) or number of wrong attempts per problem, we instead consider each student on a given day as a distinct instance.

In Tables 49 and 50 below boredom is dependent on a student’s feelings on a given day we would expect that this new “by Day” level to yield higher adjusted R squared values and greater significance.

Table 49 Interest/Boredom SOF per Problem

	R	Sig	N	Adjusted R Square
Forced Choice Interest	0.310	0.059	38	0.071
Forced Choice by Day Interest	0.301	0.008	76	0.078
Open Response Boredom	0.004	0.979	39	-0.027
Open Response by Day Boredom	0.029	0.722	150	-0.006

Table 50 Interest/Boredom Wrong per Problem

	R	Sig	N	Adjusted R Square
Forced Choice Interest	-0.135	0.418	38	-0.009
Forced Choice by Day Interest	-0.305	0.007	76	0.081
Open Response Boredom	0.274	0.092	39	0.05
Open Response by Day Boredom	0.179	0.029	150	0.025

The above results are inconsistent. In Table 49 above we see negligible change in effect size and significance. Table 50 shows a marked increase in effect size and significance in the forced choice by day interest however for the open response condition results are again inconsistent, while significance increases effect size decrease. The inconsistent results do not seem to indicate that boredom occurs as a mood on particular days.

This then leads to the hypothesis that boredom may occur due to fatigue within the day. It's reasonable to expect that if students were becoming bored over time then reports of boredom would after students had spent more time in the tutor on a given day or completed more problems within the tutor on a given day.

An independent samples T-Test was employed to test the above hypothesis comparing reports of boredom against reports of something other than boredom (but not blank responses) for students who ever reported boredom (Table 51 below). There were no significant differences for these students in terms of when boredom was reported as opposed to any other emotional state.

Table 51 Bored vs Not Bored Over Time

	Bored	Not Bored	T-Test for Equality of Means
Avg Probs	21.05	20.51	p=0.859
Avg Time (sec)	1027.06	1079.17	p=0.636
N	60	102	

6.9.4 What leads students to leave self-report prompts blank?

By finding what sorts of forced-choice reports students gave before or after choosing to leave a self-report blank we intended to find if students were going from reported feelings of disinterest into choosing to not report. Indeed, we found a higher rate of disinterest (25.6% Likert <3) and

low excitement (29.7% Likert <3) prior to choosing not to report. However, choosing to not report most frequently preceded and followed itself at rates ranging from 65% to 79%.

Table 52 Forced-Choice Reports which precede/follow reports of Blank Reports

	Prior to Confidence	Prior to Excitement	Prior to Frustrated	Prior to Interest	After Confidence	After Excitement	After Frustrated	After Interest
Blank	0.673	0.595	0.651	0.698	0.745	0.778	0.789	0.781
1	0.109	0.189	0.163	0.209	0.064	0.111	0.132	0.188
2	0.018	0.108	0.047	0.047	0.043	0.028	0	0
3	0.036	0.081	0.023	0.023	0.021	0.028	0.026	0.031
4	0.073	0	0.047	0.023	0	0.028	0	0
5	0.091	0.027	0.070	0	0.128	0.028	0.053	0
Total	55	37	43	43	47	36	38	32

For the open-response condition leaving a self-report blank was most likely to precede or follow leaving a report blank once again at a similar rate of 66%. However, students did not often describe their emotional state as bored prior to choosing not to respond. The most frequent prior response was one of neutral emotions (Table 53).

Table 53 Open Response Reports which precede/follow reports of Blank Reports

	Prior to Blank	Following Blank
Beginning/End of Day	34 (11%)	71 (24%)
annoyed	9 (3%)	3 (1%)
Bored	4 (1%)	3 (1%)
Confused	3 (1%)	1 (<1%)
DTG	6 (2%)	3 (1%)
IDK	4 (1%)	4 (1%)
Negative	9 (3%)	4 (1%)
Neutral	17 (6%)	4 (1%)
Positive	13 (4%)	6 (2%)
Blank	229 (66%)	267 (66%)

When we apply the same t-test as we did in Table 51 we find that here students do indeed seem to be more likely to leave self-report prompts blank as the session progresses (see Table 54). The average number of problems completed at the time of a blank self-report is 30.93, whereas completed emotional self-reports occur at a mean of 24.27 problems. The same follows for the amount of time spent in the tutor: the average time of a blank self-report occurs 290 seconds later into a session than a self-report where students complete an emotional self-report. Each of these differences in mean is statistically significant, suggesting that the longer students work in MathSpring and the more problems they solve the less likely they are to respond to self-report prompts.

Table 54 Blank vs Not Blank Over Time

	Blank (average)	Not Blank (average)	T-Test for Equality of Means
Avg Probs	30.93	24.27	p<0.001
Avg Time (sec)	1555.82	1265.75	p<0.001
N	491	898	

This leads to the largest group of paired emotion/attribution pairings: “neutral” report of emotion and “IDK” attribution (see Table 37). Students who report neutral feelings (e.g. “ok”) are very likely to continue reporting a neutral emotional state (see Table 55). They’re also likely to progress on to leaving self-reports blank after several self-reports. These students might not see much value in making self-reports and therefore more likely to leave them blank.

Table 55 Open Response Reports which precede/follow reports of Neutral Emotion and IDK attribution

	Prior to Blank	Following Blank
Beginning/End of Day	12 (27%)	7 (15%)
annoyed	0	1 (2%)
Bored	1 (2%)	0
Confused	0	1 (2%)
DTG	0	0
IDK	1 (2%)	1 (2%)
Negative	0	0
Neutral	25 (56%)	23 (50%)
Positive	3 (7%)	2 (4%)
Blank	3 (7%)	11 (24%)

In the case of students whose emotional self-reports were classified as “DTG” or disengaged from task we found that again students who report a particular emotional state are likely to continue reporting that same emotional state (see Table 56). In this case students who report “DTG” are likely to continue reporting “DTG”. However, these students also often progress on to leave self-reports blank.

Table 56 Reports which precede/follow reports of “DTG”

	Prior to Blank	Following Blank
Beginning/End of Day	6 (24%)	3 (12%)
annoyed	1 (4%)	0
Bored	1 (4%)	0
Confused	0	0
DTG	8 (32%)	8 (32%)
IDK	0	0
Negative	2 (8%)	2 (8%)
Neutral	0	0
Positive	4 (16%)	5 (20%)
Blank	3 (12%)	7 (28%)

Prior work hypothesizes that students may become disengaged as a means of overcoming boredom and re-engaging with a given task (Sabourin et al., 2011). Some indicators of disengaged behavior include help abuse and rapid guessing (Baker et al., 2004). Rather than operationalize these here, note Table 57 below which identifies significant (by paired samples t-

tests) differences between students' behaviors during (immediately prior to report), prior to (at prior report), and following (at following report) "DTG" reports as compared to the same students' average values of these measures.

In addition to students requesting below average amounts of hints prior to a self-report we also find that students appear to spend consistently more time to complete given problems before during and after "DTG" reports. The difference in mean is fairly large, and unfortunately may be attributable to each of these measures being taken at a problem containing a self-report.

Table 57 Mean difference between behaviors coincident with, prior to, and following reports of "DTG" with paired samples t-test significance (p)

	Current	Prior	Following
Better than Guessing	0.01 (0.77)	0.05 (0.28)	0.03 (0.46)
Hints	0.28 (0.40)	-0.08 (<0.05)	-0.05 (0.42)
Solved	-0.03 (0.75)	0.00 (0.97)	0.12 (0.14)
Quit	0.06 (0.55)	0.03 (0.77)	-0.09 (0.23)
Wrong Attempts	0.04 (0.90)	-0.11 (0.68)	-0.06 (0.78)
Seconds to Complete	135.30 (<0.01)	72.33 (<0.05)	103.14 (<0.05)

7 Discussion

7.1 Limitations

This study was originally proposed to involve at least 300 participants from a diverse set of learning environments. While the data comes from a fairly diverse public school, we only collected data from 85 participants working with a single teacher, within a single grade (8th grade). This lack of variety and number of participants prevents claims of generalizability from being made about these data, which is a particular weakness. Additionally, the method of soliciting reports of emotion and then reports of causal attributions for those emotions made students' reports particularly specific and reduced the effective sample size of students who gave a particular emotion/attribution pairing.

In Table 37 the total instances for a given pairing of emotion and attribution is fairly small and subject to being too influenced by a single student who may be responsible for a majority of those emotion/attribution pairings.

Further, because students are drawn from classes as taught by the same teacher, significant findings in the data might not generalize outside of Massachusetts or even outside of the school or grade level in which they were collected. The small sample size and lack of diversity undermine the modest results of this study; correcting this weakness is particular difficult as most of the work of this study based directly upon the small sample size. This study may only be taken as a single example of what actual teenager students in the United States are able to report for the ways they feel and their reasons for it.

7.2 Students' Perspectives in Self-Report

This work was meant to find answers to the central question, "Are we as researchers using the right categories of emotion/attribution?" by gathering data from student participants to describe

their personal experiences while working within MathSpring. The intent was that by offering open text responses students might offer unexpected responses coming from an alternative perspective to researchers. Further, students' self-descriptions of their experience might be more closely linked to actual student behaviors inside of the tutor, as their causal attributions could determine important associations between observable events and internal emotional states.

In actuality, the most striking aspect of students' self-reports of their emotional states were their simplicity. Roughly twice as many self-reports were classified into the three valence categories (N = 234) "positive", "negative", and "neutral" than more subtle categories (N = 120) of "annoyed", "bored", "confused", "DTG", and "IDK" (see Table 31). When asked generally about their feelings, students appear to be more likely to describe them in terms of simple valence rather than expressing annoyance, boredom, or confusion. This suggests that, when considering self-reports as a way to assess student emotion, simply providing students with a means of expressing their degree of pleasure/displeasure could address the most prevalent emotional states which students' would choose to report if given an open prompt.

Students' causal attributions for their emotional states were more varied and detailed than their emotional self-reports, further, these attributions were often redundant for particular tags associated to emotional self-reports. For example, in addition to describing feeling "bored", students would attribute their feelings to "boring" experiences; in addition to describing feeling "confused" students would use "hard" material as a causal attribution (see Table 37). There were also many cases where students experiencing negative valence emotions used the attribution prompt to affirm their negative feelings even more strongly.

This approach of simply considering students' emotional valence and then associating that with causal attributions and proximal behaviors provides a clearer narrative of students' feelings with regard to their work.

7.2.1 Positive Valence

Students often attribute pleasure (positive valence) to achieving performance goals (i.e. "success") and to not feeling challenged by their work (i.e. "easy" work). This can be found both at the student level (see Tables 36 and 37) and at the fine grained level (see Table 44, Hypothesis 4B). However, reporting positive emotions did not mean students were significantly more successful at solving math problems (see Table 35) unless students were specifically in the forced-choice condition and reporting their degree of confidence (see Table 35) suggesting that students report positive feelings because of specific experiences of success within MathSpring rather than overall successful performance.

7.2.2 Neutral Valence

Students who reported neutral feelings were especially likely to attribute their feelings to attributions of "IDK". Most people, might find it difficult to explain *why* they have no strong feelings at a given time. Students who reported neutral valence were also fairly likely to attribute their lack of strong feelings to "boring" or "negative" causes. These students may be particularly disengaged from responding to self-report prompts. Students who reported neutral valence emotions with "IDK" attributions were very likely to report neutral valence emotions again (Table 55) and were twice as likely to leave the following report blank than students who reported feeling bored (see tables 55 and 48).

7.2.3 Negative Valence

The only behavior variable significantly correlated with frustration, annoyance, or negative emotions at the student level was time per problem. Students who reported negative valence

emotions spent less time per problem (see Table 35). Students often attributed their negative valence emotions to the Material or MathSpring itself (see Table 39) in the case of “annoyed”, “negative”, High frustration, low excitement, and low interest. This relationship continues at the fine grained level (see Table 44) where there are no significant relationships between poor performance and criticisms of the material or learning environment (see Table 44, Hypotheses 6A & 6B) nor between poor performance and frustration (see Table 44, hypothesis 6 ALT). Students who reported frustration (Table 46) or annoyance (Table 47) showed some signs of having quit more problems than they would on average shortly before these reports, suggesting that bugs within MathSpring may have led to their feelings of frustration in accordance with their attributions of MathSpring causing their negative valence emotions. Students’ attributing their negative emotions to the material or the MathSpring itself may explain why students who reported frustration were more likely than average to remain frustrated (see Table 47 hypothesis 3 ALT) and students who reported boredom were more likely to remain bored (see Table 47 hypothesis 5). If the causes of students’ negative emotions are mathematics or MathSpring then changes in their performance while solving math problems within MathSpring are unlikely to affect their emotional state.

7.3 Priming Effects: Differences between Open Response and Forced Choice

As discussed in the introduction, many emotions related to learning can be described as epistemic emotions. Again, epistemic emotions are sometimes described as emotional or cognitive and are characterized as being partly dependent on events or cognition. This indistinct division between cognition and emotion may contribute to the differences we see differences between the forced choice and open response conditions. For example, confidence may be implied to be a result of success rather than simply having a positive outlook on one’s work. This

is evident in Table 35, students who report confidence are likely to solve more problems on the first attempt and get fewer problems wrong, whereas students who report positive emotions do not display these correlations.

This may also explain why attributions for the open-response emotional tag of “bored” differ from the attributions for the forced choice emotions of low excitement or low interest (Likert < 3) as is shown below in table 58.

Table 58 Boredom & Low Excitement/Interest vs Attributions as Percentage of total

	Boring	DTG	Easy	Hard	IDK	Matl.	Negative	Success	Website
Low Excitement	13%	9%	9%	4%	4%	10%	32%	0%	18%
Low Interest	12%	8%	12%	4%	4%	10%	34%	0%	16%
Bored	24%	0%	13%	3%	16%	16%	21%	0%	8%

Reports of “Bored” were twice as likely to have the attribution boring as reports of low excitement or interest. They were also less likely to include negative attributions or attributions of the website, they were far more likely to have “IDK” attributions. A possible explanation for this is that when given a forced choice prompt of excitement or interest students may wish to express negative emotions that are not disinterest or boredom. If a forced-choice prompt is the only lever a student is given it is possible that the responses actually capture the far more prevalence valences of positive/negative.

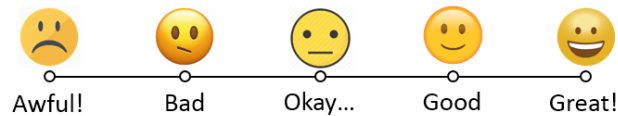
A simpler illustration of this is that in the open response condition students chose to address boredom on 28 instances out of the total set of self-report opportunities whereas for the forced-choice condition when students were only given half as many opportunities to address their excitement or interest they made 68 reports of low excitement, and 50 reports or low interest.

Admittedly, the “priming effect” is only a hypothesis of why we might see differences between the attributions reported for forced-choice and open-response self-reports. It is also possible that students might want to report boredom and it simply doesn’t occur to them to use terms that would suggest boredom. For example, students might simply use neutral emotions and “IDK” as a potential way of expressing boredom.

7.4 Improved Forced-Choice Self-Report

Given the prevalence of valence based emotional self-reports, and how most other emotions reported could be described as epistemic centered upon cognitive processes (e.g. annoyed with bugs in the system or being asked the same question over and over, bored with the material, or confused due to material being excessively challenging), an improved self-report could account extreme measures of valence while also allowing students to report common epistemic emotions and attributions.

I feel...



Because I'm...

- Annoyed
- Working on [MathSpring](#)
- Not Challenged
- Doing Math
- Getting problems right
- Bored
- Learning
- Confused
- Getting problems wrong
- Thirsty, Hungry, Tired, warm, cold, or needing the bathroom
- None of the above reasons

Figure 13 Proposed Future Self-Report Prompt

This new prompt provides (see Figure 13) students first with a sliding scale to report the valence of their emotion, and then with several possible emotions/attributions that can give greater context to their emotional valence. For example, a student might be bored and simply have a neutral “Okay...” valence of emotion or might be unbearably bored and report “Awful!” as the valence of their emotion. In this case students have access to all options and can also read each option to allow them to better describe their emotional state. Or select none of the option if they have no better explanation for the valence of their emotional state.

These prompts would require students to type less and they would cover the most common sorts of emotions/attributions which students report. Further, we have a means for students to express extreme valence in their emotions to identify when students may be indifferent rather than having strong positive or negative feelings associated with their more detailed report.

7.5 Extending this Work

In closing, the question remains as to whether there are any worthwhile avenues of research left unexplored with these data. While more exploratory analyses could be conducted at this point, the utility of conducting these analyses must be considered. As previously stated under section Lack of Participants, these data may be too few to justify more detailed analyses. We have only N=39 students who responded to open-response measures, and N=40 who responded to forced-choice measures. This limited data set must be accounted for in considering extensions to this work, starting with the potential strategy of collecting more data.

7.5.1 Collecting Additional Data

One possibility to extend this work would be to continue data collection. Collecting additional data will likely require an additional year to make contact with a teacher, schedule additional sessions with this teacher, and administer MathSpring again. After the data is collected we must

then consider where to begin in coding these new data: initial open-response coding, inter-rater agreement testing, or simply having the first author apply tags to this new data set. Starting from the first step, initial open-response coding would be the most time intensive step. It would require that we consider changing the existing coding scheme to accommodate possible new tags, which could be present in the new cohort of participants. In addition to experienced coders, new coders would have to be found who are unfamiliar with the existing coding scheme, they would have to tag the new responses, and then schedule time to discuss their coding schemes. The current data set required 3 months during the summer of 2017 to coordinate all volunteers' schedules, so it is reasonable to plan that the new coding scheme would require another 3 months.

Starting from the second step, inter-rater reliability testing, would require less time. Two coders would have to tag the new data set given the existing lexicon of tags. This would require at least one volunteer to make time for this project. Further, it's possible that despite following the existing coding scheme the coders could identify responses that would deviate from this lexicon. That would require re-considering adding or removing tags from the lexicon. While this work could extend, it is reasonable to plan that this re-application and interrater reliability check would take at least a month.

Finally, if the existing scheme were simply applied by the first author, the approach would take only a week. However, as previously stated, collecting a new data set could take up to a year.

7.5.2 Additional Analyses: Detector of Affect

A detector of affect for these data could be constructed within a month-time; however, we should consider whether such a detector would likely make a significant contribution to the field.

Detectors of affect have been built using similarly small data sets (Wixon et al, 2014) and performed relatively poorly as compared to detectors built using larger data sets. However, for the open-response condition, at least there are more potential emotional self-reports available. This means that detecting each possible affective state is more challenging because the additional specificity reduces the total possible cases to be considered.

The problem of highly specific tags could be addressed by considering only tags & tag combinations which met a minimum sample size criteria: a set minimum for the number of reports and the minimum number of students who report particular tags. Then only these more common tags would be considered. However, this still leaves the additional problem of bias due to all the analyses that have been conducted up to this point.

The approach used up to this point has been to test specific research questions using basic statistical tests regarding which events precede/follow each other. This approach has undermined the potential for building an unbiased detector: having already looked in detail at which events appear to precede particular reports means that the author has observed what features will likely predict self-reported emotions.

Finally, we should consider that the resulting affect detector would be less of a means of detecting self-reported affect in an unbiased manner than a means of describing the data and quantifying what features would have the greatest impact on students' self-reports.

7.5.3 Additional Analyses: Structural Equation Modeling

Structural equation modeling might not be a suitable approach for these data due to the particularly small sample size. As a general guideline sample sizes $N < 200$ are often excluded from SEM analyses (Boomsa, 1982). Given that we have only $N=39$ students who responded to open-response measures, and $N=40$ who responded to forced-choice measures we are far below

the expected minimum. Further, if we consult Table 44 we see that many specific constructs are only expressed by a very small subset of the total sample group of participants. Note the number of tests that were not performed because fewer than 8 participants would be counted in the given T-Test. Of the tests we could run the total sample sizes are quite small compared to the recommended total of $N > 200$. However, if more data were available, Structural Equation Modeling would be an interesting further avenue to explore, to analyze a tier of pretest incoming variables that describe the student, to behaviors and emotions inside of the tutoring system, to a third tier of post-tutor and outcome variables.

8 Conclusion and Future Work

This dissertation's initial goals were largely exploratory rather than confirmatory. Much of this work was predicated on the idea that by asking students to report their feelings, attributions, and desires relating to a learning environment in an open way and then forming categorical tags based on the content of their responses we might uncover a glimpse of how students' view their own experiences. This exploratory approach is predicated on avoiding the risks of arriving at a learning environment with a set of constructs we might ordinarily expect to find, in order to avoid running three risks: 1) expecting to find a construct (e.g. work avoidance, or boredom) which then turns out to be absent; 2) creating a previously absent construct by asking leading questions; or most importantly 3) missing an important construct which is present because our survey measures do not consider this construct initially. This goal led to a major contribution of this dissertation: designing a measurement tool to collect open-response self-reports and moreover to classify them by tags without necessarily imposing prior theory upon these data.

This work has achieved that goal of creating a measurement tool to capture students' self-reports that can avoid theoretical bias and be applied across culturally distinct populations. The open-response prompts are neutral and allow for responses that may be outside of our prior expectations. The method of coding students' responses begins by having all coders create and apply their own distinct lexicons of tags, and then find consensus among coders. While the method is borne out of well known methods of testing inter-rater agreement, the novelty resides in that coders are not given a set of constructs or prescribed measurement protocol. This method encourages the generation of independently created tags that values separate researchers' subjective understanding of students' responses. In this specific case, Cohen's kappa could arguably be called a measure of validity rather than reliability. The invention of a method for

creating construct independent coding schemes, which is adaptive to new populations and new learning environments, is the main contribution of this work.

Despite its strengths this method also comes with considerable weaknesses. Firstly, it is time and labor intensive requiring several coders to carefully review large amounts of data in a more cognitively taxing manner than traditional text coding which might come with a prior list of available tags or constructs to look out for. Then there is the time required to process those coders' data to find agreements between coders and look for emergent constructs. Fortunately, the multi coder inter-rater agreement program (Appendix H) managed much of the labor and human error inherent in finding the rate of agreement between coders, and is acts as another contribution of this work. While, it was designed for these data it could be applied by researchers in any context where several coders' responses must be compared for agreement, the fact that it is designed to robustly handle cases where different coders may be working from a different set of terms only makes it more valuable. For example, redundancy between two previously validated coding schemes could be tested to see the degree of overlap for two similar constructs (e.g. boredom & disinterest, or flow & engaged concentration) as applied by two distinct research groups.

A second and more significant next step to investigate, and somewhat a new concern that results from this work, is that it is not immediately obvious how it will improve students' learning experience or learning outcomes. Tentative future work should investigate how to "close the loop" between affect detection and pedagogical support that targets the enhancement of students' affective and cognitive states. I began with the hypothesis that by accounting for students' attributions of the causes of their emotions within the learning environment it would be possible to build better detectors of those emotions. In this hypothesis students model their own

interactions with the learning environment and their causal attributions of their emotions may refer to events that occur within that learning environment. Therefore these attributions could act as a proxy between events and emotions, which might improve the accuracy of affect detectors.

Furthermore, it is highly possible that students' open responses might help to better address student needs that had not previously been considered. In particular, I made special point to include the question "What do you wish you could do to improve this class right now?" in the open assessment within-tutor prompts. These data have not yet been analyzed in detail, and further students' beliefs about what would improve their learning environment may not actually result in learning gains or even greater enjoyment. However, my work has been designed to ultimately address students' needs by first listening to these needs from the students' own perspective. While improvements in pedagogy are important, I have largely limited my focus to exploratory measures meant to best capture students' own perspectives.

I have tried to separate my own hypotheses about the realities of students' experiences from my collection methods, so that students' naïve perspectives might remain as salient as possible throughout data collection and coding. Despite these efforts, my open-response prompts have likely impacted students' experience within the MathSpring learning environment. Firstly, because in-vivo survey measures interrupt workflow (Ocumpaugh et al., 2015). Specifically, some students may find the self-report measures themselves to be annoying: anecdotally in at least one instance when asked how they were feeling a student responded "stop asking me" this is consistent with concerns regarding negative rumination. It has been empirically shown via heart rate and cardiac output that the act of repeatedly reporting on one's own level of anger may have negative results (Kassam & Mendes, 2013). Additionally, I must acknowledge the reporting bias where I have shown that students who have spent more time in tutor or seen more

problems are more likely to leave self-report prompts blank (section 6.9.4). These measures are not passive; they influence students' responses in terms of their feelings and likelihood to respond in the first place.

Returning to the goal of comparing the content of students' open-response self-reports to the constructs included in forced choice prompts (confidence, excitement, frustration, & interest), I found that students responded in terms of simple valence (positive, negative, & neutral) most of the time. This is consistent with research on younger students who use broad categories to describe emotions in observed facial expressions (Widen & Russell 2003, 2008; Bullock & Russell 1984, 1985, 1986). This research is not conclusive to affirm that the students in this study do not possess a more nuanced understanding of epistemic emotions. Students responded with richer descriptions when asked to attribute the causes of their emotions --in many cases this cognitive attributions component captured the more subtle aspects of emotional constructs such as confusion or boredom.

This separation between valence and causal attributions solves an issue found in both this and prior work (Schultz et al., 2016), the union of self-reported emotions with attributions produces a large combinatorial amount of categories. While it could be argued that this increased specificity would allow us to better serve students' needs, having more categories to detect with fewer instances increases the difficulty of building a machine learning detector. For purposes of building a detector of emotion, simply having a self-report prompt that addresses valence (figure 13) and lets students select which common attribution best fits their situation. The only change that should be made would be to add a final box allowing students to fill in an open-response if none of the choices are appropriate. I had not expected to find the attributional category "needs"

when I began this work, but addressing immediate physical needs in the classroom is important and would have fallen through the cracks otherwise.

Finally, while I am advocating for a simplified self-report measure, which includes forced-choice measures, I maintain that the open-response coding methods should be employed when working with students from a new cultural background or within a new learning environment. Students' perspectives will likely continue to differ from our expectations and it remains important to continue listening to their needs, particularly in learning environments we find unfamiliar.

9 References

- Adolphus, K., Lawton, C. L., & Dye, L. (2013). The effects of breakfast on behavior and academic performance in children and adolescents. *Frontiers in human neuroscience*, 7, 425.
- Arroyo, I., Shanabrook, D. H., Burleson, W., & Woolf, B. P. (2012, June). Analyzing affective constructs: emotions, attitudes, and motivation. In *Intelligent Tutoring Systems* (pp. 714-715). Springer Berlin Heidelberg.
- Arroyo I., Wixon N., Alessio D., Woolf B., Muldner K., Burleson W. (2017) Collaboration Improves Student Interest in Online Tutoring. In: André E., Baker R., Hu X., Rodrigo M., du Boulay B. (eds) *Artificial Intelligence in Education. AIED 2017. Lecture Notes in Computer Science*, vol 10331. Springer, Cham
- Arroyo, I.; Woolf, B. P.; Cooper, D. G.; Burleson, W.; and Muldner, K. 2011. The Impact of Animated Pedagogical Agents on Girls' and Boys' Emotions, Attitudes, Behaviors, and Learning. In *Proceedings of the 11th IEEE Conference on Advanced Learning Technologies*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004, August). Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems* (pp. 531-540). Springer, Berlin, Heidelberg.
- Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241.
- Bieg, M., Goetz, T., & Lipnevich, A.A. (2014). What Students Think They Feel Differs from What They Really Feel—Academic Self-Concept Moderates the Discrepancy between Students' Trait and State Emotional Self-Reports. *PLoS ONE* 9(3): e92563.
- Block, C. J., Roney, C. J. R., Geeter, J., Lopez, P. D., & Yang, T. (1995, August). The influence of learning and performance goal orientations on anxiety, motivation and performance on a complex task. Paper presented at the 55th annual meeting of the Academy of Management, Vancouver, Canada.
- Boomsma A. Robustness of LISREL against small sample sizes in factor analysis models. In: Joreskog KG, Wold H, editors. *Systems under indirection observation: Causality, structure, prediction (Part I)* Amsterdam, Netherlands: North Holland; 1982. pp. 149–173.

- Boud, D., Keogh, R., & Walker, D. (1996). What is reflection in learning. In D. Boud, R. Keogh & D. Walker (Eds.), *Reflection: turning experience into learning* (pp. 7-17). London: Kogan Page.
- Broekens, J., & Brinkman, W. P. (2013). AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6), 641-667.
- Bull, S., Brna, P., & Pain, H. (1995). Extending the scope of student models. *User Modeling and User-Adapted Interaction*, 5(1), 45-65.
- Bull, S. and Kay, J. (2007) 'Student models that invite the learner in: the SMILI:-) open learner modelling framework', *International Journal of Artificial Intelligence in Education*, Vol. 17, No. 2.
- Bullock, M., & Russell, J. A. (1984). Preschool children's interpretation of facial expressions of emotion. *International Journal of Behavioral Development*, 7, 193-214.
- Bullock, M., & Russell, J. A. (1985). Further evidence on preschoolers' interpretations of facial expressions of emotion. *International Journal of Behavioral Development*, 8, 15-38.
- Bullock, M., & Russell, J. A. (1986). Concepts of emotion in developmental psychology. In C. E. Izard & P. B. Read (Eds.), *Measuring emotions in infants and children* (Vol. 2, pp. 203-237). Cambridge, England: Cambridge University Press
- Butler, R. (1993). Effects of task- and ego-achievement goals on information-seeking during task engagement. *Journal of Personality and Social Psychology*, 65, 18-31.
- Carney-Crompton, S., & Tan, J. (2002). Support systems, psychological functioning, and academic performance of nontraditional female students. *Adult education quarterly*, 52(2), 140-154.
- Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context*, 1, 39-64.
- Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences*, 11(9), 393e399.
- Clore, G. L., & Ortony, A. (1988). Semantic analyses of the affective lexicon. In V. Hamilton, G. Bower, & N. Frijda (Eds.), *Cognitive science perspectives on emotion and motivation* (pp. 367-397). Amsterdam, The Netherlands: Martinus Nijhoff.
- Clore, G., and A. Ortony (2000). "Cognition in emotion: Always, sometimes, never?" In R. D. Lane and L. Nadel (eds.), *Cognitive Neuroscience of Emotion*: 24-61. Oxford: Oxford University Press.

- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Collins, A., J.S. Brown, and S.E. Newman (1989) Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics, in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser L.B. Resnick*, Editor Lawrence Erlbaum Associates: Hillsdale, NJ. p. 453-494.
- Conati, C., Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Model. User-Adapt. Interact.* 19, 267–303.
- Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1), 3-21.
- Dimitrova, V. (2003). STyLE-OLM: Interactive open learner modelling. *International Journal of Artificial Intelligence in Education*, 13(1), 35-78.
- 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 Model. User-Adapt. Interact.* 18(1–2), 45–80.
- D’Mello, S., & Graesser, A. (2011). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145-157.
- D’Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145-157.
- Dowson, M., & McInerney, D. M. (2001). Psychological parameters of students' social and work avoidance goals: A qualitative investigation. *Journal of Educational Psychology*, 93(1), 35-42.
- Dowson, M., & McInerney, D. M. (2004). The development and validation of the Goal Orientation and Learning Strategies Survey (GOALS-S). *Educational and Psychological Measurement*, 64, 290–310.
- Dweck, C. (2006). *Mindset: The new psychology of success*. Random House.
- Elig, T. W., & Frieze, I. H. (1979). Measuring causal attributions for success and failure. *Journal of personality and social psychology*, 37(4), 621.
- Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12), 1050-1057.

- Fosnot, C. (1996). Constructivism: A psychological theory of learning. In C. Fosnot (Ed.), *Constructivism: Theory, perspectives, and practice* (pp. 21-40). New York: Teachers College Press.
- Frieze, I. H. (1976). Causal attributions and information seeking to explain success and failure. *Journal of Research in Personality*, 10, 293-305.
- Frieze, I. H., & Snyder, H. N. (1980). Children's beliefs about the causes of success and failure in school settings. *Journal of Educational Psychology*, 72, 186-196.
- Gobert, J., Baker, R., Wixon, M. (2015) Operationalizing and Detecting Disengagement Within Online Science Microworlds. *Educational Psychologist*, 50:1, 43-57.
- Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity*, 18, 81-90.
- Graesser, A., & Olde, B. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, 95(3), 524-536.
- Graesser, A., & D'Mello, S. K. (2011). Theoretical perspectives on affect and deep learning. In *New perspectives on affect and learning technologies* (pp. 11-21). Springer New York.
- Harackiewicz, J.M., Barron, K.E., Tauer, J.M., & Elliot, A.J. (2002) Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94, 562-575.
- Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, 90, 36-53.
- Henry, K. L. (2007). Who's skipping school: Characteristics of truants in 8th and 10th grade. *Journal of school health*, 77(1), 29-35.
- Henry, K. L., Knight, K. E., & Thornberry, T. P. (2012). School disengagement as a predictor of dropout, delinquency, and problem substance use during adolescence and early adulthood. *Journal of youth and adolescence*, 41(2), 156-166.
- Karumbaiah, S., Lizarralde, R., Alessio, D., Woolf, B., Arroyo, I., & Wixon, N. (2017). Addressing Student Behavior and Affect with Empathy and Growth Mindset. *Proceedings of the 10th International Conference on Educational Data Mining.*, 96-103.

- Kassam, K.S., & Mendes, W.B. (2013). The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PLOS One*, 8(6), 649-659.
- Keltner, D., & Shiota, M. (2003). New displays and new emotions: a commentary on Rozin and Cohen (2003). *Emotion*, 3, 86e91.
- Kleinman, R. E., Hall, S., Green, H., Korzec-Ramirez, D., Patton, K., Pagano, M. E., & Murphy, J. M. (2002). Diet, breakfast, and academic performance in children. *Annals of Nutrition and Metabolism*, 46(Suppl. 1), 24-30.
- Kruger, J., Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*. American Psychological Association. 77 (6): 1121–1134.
- Lepper, M. R., & Henderlong, J. (2000). Turning “play” into “work” and “work” into “play”: 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 257–307). San Diego, CA: Academic Press.
- Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement. Introduction to the special issue. *Contemporary Educational Psychology*, 36, 1–3.
- McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective transitions in narrative-centered learning environments. *Educational Technology & Society*, 13(1), 40-53.
- Mentis, H.M., 2007. Memory of frustrating experiences. In: Nahl, D., Bilal, D. (Eds.), *Information and Emotion*. Information Today, Medford, NJ.
- Nielsen, P. A. (1991). Approaches to appreciate information systems methodologies: a soft system survey. *Scandinavian Journal of Information Systems* , Volume 2, University of Aalborg.
- Ocupaugh, 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.
- Ocupaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocupaugh 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.
- Pardos, Z. A., Baker, R. S., San Pedro, M. O., 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. 3rd Int.Conf. Learning Analytics & Knowledge*, 117-124.

- Pedro, M. O., Baker, R., Bowers, A., & Heffernan, N. (2013, July). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Educational Data Mining 2013*.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18, 315-341.
- Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotions in education*. San Diego: Academic Press.
- Pekrun, R. (2010). Academic emotions. In T. Urda (Ed.), *APA educational psychology handbook*, Vol. 2. Washington, DC: American Psychological Association.
- Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102, 531-549.
- Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement* (pp. 259-282). Springer, Boston, MA.
- Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. New York: Viking.
- Popham, J., (2009) *Assessing Student Affect*, *Educational Leadership*, Vol 66, Number 8, PP 85-86.
- Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction*, 18, 125-173.
- Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.S.J.d. (2013) *Knowledge Elicitation Methods for Affect Modeling in Education*. *International Journal of Artificial Intelligence in Education*, 22 (3), 107-140.
- Rodrigo, M. M. T. (2011). Dynamics of student cognitive-affective transitions during a mathematics game. *Simulation & Gaming*, 42(1), 85-99.
- Rodrigo, M.M.T., Baker, R.S.J.d., Agapito, J., Nabos, J., Repalam, M.C., Reyes, S.S. (2010) *Comparing Disengaged Behavior within a Cognitive Tutor in the USA and Philippines*. *Proceedings of the 10th Annual Conference on Intelligent Tutoring Systems*, 263-265.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied*, 80(1), 1.

- Rozin, P., & Cohen, A. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3, 68e75.
- Russell, J., 2003. Core affect and the psychological construction of emotion. *Psychological Review* 110, 145–172.
- Sabourin, J., Mott, B., and Lester, J. Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, pp. 286-295, 2011.
- Sabourin, J., Rowe, J., Mott, B., Lester, J. (2011) When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 534-536.
- 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.
- Schultz, S. E., Wixon, N., Alessio, D., Muldner, K., Bursleson, W., Woolf, B., & Arroyo, I. (2016). Blinded by science?: Exploring affective meaning in students' own words. In *Intelligent Tutoring Systems - 13th International Conference, ITS 2016, Proceedings*. (Vol. 9684, pp. 314-319).
- Silvia, P. J. (2009). Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics Creativity and the Arts*, 3(1), 48e51.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859. doi: 10.1037/0033-2909.133.5.859
- VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3), 209-249.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4), 548.
- Weiner, B. (2010). The Development of an Attribution-Based Theory of Motivation: A History of Ideas. *Educational Psychologist*, 45:1, 28-36.
- Weiner, B., Russell, D., & Lerman, D. (1979). The cognition–emotion process in achievement-related contexts. *Journal of personality and social psychology*, 37(7), 1211-1220.

- Widen, S. C., & Russell, J. A. (2003). A closer look at preschoolers' freely produced labels for facial expressions. *Developmental psychology*, 39(1), 114.
- Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive development*, 23(2), 291-312.
- Wixon, Alessio, D., Ocumpaugh, J., Woolf, B., Burleson, W., Arroyo, I. (2015) La Mort du Chercheur: How well do students' subjective understandings of affective representations used in self report align with one another's, and researchers'? *International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015)*, 34-44.
- Wixon, M., Arroyo, I. (2014) When the Question is Part of the Answer: Examining the Impact of Emotion Self-Reports on Student Emotion. *Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP 2014)*, 471-477.
- Wixon, M., Arroyo, I., Muldner, K., Burleson, W., Lozano, C., Woolf, B. (2014) The Opportunities and Limitations of Scaling Up Sensor-Free Affect Detection. *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)*, 145-152
- Wixon, N., Woolf, B. P., Schultz, S., Alessio, D., & Arroyo, I. (Under Review) Microscope or Telescope: Analyzing Emotions at a Finer Grain. To Appear in *Proceedings of the 11th International Conference on Educational Data Mining*.
- Woolf, B.P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., Christopherson, R.M. (2010) The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. In: Alevan, V., Kay, J., Mostow, J. (eds.) *ITS 2010, Part I*. LNCS, vol. 6094, pp. 327–337. Springer, Heidelberg.

Appendix A: Goals-S Survey

Construct (Goal or Strategy)	GOALS-S Items	Alpha
Mastery: Wanting to achieve to demonstrate understanding, academic competence, or improved performance relative to self-established standards.	D2. I want to do well at school to show that I can learn new things.	0.78
	D5. I want to do well at school to show that I can learn difficult	
	D11. I try hard to understand my schoolwork.	
	D14. I work hard to understand new things at school.	
	D22. I work hard at school because I am interested in what I am	
	D24. I try hard at school because I am interested in my work.	
Performance: Wanting to achieve to outperform other students, attain certain grades/marks, or obtain tangible rewards associated with academic performance.	D3. I want to do well in school because being better than others is	0.87
	D6. I try to do well at school because I am only happy when I am	
	D9. I want to learn things so that I can come near the top of the	
	D12. I want to learn things so that I can get good marks.	
	D15. When I do good schoolwork it's because I am trying to be	
	D18. I want to do well in school so that I am one of the best in my	
Work avoidance: Wanting to achieve with as little perceived effort as possible.	D7. I choose easy options in school so that I don't have to work too	0.72
	D10. At school I want to do as little work as possible.	
	D13. If schoolwork is too hard for me I just don't do it.	
	D16. I don't ask questions in school even when I don't understand	
	D19. I don't do schoolwork if it looks too hard to learn.	
	D23. I want to do well at school, but only if the work is easy.	

Goals-S Survey Items for Mastery, Performance, and Work Avoidance Goals (Dowson & McInerney, 2004)

Appendix B: Emotion Survey Measures

Affect State	Affective Predisposition Item
Interest	How interested do you feel when solving math problems, in general?
Excitement	In general, how exciting is it to solve math problems?
Confidence	How confident do you feel while solving math problems, in general?
Anger	How angry do you get when solving math problems?
Frustration	How frustrated do you get when solving math problems, overall?
Anxiety	How anxious do you get while solving math problems?
Shame	How embarrassed (ashamed) do you get while solving math problems?
Boredom	Do you get bored when solving math problems?
Joy	Do you enjoy solving math problems?
Hopelessness	Do you feel hopeless when you solve math problems? (reverse scale, Very...Not at all)
Pride	Do you feel proud when you solve math problems?

Affect Predisposition Items to Determine Students' Baseline Affect State (Arroyo et al, 2012)

Appendix C: Pretest and Posttest Items Derived from MCAS Standards of the State of Massachusetts

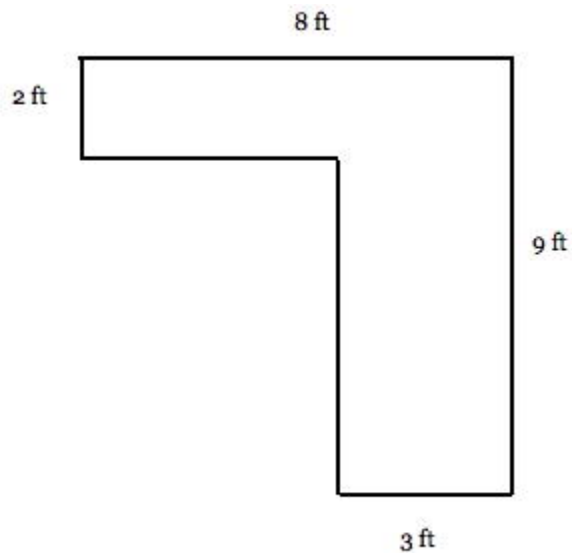
What is 150% of 48?

35) Enter your answer here for the problem ABOVE:

What is:

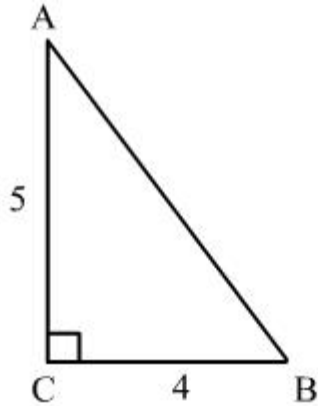
$$\frac{4}{5} \div \frac{1}{3}$$

36) Enter your answer here for the problem ABOVE:



What is the perimeter, in feet, of this figure?

37) Enter your answer here for the problem ABOVE:



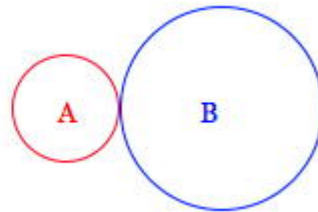
If BC is 4 and AC is 5 , what is the area of ABC?

38) Enter your answer here for the problem ABOVE:

If the perimeter of a rectangle is 6 times the width of the rectangle, then the height of the rectangle is how many times the width?

39) Enter your answer here for the problem ABOVE. (It may be helpful to make a sketch for this one.)

When wheel B turns 2 revolutions, wheel A turns 6 revolutions. When wheel B turns 40 revolutions, how many revolutions does wheel A turn?



40) Enter your answer here for the problem ABOVE:

Find the value of c in the following equation:

$$32 = \frac{1}{3}c$$

41) Enter your answer here for the problem ABOVE:

Thank you for taking our Survey!

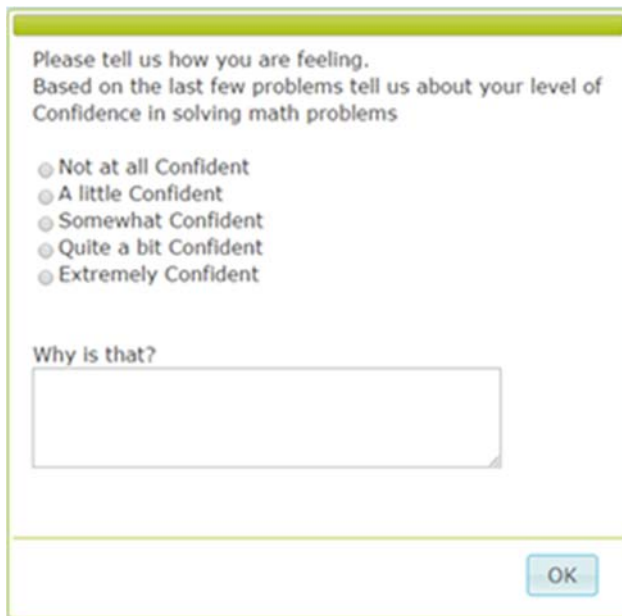
Appendix D: Coder Release Form

Survey Design and Responses: Students solve math problems within an online tutoring environment. Approximately every 5-10 minutes students are asked to answer some simple questions about their feelings. The students are not required to answer these questions.

Students were randomly assigned to receive two different sets of questions. Students were **EITHER** given one set of questions **OR** the other, no students received both sets:

The First Set: began with a multiple choice selection about how strongly they might feel one of four emotions (confidence, excitement, frustration, or interest) and then to explain why (fig 1).

The Second Set: were open ended where students were invited to describe their experiences in text (fig 2).



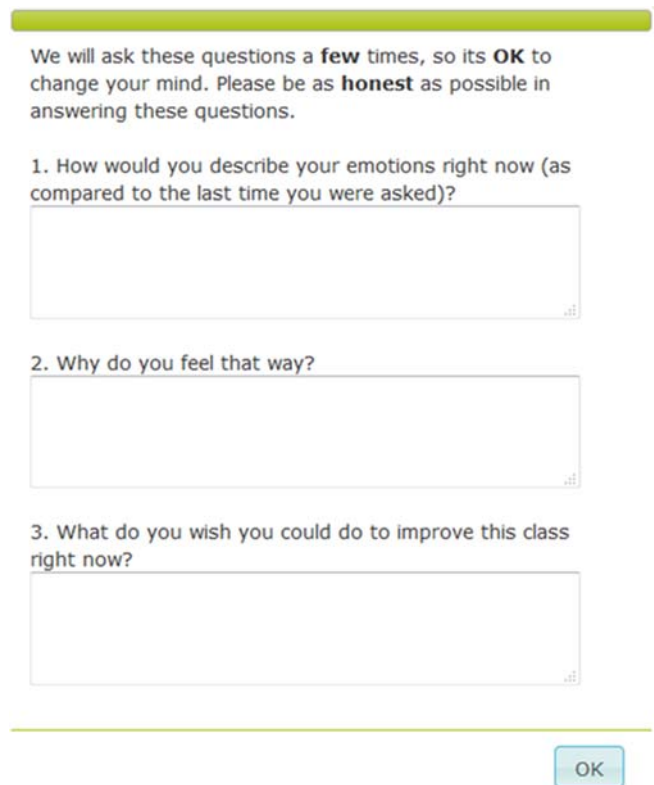
Please tell us how you are feeling.
Based on the last few problems tell us about your level of Confidence in solving math problems

- Not at all Confident
- A little Confident
- Somewhat Confident
- Quite a bit Confident
- Extremely Confident

Why is that?

OK

Figure 1 Multiple Choice Student Survey



We will ask these questions a **few** times, so its **OK** to change your mind. Please be as **honest** as possible in answering these questions.

1. How would you describe your emotions right now (as compared to the last time you were asked)?
2. Why do you feel that way?
3. What do you wish you could do to improve this class right now?

OK

Figure 2 Open-ended Student Survey

Your Task: “Code” students’ responses. In this case coding means to take students descriptions of their feelings and beliefs, and assign sets of one word tags (like hashtags) to those descriptions. As an example of how this process can work, let’s consider what it would be like if we applied a similar process to well-known films. In this case coders would watch the films and then look for simple 1-2 word tags to apply to the films. Let’s suppose we come up with tags for 5 basic qualities of films: magical elements, whether the movie is appropriate for kids, if the movie incorporates some sort of conflict within a family as a plot point, whether the movie involves moving between utterly distinct worlds, and finally if the film includes science fiction elements..

Film\Tag	Magical	Kid Friendly	Family Conflict	Different World	Sci-Fi
The Little Mermaid	X	X	X	X	
The Wizard of Oz	X	X	X	X	
The Matrix				X	X
The Godfather			X		
Lord of the Rings	X		X		
E.T.		X	X		X
Back to the Future			X	X	X
The Lion King	X	X	X		

It’s important that we can get tags that are both very descriptive of the more detailed student responses, but also tags that are not so specific that we create a great number of highly specific tags. For example, “underwater” might fit “The Little Mermaid” very well, but would only apply in a single case. This creates two problems: it will be hard to find films similar to “The Little Mermaid” if there are properties it doesn’t share with any other films, and also that by coding hundreds of films we will find ourselves with hundreds of tags. Ideally, your total list of tags for each question should be in the **single digits 1-9 with an absolute maximum of 12**. You may choose to drop or consolidate tags as you work. For example I might combine “magical” and “sci-fi” into a single tag of “supernatural” if I felt that the story elements were very much the same. If “Family Conflict” continues to describe nearly *every* film I might choose to simply note that over 90% of the films appear to involve a “Family Conflict” and choose to drop the tag for not being a good distinguishing feature of films.

Your work will be considerably easier given that you are looking for tags to apply to 1 sentence long statements by students. Please consider these statements in the context of each other, i.e. an explanation for **why** a student feels a particular way may help you understand how that student feels if the student’s description of their feelings alone appears noncommittal or unclear. You are encouraged to have a different set of tags for each question. The tag list for the question “How would you describe your emotions right now (as compared to the last time you were asked)?” will likely be different from the tag list for “What do you wish you could do to improve this class right now?”.

A later task: Your tag list will be compared with other coders’ tag lists. You may be asked to come back and discuss your coding decisions with other coders, our goal will be to reach a consensus set of tags for

each question that best reflects students' intended meaning. For example, some coders could argue that "Lord of the Rings" involves movement to a "Different World" based on the characters' long journey, or that "E.T." should be classified as a "Different World" if we consider E.T. the extra-terrestrial to **be** the film's protagonist. The goal of this process is to overcome personal biases and defer to the most accurate and descriptive summary of students' intended meaning.

Once consensus is reached you may be contacted again to re-code these data using the new consensus-based coding scheme.

Finally, please check all that apply:

I am an educational psychology researcher

I am an educational data miner

I am a teacher

I do not identify as working now (or having spent a substantial period of time working in the past) in any of the aforementioned professions

I agree to have my codes and responses to these survey questions published in academic work provided my name and identifying information remains anonymous in any publication

I agree to have my codes and responses to these survey questions published in academic work and would like my name to be included in a "Special Thanks" section of Coder N Wixon's dissertation

I, the undersigned, agree to have the data which I provide appear in academic work

Appendix E: Coder Instructions

Here are the types of tags which can be applied to each set of data. The Attribution Tags data set can be applied to both the Open Attributions and the Forced Choice Attributions. The “Agency” tags for easy and hard should mean if the student ***wants*** the material to be easier or harder, but for attribution they should only be if the student ***references*** easy or hard in task difficulty.

Feelings Tags	Attribution Tags	“Agency” Tags
bored	bored	
DTG	DTG	DTG
	easy	easy
	failure	
	growth	growth
confused	hard	hard
IDK	IDK	IDK
annoyed	material	material
neutral	needs	needs
negative	negative	quit
positive	positive	bugs
	success	design
	website	fun

Tag	Meaning	Frequently Occurs With Following Prompts
Annoyed	Refers to Annoyance or feeling Annoyed	Open Feelings
Bored	Refers to Boredom or feeling Bored	Forced Choice Attributions & Open Attributions
Bugs	Refers to parts of MathSpring which appear to be broken or buggy	Open "Agency"
Design	Refers to requesting aesthetic improvements/changes to MathSpring	Open "Agency"
DTG	Disengaged from Task: May appear as random or unrelated to learning task	All Prompts
Easy	Refers to low difficulty level; in "Agency" prompt is a request for easier material	Forced Choice Attributions, Open Attributions, & Open "Agency"
Failure	Refers to perceiving one's own performance to be poor/insufficient	Open Attributions
Fun	Refers to requesting more enjoyable or game-like design elements	Open "Agency"
Growth	Refers to a focus on improving one's own growth/learning in MathSpring	Open Attributions, & Open "Agency"
Hard/Confused	Refers to low difficulty level; in "Agency" prompt is a request for more challenging material; in Open Feelings prompt "confused" tag is used to refer to student experience of confusion rather than content.	All Prompts
IDK	Refers to "I don't know" or any instance where a student states that they don't know how to respond to a given prompt.	All Prompts
Material	Refers to the material (i.e. math problems) itself without directly addressing difficulty level.	Forced Choice Attributions, Open Attributions, & Open "Agency"
Needs	Refers to basic human needs/experiences which are indirectly related to the task at hand (e.g. food, water, temperature, exhaustion, lighting, time of day).	Forced Choice Attributions, Open Attributions, & Open "Agency"
Neg	Negative: may refer to negative valence in emotion OR as a modifier of other tags (e.g. self-report "I hate Math" = "neg" + "material")	Forced Choice Attributions, Open Feelings, & Open Attributions
Neutral	Refers to a student describing their emotional state as "Ok", neither positive nor negative	Open Feelings
Pos	Positive: may refer to positive valence in emotion OR as a modifier of other tags (e.g. self-report "I'm glad these problems are easy" = "pos" + "easy")	Forced Choice Attributions, Open Feelings, & Open Attributions
Success	Refers to perceiving one's own performance to be good (e.g. many right answers)	Forced Choice Attributions, & Open Attributions
Website	Refers to the MathSpring website itself, may be modified with "pos" or "neg" depending on whether MathSpring is seen as good/likable or bad/unlikable	Forced Choice Attributions, & Open Attributions
Quit	Refers to requesting to quit or leave the learning task, may refer to quitting work within MathSpring, quitting math class, or leaving school entirely.	Open "Agency"

bored – “bored” was a fairly common and self-explanatory tag. In these cases students would discuss feeling bored, as well as states that could be described as boredom or disinterest. Every coder had a tag that roughly reflected boredom with the exception of Coder T who only coded a portion of the data set. “bored” is highlighted in dark red in appendix X.

IDK – “IDK” is an abbreviation of “I don’t know” meant to identify instances where students claimed they didn’t know why they felt a particular way, sometimes students would simply answer “nothing” or “because I do” when asked why they feel a particular way. “IDK” is highlighted in orange in appendix X.

IDK is one of the many tags that was used in prior work (cite) where it occurred quite frequently. However, it is important to note that in this prior work “IDK” was used as a catch-all tag that could also include responses that would currently be tagged as “DTG” or “Disengaged from Task Goal”.

DTG vs needs – “DTG” or “Disengaged from Task Goal”. In prior work (cite), this construct typically means that students are engaging in a task in a way that is not related to the goal of the intended goal of the task. In this context the students’ reports illustrate a focus on something unrelated to working within MathSpring. These responses often seem absurd, for example responding with “eating chicken”, “ya”, “swagger”, or “cats”. Distinguishing between “DTG” and “IDK” was a common element in the group discussions about coding (as found in Appendix X):

“I used to code IDK being like “Nonsense” or “Uninterpretable” but it’s a little bit different. IDK can mean “I don’t know why I feel that way” you can also have a student typing like “bbbbbbbbbbbbbbbbbb” or just a nonsense set of text that doesn’t seem to be made to communicate something”

“Coder D: I just, I sort of indicated it in one place but my “IDK” straight is “I don’t know”, IDK with a question mark which is like random text which is off task.”

“DTG” is highlighted in dark brown, while “needs” is highlighted in light brown in appendix X.

Another tag often grouped with “DTG” or “IDK” is “needs”. The “needs” tag refers to students asking for accommodations to allow them to perform a task. For example, a student might attribute their emotions to: “it’s first period” or “because i haven’t eaten anything”. In these cases, students can complain about the amount of heat in the classroom, the fact that they feel tired or thirsty or hungry. These responses are only tangentially related to the learning task, however they are not the absurd self-reports found in DTG.

pos – Stands for “positive”. This tag was used to indicate a positive valence. While the tag was relatively simple, it could be used as a modifier in conjunction with other tags. “pos” or “positive” is highlighted in light green in appendix X.

neg – Stands for “negative”. This tag was used to indicate a negative valence much like the previously mentioned positive tag. Again while this tag could simply mean that a student was “unhappy”, it could

also be used as a modifier in conjunction with other tags. “neg” or “negative” is highlighted in yellow in appendix X.

easy – Easy referred to instances where students described a low difficulty level. This label can be used in combination with “pos” or “neg” to in cases where students either like or dislike the fact that they find the material to be “easy”. For example, one student explained why they felt bored with the single word response “Unchallenged”... this would be a case of “easy” + “neg”. A case of “easy” + “pos” would be when a student reports feeling calm “because i know how to do this”. “easy” is highlighted in dark blue in appendix X.

hard/confused – Hard referred to instances where students described a high difficulty level. It operates basically the same way as “easy”. “Hard” can also be used in combination with “pos” or “neg”. For example “hard” + “neg” could be used to describe an instance where a student described feeling annoyed “because i am not able to understand the problem”. “hard” and “confused” are highlighted in gray blue in appendix X.

While “hard” and “confused” could be interchangeably, it bears mentioning that “confused” shows up when students are asked about how they feel, while “hard” occurs when students are discussing their attributions for why they feel a particular way. In this sense it makes sense to have a distinct “confused” tag for feelings, even if challenging “hard” material causes students to feel “confused”.

material – Refers to the content. It could refer to “mathematics” in general, or a specific unit like “fractions & decimals”. It’s distinct from “easy” or “hard” because some student may claim to dislike math regardless of difficulty or the way it’s presented in MathSpring, this would be represented by “material” + “neg”. An instance of this would be a case where a student said they felt “terrrrrrrrrrribleeeeeeee” and then explained that this was “cause of the inventor of fractions”. “material” is highlighted in light blue in appendix X.

success – When students are doing well and answering several questions correctly. This can be related to “easy”, but not necessarily. Sometimes students don't address difficulty (e.g. “i got the problem right”), or may even feel pride at being successful despite adversity. “success” is highlighted in dark green in appendix X.

Growth – When students attribute their feelings to personal improvement or learning (e.g. “i feel like i'm learning new stuff”). This can also be used in the context of the agency prompt when students take responsibility for improving their interactions with MathSpring themselves by learning more or working harder (when given the “Agency” prompt students might respond with “study” or “work harder”). “growth” is highlighted in pink in appendix X.

website – if a student references the MathSpring website itself. Again these references can be positive or negative. For example an instance of “website” + “neg” would be “these problems are hard to read and i keep getting the same problem over and over”. “website” is highlighted in purple in appendix X.

failure – When students are doing poorly and or feel they are failing. This can be related to “hard”, but not necessarily. Rather than assessing item difficulty, students may focus instead on their own ability level. For example “Im not good at math”.

annoyed – refers to when students describe a feeling of annoyance. Many coders specifically used variations of the phrase “annoyed” as opposed to frustration. Annoyance may also imply a distinction in where the student places themselves in import in relation to the learning environment. Frustration implies negative affect due to a lack of one’s own ability to affect a change in one’s environment, while annoyance implies a negative affect due to an unimportant or trivial element of one’s environment.

neutral – when students describe their feelings as neither positive nor negative, simply “fine” or “ok”.

bugs – refers to when students identify some sort of error in the system. These are not intentional design features, but rather issues like receiving the same problem repeatedly.

design – refers to criticisms or suggestions about design improvements. These design elements are largely aesthetic about layout, color, sound, or the way the learning companions talk to students.

fun – refers to requests that MathSpring be more fun or include more game-like elements.

quit – refers to requesting to quit or leave the learning task, may refer to quitting work within MathSpring, quitting math class, or leaving school entirely.

Appendix F: Full Lexicon of Tags for Each Coder and Total Instances

Forced Choice Attribution Tags

Naomi Totals	Naomi Forced Attribution	Rashid Forced Attribution	N	Kappa	Colleen Forced Attribution	N	Kappa	Sarah Forced Attribution	N	Kappa	Tamisha Forced Attribution	N	Kappa	Taylyn Forced Attribution	N	Kappa	Danielle Forced Attribution	N	Kappa
46	IDK	#not relevant	36	0.736	Disinterested	37	0.646	nonsense	30	0.616	blank	36	0.757	xxx	41	0.097	idk	35	0.608
35	software	#Website Problems	28	0.681	Website	14	0.505	problem with system	3	0.142	system is repetitive	4	0.142	tech	3	0.139	external	24	0.287
31	bored	#Imbored	23	0.707	Not challenging	18	0.283	boring	25	0.727	bored	22	0.617	Negative engagment	3	0.138	boring	16	0.586
26	success	#Im good at Math	19	0.752				successful	16	0.689	success	18	0.513	positive self	3	0.178	internal	24	0.430
25	hate	#l'mlearning	0	0	Personal	12	0.183	negative	21	0.504	frustration	16	0.519	challenge	0	0	negative	22	0.356
16	too easy	#tooeasy	16	0.653	Unsure	0	0	easy	16	0.571	needs challenge	15	0.621	understanding	5	0.255	easy	12	0.517
15	easy	#IDK	0	0				easy	0	0	does not know	0	0	easy problems	2	0.176	supportive	6	0.150
15	confused	#I don't understand	12	0.788	Confusion	3	0.320	hard	8	0.573	math	12	0.402	hard problems	2	0.176	hard	5	0.406
12	xxx	#lts challenging	2	0.183	xxx	1	0.148	proficiency	3	0.137	neutral	0	0	idk	0	0	confident	2	0.084
11	fun	#MathisFun	6	0.500	Good	10	0.480	fun	5	0.616	fun	7	0.472	positive engagement	1	0.143	positive	8	0.192
10	dislike math	#I don't like Math	9	0.601				domain	10	0.628	help	0	0	negative math	2	0.325	negativ e	1	0.179
7	like content	#ZPD	1	0.213				positive	7	0.544	math	0	0	not engaging/ interesting	0	0	?	1	0.145

Rashid Totals	Rashid Forced Attribution	Colleen Forced Attribution	N	Kappa	Sarah Forced Attribution	N	Kappa	Tamisha Forced Attribution	N	Kappa	Taylyn Forced Attribution	N	Kappa	Danielle Forced Attribution	N	Kappa
46	#notrelevant	Disinterested	37	0.636	nonsense	32	0.668	blank	37	0.779	xxx	44	0.169	idk	37	0.651
30	#tooeasy	Not challenging	28	0.550	easy	29	0.863	needs challenge	25	0.825	understanding	10	0.366	external	24	0.370
30	#WebsiteProblems	Website	14	0.569	negative	16	0.401	frustration	15	0.510	tech	3	0.162	negative	21	0.371
27	#Imbored				boring	23	0.802	bored	18	0.539	negative engagment	3	0.157	boring	17	0.728
23	#ImgoodatMath	Good	9	0.271	successful	11	0.517	success	18	0.595	positive self	3	0.199	internal	22	0.421
18	#IdontlikeMath	Personal	14	0.293	domain	10	0.526	neutral	1	0.055	negative math	2	0.187	bored	2	0.136
13	#Idontunderstand	Confusion	3	0.338	hard	6	0.479	math	10	0.359	content	3	0.293	hard	4	0.370
12	#MathisFun				positive	6	0.417	fun	8	0.522	engaging	1	0.130	positive	10	0.254
11	#IDK	Unsure	4	0.452	idk	6	0.555	does not know	5	0.613	idk	3	0.417	unsure	2	0.240
6	#ltschallenging				lack of proficiency	0	-0.019	system is repetitive	0	-0.025	easy problems	1	0.162	supportive	2	0.069
4	#HelpfulWebsite				website	4	0.318				personal skill	0	-0.007	excitemet	1	0.398
2	#lmllearning				helpful	1	0.395	math	0	-0.006	thoughts on math	0	-0.005	confident	1	0.068
2	#ZPD	xxx	0	0.006	xxx	1	0.157	help	0	-0.009	positive engagement	1	0.496	WOW! External	0	-0.002

Colleen Totals	Colleen Forced Attribution	Sarah Forced Attribution	N	Kappa	Tamisha Forced Attribution	N	Kappa	Taylyn Forced Attribution	N	Kappa	Danielle Forced Attribution	N	Kappa
59	Not challenging	easy	29	0.530	needs challenge	26	0.507	understanding	14	0.288	external	28	0.302
57	Disinterested	nonsense	39	0.729	blank	36	0.633	xxx	54	0.206	idk	43	0.689
56	Personal	negative	24	0.408	math	27	0.509	content	5	0.130	internal	31	0.412
28	Good	positive	9	0.372	success	18	0.528	positive engagement	2	0.119	positive	24	0.518
16	Website	boring	4	0.108	frustration	9	0.413	negative engagement	1	0.065	negative	14	0.292
6	Unsure	idk	5	0.614	does not know	3	0.534	idk	3	0.661	unsure	1	0.174
3	Confusion	proficiency	2	0.165	neutral	0	-0.017	hard problems	1	0.275	supportive	2	0.084
1	xxx	n/a	1	0.665	system is repetitive	0	-0.007	thoughts on math	0	-0.004	xxx	1	0.499

Sarah Totals	Sarah Forced Attribution	Tamisha Forced Attribution	N	Kappa	Taylyn Forced Attribution	N	Kappa	Danielle Forced Attribution	N	Kappa
43	nonsense	blank	32	0.684	xxx	40	0.127	idk	36	0.652
38	negative	frustration	19	0.577	content	2	0.047	negative	35	0.492
36	easy	needs challenge	27	0.807	understanding	11	0.350	external	26	0.318
29	boring	bored	20	0.598	negative engagement	3	0.150	boring	18	0.680
18	domain	math	16	0.592	negative math	2	0.162	unsupportive	1	-0.001
16	successful	success	12	0.432	positive self	3	0.283	internal	16	0.260
15	positive				solving	0	-0.007	positive	12	0.258
14	proficiency	neutral	2	0.141	easy problems	1	0.071	confident	8	0.363
10	hard	math	1	0.161	hard problems	3	0.333	hard	6	0.564
10	idk	does not know	5	0.658	idk	3	0.452	?	0	-0.013
10	xxx				thoughts on math	1	0.176	internal positive	1	0.162
5	fun	fun	5	0.438	positive engagement	1	0.278	fun	4	0.888
5	lack of proficiency				not engaging/interesting	0	-0.011	unsure	2	0.327
3	problem with system				too easy	0	-0.009	supportive	3	0.112
2	n/a				uninterpretable	0	-0.005	xxx	1	0.398
1	repetitive	system is repetitive	1	0.329	tech	1	0.497	neutral	0	-0.002
1	unsuccessful	help	1	0.665	engaging	0	-0.006	negative	0	-0.002

Tamisha Totals	Tamisha Forced Attribution	Taylyn Forced Attribution	N	Kappa	Danielle Forced Attribution	N	Kappa
44	blank	xxx	42	0.157	idk	35	0.625
33	math	content	4	0.170	internal	22	0.358
33	success	positive self	3	0.136	positive	28	0.576
31	bored	not engaging/interesting	1	0.055	boring	14	0.517
29	needs challenge	understanding	10	0.379	external	22	0.334
23	frustration	positive engagement	0	-0.016	negative	22	0.432
17	fun	negative engagment	3	0.248	fun	4	0.374
5	does not know	idk	3	0.746	unsure	1	0.174
5	neutral	thoughts on math	1	0.329	confident	2	0.120
5	system is repetitive	tech	1	0.238	supportive	5	0.200
2	help	easy problems	0	-0.012	unsupportive	2	0.112
1	math	dont understand content	1	0.665	neutral	0	-0.002

Taylyn Total Count	Taylyn Forced Attribution	Danielle Forced Attribution	Danielle N	Danielle Kappa
162	xxx	idk	55	0.186
15	understanding	negative	5	0.052
5	negative engagment	boring	2	0.128
4	positive self	iidk?	1	0.395
2	misunderstanding	xxx	0	-0.008
2	negative math	idk?	0	-0.011
1	challenge	?	0	-0.008
1	engaging	bored	1	0.214
1	negative content	boirng	0	-0.006

Open Feeling Tag

Naomi Total Count	Naomi Feelings	Rashid Feelings	N	Rashid Kappa	Colleen Feelings	N	Colleen Kappa	Sarah Feelings	N	Sarah Kappa	Tamisha Feelings	N	Tamisha Kappa	Taylyn Feelings	Taylyn N	Taylyn Kappa	Danielle Feelings	Danielle N	Danielle Kappa
119	bored	#notrelevant	47	0.268148	disinterested	80	0.462144	deactivating	28	0.25149	neutral	69	0.39464	Bored tired meh	11	0.018651	idk	40	0.324951
93	annoyed	#annoyed	20	0.28634	annoyed	30	0.388227	negative	66	0.485381	frustration	64	0.69039	Annoyed confused not ok	17	0.088725	negative	71	0.452978
85	ok	#content	59	0.39934	satisfactory	72	0.612946	neutral	52	0.439317	discomfort	0	-0.00946	calmfineok	43	0.205433	calm	5	0.096081
74	good	#happy	14	0.206024	optimistic	60	0.790895	positive	66	0.657724	positive	59	0.760511	Good awake better	22	0.232	positive	66	0.436251
16	confused	#confused	7	0.524285	confused	8	0.579313	confused	7	0.548307				Agitated frustrated bad stressed	0	-0.00931	confused	10	0.681614
13	null	#noemotion	2	0.208379	angry	0	-0.0412	n/a	4	0.348054	blank	12	0.283046	missing	5	0.506452	nothing	2	0.12131
11	depressed	#sad	3	0.391354	amused	0	-0.01441	null	1	0.042694	negative	7	0.44823	Idk silly	1	0.006369	Negative?	2	0.086541
7	anxious	#frustration	4	0.290107	sad	5	0.464079	nonsense	0	-0.02542				Agitated frustrated bad stressed	3	0.211679	hard	1	0.217939

Rashid Total Count	Rashid Feelings	Colleen Feelings	Colleen N	Colleen Kappa	Sarah Feelings	Sarah N	Sarah Kappa	Tamisha Feelings	Tamisha N	Tamisha Kappa	Taylyn Feelings	Taylyn N	Taylyn Kappa	Danielle Feelings	Danielle N	Danielle Kappa
113	#content	satisfactory	64	0.476728	positive	62	0.420884	neutral	52	0.158182	Calm fine ok	36	0.444079	positive	106	0.754303
73	#notrelevant	disinterested	46	0.321124	nonsense	21	0.393126	blank	30	0.313255	Good awake better	6	0.019884	idk	33	0.450988
30	#bored				negative	29	0.197394	negative	5	0.147581				negative	29	0.27966
21	#happy	optimistic	14	0.25392				positive	14	0.225942	Grea tpleased	3	0.348497	silly	2	0.167345
20	#annoyed	annoyed	17	0.641782				frustration	19	0.331256	Annoyed confused not ok	12	0.591928	annoyed	17	0.866162
16	#frustration	sad	6	0.38874							Agitated frustrated bad stressed	5	0.451461	frustrated	3	0.273763
13	#upset	angry	8	0.430553							Agitated frustrated bad stressed	1	0.157527	hungry	1	0.115689
9	#disengaged													idkquestmark	1	0.10986
9	#idk				idk	9	0.946026				idsilly	7	0.584767	thirsty	0	-0.0039
8	#confused	confused	6	0.744152	confused	6	0.796002							confused	7	0.773428
8	#relaxed				deactivating	6	0.182189							calm	5	0.766122
6	#indifferent				neutral	6	0.126476				boredtiredmeh	5	0.533471	Negative?	4	0.351411
5	#noemotion				null	4	0.492				missing	1	0.179612	nothing	4	0.526516
4	#engaged													awake	1	0.280385
4	#sad													sad	1	0.397919
3	#ambivalence													positve	0	-0.00652
3	#disdain	amused	0	-0.00989	n/a	0	-0.01226	discomfort	0	-0.00688				bored	0	-0.01184
2	#angry													same	0	-0.00289
2	#anxious													hard	0	-0.00289
1	#agitated													null	0	-0.00216
1	#tired													tired	1	0.397919

Colleen Total Count	Colleen Feelings	Sarah Feelings	Sarah N	Sarah Kappa	Tamisha Feelings	Tamisha N	Tamisha Kappa	Taylyn Feelings	Taylyn N	Taylyn Kappa	Danielle Feelings	Danielle N	Danielle Kappa
116	disinterested	negative	48	0.168805	blank	41	0.298773	Bored tired meh	12	0.066667	negative	49	0.175178
91	satisfactory	neutral	58	0.628013	neutral	65	0.445716	Calm fine ok	54	0.536232	positive	68	0.437668
66	optimistic	positive	64	0.710505	positive	56	0.757779	Good awake better	24	0.325467	excited	2	0.046148
31	annoyed	n/a	0	-0.02718	frustration	29	0.470114	Annoyed confused notok	14	0.388247	annoyed	16	0.620699
22	angry	idk	0	-0.03754				Agitated frustrated bad stressed	8	0.334252	frustrated	1	0.057487
13	sad	deactivating	4	0.108076	negative	6	0.346161	Agitated frustrated bad stressed	0	-0.01118	sad	1	0.139405
8	confused	confused	6	0.795991				missing	0	-0.02667	confused	8	0.886714
4	amused	nonsense	2	0.1385	discomfort	2	0.664107	idsilly	3	0.272277	idk	2	0.054993

Sarah Total Count	Sarah Feelings	Tamisha Feelings	Tamisha N	Tamisha Kappa	Taylyn Feelings	Taylyn N	Taylyn Kappa	Danielle Feelings	Danielle N	Danielle Kappa
113	negative	frustration	62	0.544199	Annoyed confused not ok	20	0.055808	negative	106	0.703893
100	positive	positive	65	0.687633	Good awake better	27	0.200939	positive	86	0.605768
73	neutral	neutral	54	0.410803	Calm fine ok	39	0.387932	Negative?	12	0.236023
40	activating									
39	deactivating	negative	8	0.224386	Bored tired meh	6	0.084555	bored	24	0.489799
22	nonsense	blank	22	0.479941	Agitated frustrated bad stressed	0	-0.12437	idk	21	0.523897
12	null				Agitated frustrated bad stressed x2	0	-0.01042	nothing	8	0.757081
10	idk	discomfort	0	-0.00873	idksilly	7	0.504384	calm	0	-0.01969
10	n/a				missing	6	0.691698			
7	confused				greatpleased	0	-0.03369	confused	7	0.820679

Tamisha Total Count	Tamisha Feelings	Taylyn Feelings	Taylyn N	Taylyn Kappa	Danielle Feelings	Danielle N	Danielle Kappa
123	neutral	Calm fine ok	39	-0.03578	bored	21	0.199065
77	frustration	Annoyed confused no tok	19	0.184656	negative	69	0.565392
73	positive	Good awake better	23	0.253405	positive	69	0.535363
62	blank	idksilly	7	0.037089	idk	33	0.513768
19	negative	Agitated frustrated bad stressed	3	0.0696	sad	1	0.096283
2	discomfort	Bored tired meh	0	-0.02034	silly	2	1

Taylyn Total Count	Taylyn Feelings	Danielle Feelings	Danielle N	Danielle Kappa
186	xxx			
57	Calm fine ok	positive	37	0.396765
31	Good awake better	calm	0	-0.01762
26	Annoyed confused not ok	negative	23	0.441513
16	Agitated frustrated bads tressed	frustrated	2	0.209266
16	Idk silly	idk	10	0.410535
12	Bored tired meh	bored	5	0.538543
7	missing	null	1	0.243902
5	Great pleased	confident	1	0.328173

Open Attribution Tag

Naomi Total Count	Naomi Attributions	Rashid Attributions	Rashid N	Rashid Kappa	Colleen Attributions	Colleen N	Colleen Kappa	Sarah Attributions	Sarah N	Sarah Kappa	Tamisha Attributions	Tamisha N	Tamisha Kappa	Taylyn Attributions	Taylyn N	Taylyn Kappa	Danielle Attributions	Danielle N	Danielle Kappa
186	null	#notrelevant	68	0.239417	avoidance	106	0.585955	idk	43	0.170602	blank	120	0.630982	xxx	115	-0.0054	idk	47	0.32231
35	math	#idon'tlikemath	19	0.585221	frustration	9	0.141367	domain	21	0.667452	math	20	0.403351	doesn't like math	12	0.486486	negative	22	0.274386
33	success	#tooeasy	9	0.252454	good	16	0.426689	successful	20	0.524799	experience is positive	25	0.575922	successful	11	0.404855	positive	29	0.511512
32	design	#changeavatar	5	0.239735	website issue	8	0.306247	negative	9	0.214925	frustration with the system	21	0.5523	doing task	2	0.101312	external	19	0.320487
16	bugs	#websiteproblems	16	0.600718	system issue	3	0.285938	problem with system	13	0.7046	repetition	3	0.27188	tech issues	5	0.364017	unsupportive	9	0.502446
15	too hard	#idon'tunderstand	10	0.585375	confusion	4	0.309429	lack of proficiency	6	0.458182	discomfort answering feeling questions while doing math	0	-0.01263	answering question	0	-0.00496	unsure	9	0.552735
14	learning	#i'mlearning	10	0.759766				improvement	7	0.593759	experience is neutral	1	0.030565	likes task	1	0.106023	internal	9	0.098397
14	out	#nochange	4	0.266133	physical	4	0.384081	unrelated	3	0.195899	neutral	6	0.119665	doesn't like task	0	-0.02753	hungry	7	0.66235
12	failure	#lowachievement	4	0.405013	environment	0	-0.01881	unsuccessful	4	0.3217	experience is inconsistent	3	0.26691	stuck	6	0.407176	frustrated	5	0.331953
12	too easy	#i'mbored	3	0.171353	not challenged	10	0.310691	easy	7	0.39314	unchallenged	3	0.392731	too easy	1	0.135053	too easy	3	0.396029
11	easy	#i'mgoodatmath	4	0.324176				proficiency	4	0.166822	confidence	3	0.252214	outside influence	0	-0.01256	easy	6	0.374944

Rashid Total Count	Rashid Attributions	Colleen Attributions	Colleen N	Colleen Kappa	Sarah Attributions	Sarah N	Sarah Kappa	Tamisha Attributions	Tamisha N	Tamisha Kappa	Taylyn Attributions	Taylyn N	Taylyn Kappa	Danielle Attributions	Danielle N	Danielle Kappa
83	#notrelevant	avoidance	75	0.473637	nonsense	13	0.560531	blank	55	0.298086	xxx	47	-0.15593	because	23	0.349894
27	#idk				idk	25	0.662215				idk	9	0.445253	idk	26	0.647766
27	#websiteproblems	website issue	12	0.558212	problem with system	16	0.669635	frustration with the system	18	0.57156	tech issues	7	0.360587	external	23	0.507135
25	#idon'tlikemath	frustration	15	0.4	domain	19	0.601218	math	21	0.586367	doesn't like math	11	0.57183	negative	24	0.384541
22	#tooeasy	not challenged	19	0.519868	easy	17	0.733474	experience is positive	11	0.235219	successful	4	0.175994	easy	17	0.783632
17	#i'mbored	physical	1	0.035813	boring	10	0.69912	boredom	13	0.728236	emotional state	1	0.084695	boring	7	0.552602
15	#idon'tunderstand	confusion	6	0.480519	hard	7	0.444153	unchallenged	0	-0.01579	wants challenge	0	-0.00619	unsure	9	0.612153
13	#nochange				unrelated	7	0.688724	neutral	13	0.44795	answering question	0	-0.00613	internal	12	0.160735
11	#personalproblems				negative	2	0.087401	unrelated to system	3	0.247467	uninterpretable	1	0.010543	questmark	2	0.130929
10	#i'mlearning	good	7	0.28115	improvement	6	0.742343	repetition	0	-0.02161	likes task	1	0.140845	learning	4	0.567674
9	#i'mgoodatmath				proficiency	8	0.570755				silly	0	-0.02153	positive	9	0.211693
7	#igotheanswers	system issue	0	-0.01504	successful	6	0.40749				isn't good at math	0	-0.00577	confident	6	0.272821
7	#indifferent				not helpful	0	-0.00651				same	0	-0.00577	math	1	0.056371
6	#lowachievement				discomfort answering feeling questions while doing math	4	0.425163	unsuccessful	1	0.189526	stuck	3	0.359244	frustrated	5	0.52033
5	#changeavatar				companion	5	0.657357	experience is inconsistent	0	-0.01707	doesn't like task	1	0.136422	avatar	4	0.798469
5	#nothing				experience is neutral	3	0.224549	outside influence	3	0.746888	neutral	3	0.541655	neutral	3	0.541655
4	#igotheanswers				helpful	0	-0.00607	confidence	3	0.418345	missing	0	-0.02215	hard	0	-0.00524
3	#mathisfun				neutral	0	-0.00379	likes math	1	0.329897	doing task	1	0.395241	school	1	0.397804
3	#timeofday	environment	2	0.566474	environment	2	0.798172	outside factors	1	0.497682	bored	1	0.497529	early	3	1
2	#justfine				bored	0	-0.00379	no changes	1	0.496904	not changed	1	0.665203	repetitive	0	-0.00436
1	#indiffernt				lack of proficiency	0	-0.00675							survey question	0	-0.00228
1	#nobenefits										too easy	0	-0.00439	supportive	1	0.1225
1	#switchtopics													idkquestmark	0	-0.00284
1	#zpd										other person influence	0	-0.00329	bored	1	0.283848

Colleen Total Count	Colleen Attributions	Sarah Attributions	Sarah N	Sarah Kappa	Tamisha Attributions	Tamisha N	Tamisha Kappa	Taylyn Attributions	Taylyn N	Taylyn Kappa	Danielle Attributions	Danielle N	Danielle Kappa
146	avoidance	idk	38	0.206747	blank	74	0.529087	xxx	94	0.021825	idk	48	0.431894
45	not challenged	easy	17	0.478018	experience is positive	17	0.295783	too easy	1	0.030255	easy	17	0.492878
40	frustration	negative	16	0.404313	math	23	0.471687	doesn't like task	8	0.30254	negative	32	0.449525
34	good	successful	15	0.395166	neutral	12	0.155729	successful	6	0.197283	internal	28	0.329953
14	website issue	problem with system	8	0.505127	frustration with the system	11	0.456321	0	0	0	external	13	0.333014
9	confusion	hard	4	0.484506	confidence	0	-0.02875	stuck	2	0.184959	hard	3	0.496525
6	physical	unrelated	3	0.312336	boredom	2	0.142705	uninterpretable	1	0.03867	hungry	3	0.456149
4	environment	environment	3	0.855249	unrelated to system	3	0.418248	likes task	1	0.277512	early	2	0.569163
4	system issue	repetitive	0	-0.00614	experience is inconsistent	2	0.390921	tech issues	4	0.608199	supportive	4	0.41539

Sarah Total Count	Sarah Attributions	Tamisha Attributions	Tamisha N	Tamisha Kappa	Taylyn Attributions	Taylyn N	Taylyn Kappa	Danielle Attributions	Danielle N	Danielle Kappa
45	idk	blank	23	0.034125	xxx	36	0.090702	idk	26	0.610082
27	negative	repetition	2	0.093648	doesn't like task	4	0.201807	negative	25	0.323126
24	domain	math	19	0.500537	doesn't like math	11	0.469141	math	20	0.688603
22	successful	experience is positive	16	0.347477	successful	11	0.530093	internal	29	0.358056
19	easy	experience is neutral	4	0.147462	too easy	2	0.179268	easy	16	0.815035
16	problem with system	frustration with the system	9	0.323254	tech issues	7	0.493606	external	15	0.370571
14	proficiency	confidence	2	0.074997	isn't good at math	1	0.113289	positive	12	0.30486
13	nonsense				silly	5	0.543103	idkquestmark	4	0.413737
12	unrelated	neutral	8	0.252119				hungry	3	0.395591
11	boring	boredom	9	0.60117				boring	7	0.630181
8	lack of proficiency	outside factors	0	-0.00649	idk	0	-0.03302	unsure	7	0.38853
8	unsuccessful				stuck	6	0.47903	frustrated	8	0.507239
7	hard	likes math	0	-0.00631				hard	3	0.59743
7	improvement							learning	3	0.59743
6	companion							avatar	5	0.620676
3	environment	unrelated to system	3	0.452481	other person influence	0	-0.00569	early	2	0.799268
1	bored				emotional state	1	1	bored	1	0.398537
1	helpful	no changes	0	-0.00722	likes task	1	0.664981	supportive	2	0.247857
1	neutral							repetitive	0	-0.00253
1	not helpful	experience is inconsistent	1	0.281298		0	0	unsupportive	1	0.130868
1	repetitive	discomfort answering feeling questions while doing math	0	-0.00541	wants challenge	0	-0.00379	survey question	1	0.66599
1	xxx	unchallenged	0	-0.00541	uninterpretable	0	-0.00506		0	0

Tamisha Total Count	Tamisha Attributions	Taylyn Attributions	Taylyn N	Taylyn Kappa	Danielle Attributions	Danielle N	Danielle Kappa
126	blank	xxx	74	-0.05026	idk	39	0.372808
42	experience is positive	successful	11	-0.05026	positive	34	0.573796
41	math	doesn't like math	11	0.383382	negative	38	0.536728
41	experience is neutral	outside influence	3	0.306077	neutral	5	0.447358
41	neutral	idk	3	0.070733	internal	18	0.15102
32	frustration with the system	tech issues	4	0.162531	external	23	0.460897
18	boredom	emotional state	1	0.082162	boring	6	0.4525
16	experience is inconsistent	other person influence	1	0.282306	unsupportive	4	0.392394
10	confidence	too easy	0	-0.00932	confident	6	0.249263
10	unrelated to system	wants challenge	0	-0.01881	early	2	0.302897
6	discomfort answering feeling questions while doing math	stuck	1	0.131516	survey question	1	0.282046
5	repetition	doesn't like task	2	0.295686	repetitive	1	0.193994
3	unchallenged	missing	0	-0.01456	too easy	2	0.66517
2	no changes	not changed	1	0.665431	because	2	0.106078
1	likes math	doing task	1	0.397329	school	1	0.49888
1	outside factors	bored	1	1	slow	1	1

Taylyn Total Count	Taylyn Attributions	Danielle Attributions	Danielle N	Danielle Kappa
226	xxx	internal	67	0.213522
23	uninterpretable	because	12	0.404097
19	missing	not fun	0	-0.00289
14	successful	positive	11	0.238193
12	doesn't like math	negative	11	0.187489
11	idk	idk	9	0.270315
10	stuck	frustrated	7	0.601877
9	tech issues	external	8	0.213225
8	doesn't like task	math	1	0.052329
5	silly	confident	1	0.037236
3	emotional state	bored	2	0.441057
3	likes task	supportive	1	0.104326
3	outside influence	neutral	3	0.664634
2	doing task	school	1	0.49848
2	too easy	easy	2	0.195612
1	answering question	nothing	0	-0.00202
1	bored	slow	1	1
1	isn't good at math	unsure	0	-0.00284
1	not changed	boring	0	-0.0027
1	other person influence	unsupportive	1	0.130875
1	same	no fun	0	-0.00152
1	wants challenge	early	0	-0.00228

Open Agency Tag

Naomi Total Count	Naomi Agency	Rashid Agency	Rashid N	Rashid Kappa	Colleen Agency	Colleen N	Colleen Kappa	Sarah Agency	Sarah N	Sarah Kappa	Tamisha Agency	Tamisha N	Tamisha Kappa	Taylyn Agency	Taylyn N	Taylyn Kappa	Danielle Agency	Danielle N	Danielle Kappa
153	idk	#nothing	77	0.479221	no change	84	0.583626	none	79	0.4912	neutral	74	0.461552	xxx	137	0.218201	nothing	78	0.575936
34	content	#difficultylevel	11	0.419149	content	29	0.751509	questions	27	0.714127	more challenges	13	0.456804	content	4	0.191553	content	26	0.771424
25	disengage	#notrelevant	15	0.280835	personal	18	0.276215	leave	13	0.664122	quit	14	0.679711	no change	2	0.092455	leave	7	0.423738
23	design	#aesthetics	7	0.3804	structure	17	0.565527	display	8	0.395653	aesthetics	13	0.488428	flow	2	0.13886	external	10	0.270971
17	more fun	#morefun	11	0.746489				more fun	13	0.726112	fun	9	0.537603	more engaging	1	0.07561	fun	8	0.581074
15	internal	#improvemyself	13	0.859498				outside	7	0.517028	change self	6	0.504263	challenge	1	0.096296	internal	11	0.778993
12	lc	#chnageavatar	4	0.428904				companion	8	0.652068	blank	4	0.013333	avatar	2	0.277594	avatar	7	0.659009
7	bugs	#debugit	7	0.818444	error	5	0.829826	system	6	0.586854	frustration with system	7	0.481611	completion	3	0.594415	finish	3	0.596298
5	basic needs	#ammeneties	4	0.887183				environment	5	0.707716	sustenance	3	0.661585				water	3	0.6634
5	hints	#hints	4	0.79654	unsure	0	-0.03213	hints	5	0.829863	asks for help	3	0.298322	hint	4	0.796667	hint	4	0.8878

Rashid Total Count	Rashid Agency	Colleen Agency	Colleen N	Colleen Kappa	Sarah Agency	Sarah N	Sarah Kappa	Tamisha Agency	Tamisha N	Tamisha Kappa	Taylyn Agency	Taylyn N	Taylyn Kappa	Danielle Agency	Danielle N	Danielle Kappa
78	#nothing	no change	77	0.921225	none	77	0.964157	neutral	49	0.500473	xxx	73	0.126126	nothing	77	0.98285
55	#notrelevant	personal	34	0.458125	nonsense	15	0.4926	blank	27	-0.04653	emotional state	1	0.029161	idkquestmark	11	0.274129
35	#idk	unsure	35	0.967989	idk	35	0.967934				idk	7	0.305489	idk	35	0.674521
15	#improvemyself				outside	7	0.621393	change self	6	0.50343	easierquestmark	1	0.119326	study	2	0.226459
12	#difficultylevel	content	10	0.339806	domain	1	0.083865	asks for help	6	0.456011	easier	2	0.254368	content	10	0.445423
11	#morefun							fun	7	0.592154	more fun	1	0.161383	fun	7	0.728709
10	#debugit	error	5	0.658291	system	6	0.526742	frustration with system	8	0.51026	completion	3	0.452861	finish	3	0.453374
9	#morechallenges				questions	9	0.366064	more challenges	7	0.518434	challenge	1	0.153578	math content	4	0.267864
9	#switchtopics							negative	2	0.247522	questions	1	0.153578	less survey questions	1	0.195313
8	#aesthetics	structure	7	0.308735	display	8	0.792278	aesthetics	7	0.478958	more engaging	1	0.169363	color	7	0.931683
6	#notrelvant				leave	3	0.363464	quit	3	0.278587	uninterpretable	4	0.796646	leave	3	0.450038
5	#chnageavatar				companion	4	0.520499	positive	0	-0.01343	clarity	0	-0.00576	change time	0	-0.00542
5	#everything				everything	3	0.537959				more challenge	0	-0.00992	everything	3	0.746929
4	#ammeneties				environment	4	0.607299	sustenance	3	0.746416	challenging	0	-0.00553	water	2	0.56662
4	#hints				hints	4	0.796366				hint	4	1	faster	1	0.32753
4	#music				music	3	0.855299	location	0	-0.00569	add audio	2	0.663584	more fun	0	-0.00871
3	#changeavatar										avatar	2	0.798337	no class	0	-0.01311
2	#gameification										timing/flow	0	-0.00692	games	1	0.496743
2	#ungamethesystem										tech issues	0	-0.0046	fix external	2	0.663765
1	#gamification				n/a	0	-0.00498	more of same	0	-0.00355	0	0	0	interactive	1	1
1	#hardwareimprovements				computer	1	1	unrelated to system	1	1	content	0	-0.00553	better computer	1	1
1	#leaveit										no change	1	0.245462	harder math content	0	-0.00542
1	#morecontent				more fun	11	0.838535	variety	1	1	amount of content	0	-0.00345	more math content	1	0.396877
1	#moretechnology							clarity	0	-0.00355	hints	0	-0.0046	breaks	0	-0.00325
1	#softwareissues							satisfied	0	-0.00473	flow	1	0.497409	new math content	1	0.665222

Colleen Total Count	Colleen Agency	Sarah Agency	Sarah N	Sarah Kappa	Tamisha Agency	Tamisha N	Tamisha Kappa	Taylyn Agency	Taylyn N	Taylyn Kappa	Danielle Agency	Danielle N	Danielle Kappa
85	no change	none	79	0.94723	neutral	54	0.53304	xxx	79	0.138641	nothing	78	0.941573
72	personal	nonsense	15	0.269253	blank	24	0.087942	easierquestmark	1	0.020497	idkquestmark	10	0.176196
40	content	questions	31	0.78532	more challenges	12	0.365385	content	4	0.16	content	27	0.742693
37	unsure	idk	36	0.968625	negative	1	0.008887	idk	7	0.288256	idk	36	0.678783
33	structure	display	8	0.333795	aesthetics	14	0.480757	hint	4	0.195721	color	7	0.32459
5	error	system	5	0.543199	frustration with system	5	0.39918	completion	3	0.746606	finish	3	0.746908

Sarah Total Count	Sarah Agency	Tamisha Agency	Tamisha N	Tamisha Kappa	Taylyn Agency	Taylyn N	Taylyn Kappa	Danielle Agency	Danielle N	Danielle Kappa
79	none	neutral	50	0.538308	xxx	74	0.12746	nothing	78	0.982936
37	idk	blank	17	0.150515	idk	7	0.287803	idk	36	0.735708
36	questions	more challenges	13	0.430688	content	4	0.178571	content	26	0.748768
16	nonsense	negative	0	-0.03353	flow	0	-0.01865	idkquestmark	10	0.674391
15	more fun	fun	10	0.698684	more fun	1	0.119015	fun	8	0.653944
13	leave	quit	10	0.727062	no change	2	0.172724	leave	7	0.690805
13	system	frustration with system	10	0.624063	completion	3	0.363762	fix external	3	0.364882
10	display	aesthetics	8	0.456695	more engaging	1	0.139456	color	7	0.771129
9	companion	variety	0	-0.00674	avatar	2	0.356	everything	1	0.138249
9	environment	sustenance	3	0.450219	0	0	0	water	3	0.492537
9	outside	positive	3	0.491493	more challenge	0	-0.00656	study	2	0.356757
6	domain	clarity	0	-0.00641	questions	0	-0.01471	no math	4	0.722323
6	hints	asks for help	4	0.454853	hint	4	0.79646	hints	2	0.49505
3	everything	more of same	0	-0.00561	uninterpretable	0	-0.00546	everything	3	1
3	music	location	0	-0.00561	add audio	2	0.798246	0	0	0
2	n/a	satisfied	0	-0.00749	hints	0	-0.0073	more help	0	-0.00438
2	self	change self	2	0.437944	easierquestmark	1	0.665049	faster	1	0.496711
1	computer	unrelated to system	1	1	amount of content	0	-0.00364	better computer	1	1

Tamisha Total Count	Tamisha Agency	Taylyn Agency	Taylyn N	Taylyn Kappa	Danielle Agency	Danielle N	Danielle Kappa
157	blank	xxx	141	0.160175	idk	35	0.044852
76	neutral	idk	7	0.138415	nothing	49	0.515701
20	aesthetics	more engaging	2	0.161934	color	6	0.425152
20	frustration with system	questions	3	0.250151	fix external	4	0.318633
17	more challenges	challenge	2	0.188664	content	10	0.382137
14	quit	no change	2	0.169233	leave	7	0.656285
13	asks for help	hint	3	0.341911	easier content	5	0.544919
12	fun	tech issues	0	-0.00508	fun	6	0.587174
8	change self	easierquestmark	1	0.218415	math content	2	0.122386
6	negative	content	1	0.189339	no class	3	0.490099
4	sustenance	flow	0	-0.00948	water	2	0.56662
3	positive	hints	0	-0.00662	study	2	0.798434
2	satisfied	more challenge	0	-0.00551	harder math content	0	-0.00933
1	clarity	clarity	1	1	breaks	0	-0.00325
1	location	uninterpretable	0	-0.0044	idkquestmark	1	0.137674
1	more of same	math	1	1	faster	0	-0.00433
1	unrelated to system	missing	0	-0.00508	better computer	1	1
1	variety	amount of content	0	-0.00275	more math content	1	0.396877

Taylyn Total Count	Taylyn Agency	Danielle Agency	Danielle N	Danielle Kappa
291	xxx	nothing	73	0.020122
12	missing	more fun	0	-0.0108
7	idk	idk	7	0.170387
7	no change	leave	2	0.269791
4	uninterpretable	negative	1	0.327749
3	completion	finish	3	1
3	content	content	3	0.152139
3	more engaging	colors	1	0.327044
3	questions	less survey questions	1	0.497653
2	add audio	study	0	-0.00627
2	avatar	games	0	-0.00627
2	easier	easier content	2	0.56758
2	hint	faster	1	0.327749
2	hints	hints	1	0.39548
2	more challenge	harder math content	1	0.242925
2	timing/flow	energy	0	-0.00417
1	amount of content	math content	1	0.08547
1	challenge	more challenging	1	0.497653
1	challenging	no class	0	-0.00537
1	clarity	change time	0	-0.00312
1	easierquestmark	breaks	0	-0.00312
1	emotional state	everything	0	-0.00469
1	flow	new math content	1	0.497653
1	math	idkquestmark	0	-0.00582
1	more fun	fun	1	0.217891
1	tech issues	fix external	0	-0.00501

Appendix G: Coder Discussion Transcript

Day 1 Discussion

Coder N: So right now I've got my screen up on screenshare. So everyone can see what I'm looking at and what I'm working on right now. Also just as a quick note right now. I am actually recording. I'm not releasing the video but I am transcribing for only those who have given permission to have transcriptions of their work included here. So if you're not included you'll just have redacted text.

Anyway, that's my quick disclaimer there for everyone. Welcome, thank you so much for joining in, we're going to be here for the next 45 minutes or less. I'm just going to go through some of the most common codes people used and where I saw them to be relatively correlated. It was a bit of a fuzzy technique I used looking for things that were highly correlated and also things that had similar meaning. So in some ways it was almost like a meta coding experience. I'm going to start with the very first set we had which was the forced choice set. The forced choice set had students give a multiple choice response you can see right here students got to choose from the "Not at all Confident" through the "Extremely Confident". And that was a "Forced choice" and then they would have to put in why that is. So these are the codes we came up with for that.

Some of the really common ones I saw.

Coder S: I'm not seeing the screen there.

Coder N: Oh! Thank you I must be having a problem with my screen share.

.....

Coder N: Ok, now is this visible to people.

Coder T: That's good.

Coder N: All right thank you so much! So quick run down the first set we'll be going though is based on the forced choice tags where students would respond to this multiple choice question, it looked like this, and then describe why they felt that way. The technique I used to get peoples' codes out was I found there had to be at least 10 codes out of the total code set. So these were all things that were somewhat frequent, then I looked for codes that were highly correlated and descriptively similar. Anything that I've colored in are things that I'm thinking of preserving. I'm a little fuzzy about whether I want to keep "confused" because that was pretty rare, it didn't happen very often or a lot of people didn't include a code of that. Also distinguishing between two things "positive" and "negative"; you'll notice we've both "dislike math" and then we've got "math is fun". So having a "positive" or "negative" code and separate code like "content" you could get "content"/"positive" or "content"/"negative" which means that you'd have a few more codes for then, but you could also use the "positive" and "negative" around things like

“easy”. Where in some cases students said it was easy but it was “too easy” or “not challenging” so that would be “easy”/“negative”. Whereas a more general easy statement might be “easy”/“positive”.

Coder S: I think I had a code for both of them separately.

Coder N: So Coder S and Coder D have actually worked with me on a prior paper for this which had some prior codes that included that “positive”/“negative” separation. So in your case you got “domain (negative)” right over here. Is that what you’re describing?

Coder S: That was usually the “I don’t like math” but I’m saying that if they said “The problems are easy, and I’m getting them all right”. I think I tagged it as easy and internal.

Coder N: Oh so you did do that.

Coder S: It was both the problems and an internal “because I am succeeding”.

Coder N: Yeah I think that actually makes sense. I like that if you’re doing a thing around “positive” there. That involves tagging along with “easy”... I’m sorry I’ve just got easy right here. I tried to separate out a lot of individual codes rather than getting combos of them for these charts.

Coder S: There’s also I guess the blue one is “Successful” which other people also had. “Success”: I’m good at math.

Coder N: I’m also thinking maybe of distinguishing between the idea of “enjoyment of math” and “success at math”. I think those are two distinct ways students can look at this. I think for now that having a tag for “success” and then tags for “positive” and “negative” makes sense? Because “positive”/“negative” content makes it sound like I like or dislike it while success is a bit of a self-assessment.

Coder S: I think on the previous paper we had “internal”/“external”. I guess you have them here but they’re not highlighted because not enough people used those. If it was the domain or the website it was “external”, whereas if it was something about themselves it was tagged as “internal”.

Coder N: You are correct. One of the things I was doing differently on this-

Coder D: I coded it that way.

Coder N: I see right here actually, Coder D, you’ve got internal and external.

Coder S: I think you’re the only one.

Coder T: I coded “tech issues” as well.

Coder N: I’m sorry?

Coder S: Yeah I think that might be the “website” one.

Coder N: Yeah the “unsupportive”.

Coder S: I think mine included “tech issues”.

Coder N: Yeah I think I'm trying to get a little more verbose than we had with “internal”/“external”. So the idea of “content” is an external thing but “software” is also an external thing. So I'm trying to do them distinctly here. One reason I tried to get away from it a little bit I'm trying to be more driven by what we saw in the individual codes than what we recall from the prior paper. So I tried to get some people [coders] who had worked on this previously, and other people who had not. But yes I do agree these things are “external”. I don't know students would view them so much as “external” or “math” in general. I did get times students would say “I really like math, but I don't like this website” or something along those lines.

Coder S: Right... Yeah that's true. There was sometimes the “I'd rather just do my math on paper, I don't like this site”.

Coder N: Or also some students liked some math content, but not other math content.

Coder S: Even the domain one it would be “I don't like math” or some would be like “I don't like fractions”.

Coder N: Yes! I remember that! I think in some ways “content” is kinda nice because it is vague, and it's a bit of a catch all whereas fractions seems specific and it would be harder to find a specific example of?
...

Coder N: Yeah... umm... So what I'm currently thinking of is going with the “IDK” value, for students saying they just don't know. “Bored” I feel was really really frequent. There were a lot of students that would bring up disinterest or boredom. And doing the “positive”/“negative” split as sort of general “positive”/“negative” things that could be applied to different things. Content, which would be largely mathematics but could also involve subsections like fractions. Success which is like self-assessment a student would have. So they might not like something, but think they are doing well at it, but I think most often you'll get “success” + “positive” that they feel good about success. And then not all the time because “easy” would be a thing where students might be being successful or say something is “easy” but they're being really critical of their success because they wish they were more challenged. And that's really valuable. If you get easy and negative I think that tells you a little bit more about a student that they're actually challenge seeking. Finally, you've got “software” here, that could be a thing where you're talking about that article. Actually, this other weird one I got was that some students, I guess we might just use “negative” for this, but I felt that some students had really strong negative feelings, like really intense negative feelings. That tended to be correlated between a few of us. I'm sorry? Coder S?

Coder S: hm?

Coder N: Oh. Really really negative feelings some students had.

Coder S: Yeah... yeah, not just “I don't like math” but “This website is terrible, it's the worst thing ever”.

Coder N: Yeah... yeah... or even like things that wouldn't have an article. They wouldn't have a direction just: "I hate this" really just really a lot of anger and kind of hatred I would get sometimes. It wouldn't even seem directed. So I'm not sure if I want to include a separate category for that... I think I might... for that sort of thing here... but maybe not 'cause we could put that under "negative"...

Coder N: I guess finally "confused" is sort of a thing where it's pretty rare but I do think it's a distinct cognitive state. But I think we can leave it out because it's included in later types of tags or other types of tags. I don't know. How do people feel about the confused state or the extreme negative state?

Coder T: If you don't include the "confused" state do you have a state that they're *too* challenged? Is that just...

Coder N: Ahh... that's valid.

Coder T: Because you talked about "success" + "positive"/"negative" but then there's the opposite of success. "Positive"/"negative" challenge or it's too hard.

Coder N I think you're right. I think you're right. Actually we included a tag for "hard" before. Putting down "hard" "positive" or "hard" "negative" or even just "hard" might be a good stand in for confused, showing there's a mix between the student and the material. So yeah, that actually... it think hard would be a good thing to put in there-

Coder S: Did we see... did we just not see a lot of those. I guess I coded some things with "hard" but it doesn't look like anyone else did.

Coder T: I had... I think I remember writing "stuck"

Coder N: I included "confused" and that had at least 10. Coder R had a code for "I don't understand" and actually Coder S, it should be up there although I think part of it was that "hard" could be positive or negative. You did have a code for "hard" that I think could work here. So... we might actually have enough. That's three things that could go towards hard that I think would make some sense here. If people are up for it, I'm going to include "hard" as an option here as opposed to confused. And then "hard" can be "hard" "positive", "hard" "negative", or this is just plain well this is really "hard". And I think... yeah I think that might be it unless we want to have an "extreme negative"... "negative" "negative"... "double plus negative"? Meh... I think that's okay... we've got negative and that would encompass the really difficult or "hate" kind of thing. We've now got a total of one, two, three, four, five, six, seven, eight, nine codes for this particular set. Are we good with moving on?

Coder T: The ones where... responses that were just "silly" because I think I saw them in pretty much every category.

Coder S: I think maybe I coded those under "IDK" for this one.

Coder N: Yeah, that's actually a fair point, Coder T. That I do have silly or I guess disengaged, unrelated...

Coder S: yes I did code IDK

Coder N: But Coder S, you're right we used to do for the old coding scheme IDK as a catchall for things that were silly but also "I don't know". I think that it might make sense to have a distinct value for it the sort of silly response. I think it might have happened more often at least in the 10+ instances... with the umm... yeah Coder S has "nonsense" here. So in your case it might have happened more often in the other student sets that were not just "forced choice" because I do have it later on.

Coder N: But um... yeah I'm comfortable with including a code for just sort of "nonsense" or "unrelated", would "unrelated" be OK?

Coder T: "Unrelated" sounds good.

Coder N: Okay. That might also include "I want to get a drink of water" or something like that. Like it's just like not related to the task at hand. It could also be something silly.

Coder S: Sometimes they even do things about their physical environment too though. And like that's related, exactly what they want them to talk about.

Coder N: I'm not sure... actually... that's a fair point.

Coder S: Like "I'm frustrated because this room is too hot".

Coder N: It's true, it's a creature comfort thing... hang on I'm going to jump to a different set here... Yeah there's some stuff like this. Not necessarily personal problems but like I had a really difficult morning or I'm thirsty or something like that could be in there. There were also basic needs things like "it's too hot" or something along those lines. So... I get what you're saying. You're making the point that those things would be still related, right?

Coder S: Right, like it does directly impact how they're feeling. But when we ask them we really meant about... you know, this thing right now on the system.

Coder N: What about a code like "off task/disengaged" would that be something you think could respond to that?

Coder D: Say it again?

Coder N: "off task/disengaged"

Coder D: I was going to say "off task".

Coder N: Right, the only reason I'm adding "disengaged" in there is that my Masters' thesis was around it and so they're like "oh well it's still kind of on task they could just be messing around or doing something like that" but "off task" works because I think more people will get it. I'll do "off task", how do you feel about that Coder S, Coder D?

Coder S: I guess I feel like "Off task" is the one that's like "I like cats" whereas if they're talking about "I'm thirsty" or "I'm hot" it's not so much ... like that *is* answering the question. Like "Why do you feel

that way?" "oh, because I'm thirsty" it's not a completely just random thing, it actually does have to do with why they feel that way at that time.

Coder N: That's valid... I'm not saying that it's unrelated to it. I do think that "I like my cat" is an "off task" answer and I need to go to the bathroom is related to the task in that it's a feeling that your basic needs are required to do the task... "if the room were cooler I'd be good at math"? I don't know. I guess part of the problem I'm having with it is I feel in some ways it's like if you tell a kid to go to bed they'll ask for a glass of water and I'm not sure if the motive is that I'm actually thirsty or that I'm forestalling the bed thing. But that's reading into the student's response. The other thing about it is I feel both of these are relatively rare? They're a little bit of an edge case so I'm grouping them together.

Coder S: That's fair. It's probably rare enough we can put them into the same category and say alright they're saying something that's not relevant to the task at hand specifically.

Coder N: I think you make an important distinction here, it could be about my cat or it could be about hydration. It could be a basic thing you need or it could also be an unrelated thing. That it is a grouped category here. I think that distinction you make is important if we could find some kind of difference there later on.

Coder D: So for the second half of coding I did split it out. I actually put a little chart to my coding and I have "IDK" and I have "IDK?" And "IDK?" is "I don't know" and "IDK" is "off task". But I also did some of my coding about like if the room is too hot I did "external" + "negative". If it was like I'm hungry it's "internal" + "negative". So I tried to code out these specific differences and "IDK"s, and I think it's important.

Coder N: No, I agree, I think that you're right. Let me think about this, you're doing an internalized state whereas the other is an externalized state. So it's "do you have control over the situation or not?" is generally how I think of internal and external in those terms. So the idea of like wanting a snack and the room being hot, I can't really parse the degree of internal control the student has over those things. Whether they would be capable of asking the teacher to turn down the heat vs if they could ask a teacher "can I go get a granola bar". I'm not 100% sure on what the class permissiveness is as far as those things go or the responsiveness to them. Just because one of them is an external environment and the other is internal to the body I think it's more a matter of locus of control. If something is being done to you or if you're capable of doing something. But I do agree... I think that when you had the "IDK" vs the "IDK?" that's kind of what we're addressing here with "IDK" vs "off task", whereas "IDK?" would be "off task" and "IDK" would be the student saying they don't know how they feel at the time. So I think we're capturing that distinction. I don't know that we're capturing the "internal" and "external" the same way because we're applying a more specific label for some cases of "external" like it being "content" or "difficulty" or "software" or "success" which could be a little bit more internal. So they're a little less abstract than "internal"/ "external" except of course the "off task" which could be unrelated to the task but agnostic about the degree of control students have over that. I think there's less control for those things because teachers are less likely to be willing to do stuff that is not task related. Or grudgingly like "I guess you can go to the bathroom" along those lines? OK. I think I'm going to the

next one if that's OK? Because I've only got about 25 minutes left and I've got 3 more of these to go through. So the second set come down to the 3 questions: "how would you describe your feelings?", "why do you feel that way?" and "what do you wish you could do to improve the class right now?" and the intent I was hoping for on these was the one. The first one is the same sort of forced choice like it's a feelings question, the second is an attributional question like what we had before... The third one is almost trying to get at more of that agency sort of thing "Well if you could control your environment what would you *want* to control?" so it's a little bit like "why do you feel that way?" but also has this agency or aspirational aspect to it. At least that was the goal I was going for. So we're dealing with the first right now which involves students' feelings. And we really similarly to before have a whole lot of "bored". We have a "annoyed" and then "frustrated" kind of thing going on here which is similar to what we had before with "negative" I think that might also work here. Then we also had the "negative" category existed for some people. These are all somewhat correlated but also distinct ideas because annoyance or frustration is a little more feelsey or specifically a feeling. We then had a whole row of things that were like "Ok", "content", "satisfactory" or "neutral" that tended to hang together. That's almost a third category that we didn't see as much of last time, where students would say that things weren't really positive or negative, it tended to have a bit more of a positive valence if someone was saying they were content or satisfied with the situation then they're more likely to be correlated with "I'm feeling good" about something, but it could also just be a neutral state where "well, It's OK I mean it's classwork, I'm working on it".

Then we have the "positive" category here. What's interesting too is there's a different set of "negative" valence. A lot of the ones we saw before were really activating the "annoyed" and "angry" ones. There were also "negative" sort of "deactivating" that Coder S pointed out. I called "depressed", and Coder R called it "upset" where students could have negativity.

Coder S: So I'm a little surprised that my "deactivating" was specifically correlated with those because I think I tagged "bored" with anything that was bored related instead of using "bored" I put a tag of "deactivating" + "negative".

Coder N: Yeah, I'm not going to say that... you're correct about that. And that might be the most frequent thing under bored is deactivating. I think it would occur under both of these scenarios, but I tried to repeat them less. I suppose I had in some cases where Coder SH had multiple variants of frustration? But I think she might have been measuring the first instance of frustration through the fifth instance of frustration? You're correct, that your "deactivating" should also be up under "bored". The question is also if we want it under the same label. I dunno. I disagree a little bit because I feel that bored is distinct from a deactivating emotion, students will sometimes put bored in with multiple exclamation points and be active about describing their boredom or disinterest. But I think what you're describing about a deactivating negative emotion where students are maybe failing at their work and feel bad about it rather than maybe angry about it. It might also come down to an "internal" / "external" thing. Students blaming the system.

Coder S: I feel like I would have said if they had said that they were sad or upset I would have said that was it. I don't remember saying that was "deactivating" but it doesn't seem like it's either "activating" or

“deactivating” exactly. Because it’s not like “frustration” or “anger”, like making your heartrate rise. I dunno, it’s hard to classify. Maybe when I was doing it I was thinking of when you’re sad you’re kind of the opposite arousal of when you’re frustrated.

Coder N: I can pull it really quick. Let me take a quick look here at the main. This is basically the whole sheet of everyone’s codes I'm just going to look for where I said something was “depressing”. [searching noises] ... Depressed would be something like “I'm not good at math” “I need to work harder” is their description of how they ought to do. “Avoidance”, “lack of proficiency”, let me find your column... you may not have used deactivating.... I might be off and just putting them together because I like them. Oh so you’re doing that one as a same similar to what they’re doing the prior time. ... it’s a reasonable time to bring it up: If I find an instance of “same” I'm most likely to duplicate what the prior codes are rather than giving the same code there.

Coder S: That’s fair. I think I just didn’t want to try and look back and find the same student again and see what they said last time. So I just put “same” they’re being non-descriptive.

Coder N: Oh yeah the thing you did was appropriate I would have done it similarly. So you said this was “negative” whereas I said “depressed”. “I don’t think this is much help because I believe that”. And then is there anything you think you can change to make this better? “No”. I put that as “depressed” + “bored” basically. I think I read too much into it, maybe it’s less of an upset or depressed. Or bored might just work for this. [More searching noises]... you know I'm starting to be convinced.... oh here’s one! “my cat died”... ummm “more fish”... which I think is more of an absurd thing. Student messing with us.

Coder S: I guess they did say “my cat died”, that’s pretty depressing, but off topic.

Coder N: Yeah I felt that was definitely a depressing kind of response. Here’s another one who actually says depression. This is a nihilist “nonexistent depression thing” some students said depressed, it was relatively rare though is the other thing I’d say about depressed. Here’s another one “I'm depressed because I'm doing math”. The way I could improve it is by “leaving the situation” just leaving math entirely, that would be good. I feel “bad” “because I got all my questions wrong”. So I think it is somewhat rare that “depression” showed up, it was more common for me. But we’ve got “depressed” and “upset” are relatively rare, it might be better to just put that under “negative”?

Coder N: Hello?

Coder S: Hm?

Coder N: Do you think it might be better to just “negative” or something like that for these?

Coder S: I think that might make sense? I don’t know. I feel like I did something weird on this one, like most people actually had “bored”, but I did the thing with the two dimensional activation and put two tags on everything instead of using “bored” and “annoyed” so...

Coder N: Yeah, but I feel like your stuff makes sense for like some of this second pass. I feel like in some cases it's more descriptive because having a specific thing for "it's too easy" vs "easy" and then also having "I like math" vs "I don't like math". I think these separate tags you have here help and makes it more efficient in terms of coding um... just also I'm trying to preserve what the student intent was and I feel like some things along the lines of.... I don't know! I think it's harder to get students to think in terms of activating/deactivating or internal/external they tend to be a little bit more descriptive but positive/negative tends to be a thing they talk about. I suppose I can drop "depressed" and just go with "negative" here. Let's see what we've got "Bored", "negative", "neutral", "positive", now we're at the part where we're talking about "deactivated" or "depressed" kind of state. I think that maybe we can just do a combo... that just seems like "negative" + "bored"... or... it might also be "hopeless" like some students described they didn't have much opportunity beyond that. But I do think in those cases we're going to be talking about what their attribution is afterward. So maybe negative is just fine for this and we'll take on that in the attribution section. "IDK" which we did before so I think we can keep that because "IDK" is pretty common. Now we also do have this "off task" or "not relevant" or "silly" kind of situation so I think I'm going to preserve the off task we used before for that. And then "confused"... well that would be on the matter of "hard" vs "easy" Students find something hard then they would be "confused." I think for consistency sake it makes sense to keep the "hard" attribution here, that would be fine. So that's not bad we've got away with about.... "bored", "negative", "neutral", "positive", "IDK", "off task", and "hard". A total of only 7 possible attributions here, is there anything you think I'm missing with this?

Coder T: There's the whole being "stressed" or "anxious". I know I saw a couple of those.

Coder N: Okay, so that's interesting that would be an active kind of state there.

Coder T: I think it was something like... I can't think of a specific example now.

Coder N: Oh I can pull it up for you!

Coder T: I had number so my number would have been 4.

Coder D: There was a [uninterpretable] who was referring to her state over and over again.

Coder N: Ok your number would have been 4, right? And Coder D, I will get to you... actually can you repeat that, is that related to what Coder T just said?

Coder D: I said that reiterating her state over and over again.

Coder N: You felt that was one of these anxious situations where she was reiterating her state? So one of these anxious states looks like we have...

Coder S: I have to go to a lecture.

Coder N: You have to go? Ok. Thank you so much for taking so much time out and I understand if other people have to go because we're running up on time right now.

Day 2 Discussion

Coder N: Associated, null, blank, or IDK – so the student doesn't really know how they're feeling or why they feel that way. Or they leave it blank which is another code. Personally I'd prefer to just leave blank as blank and just don't code it and maybe have IDK be a separate thing, but that isn't there. Many students talked about math or the domain, which you know we have a line for right here so the cause of their feelings or attributions be mathematics. Ah, there's a conditional one I think is useful "domain negative: I don't like math" or something along those lines, and just from what we discussed last time: I like the idea of having "negative" be present in these codes, but also have the domain so that you can couple them and have negative but it's also a domain related thing. So you can put a qualifier or modifier on that.

There are a lot of students who talked about feeling good due to success or proficiency. So that was a pretty common thing that students measured during their performance and saying "I am doing well".

Finally, or actually the next thing is that students would also refer to the system. That could be design or bugs that I pointed out, but I'm willing to merge those into "software" related issues for both of those. Website problems, website issues, problems with the site, frustration with the site... I'm not solely saying those are all "bad things"... so coupling up system with positive or negative would give us more descriptive power of "the software is good" or "the software is bad" to me.

One thing that's kind of interesting here is we had both "hard" and "easy" but also "too easy". So once again that seems like having "negative" and "positive" are helpful. In cases where students are saying something is easy but they don't like it they can say "easy" but also "negative" as two codes. Or if they're saying it's "hard" and they don't like it that would be "hard" and also "negative". Otherwise there's cases where students find something challenging and may find it rewarding. I don't think it's very common but it does happen sometimes with students, that students will say something along those lines.

.....

Coder N: because this is a thing that they don't want to do. I don't want to make a guess as to what the... you know that it's not as directly related to cognitively solving math problems or working with the system. I mean it could be that I [the student] am just having a really bad day and that would affect my work with the system without being directly related to it.

Boredom once again is a really common thing that we saw a whole lot of here. And the other things, well these are where we start getting more rare: Coder R and I found pretty commonly that some students would say that they were learning or improving, which is similar to success but a little bit different because you can have a growth mindset with that. And then finally over here I've got a couple

from Coder S and Coder T where you have something that would be almost how I used to code IDK being like “Nonsense” or “Uninterpretable” but it’s a little bit different. IDK can mean “I don’t know why I feel that way” you can also have a student typing like “bbbbbbbbbbbbbbbbbbbb” or just a nonsense set of text that doesn’t seem to be made to communicate something. So... are there any things that I’ve left out....?

Coder D: Can you hear me?

Coder N: Are there any things I’ve left out do people feel?

Coder D: Can you hear me? Can you hear me? Can you hear me? Can you hear me?

Coder R: No, I don’t think so I mean. I kind of just reviewed the coding that I did from before, and.... I like that you have the same similarity of the learning and differentiating that from success. Because you can have a growth mindset vs if the student is only interested in getting things right vs learning from the software. So I think that’s a good idea to distinguish between both of them. I also like the idea of the domain negative and positive where if the person likes math they’re more likely to enjoy the software vs if the person doesn’t like math they might be less inclined to like the software so I tend to agree for the most part with what you have here.

Coder N: Oh cool. I appreciate that. I wasn’t sure if there was anything I was missing but I appreciate that you seem to like what this is. Coder D, do you have any opinion?

Coder D: I just, I sort of indicated it in one place but my “IDK” straight is “I don’t know”, IDK with a question mark which is like random text which is off task. So I kind of have two IDK tags anyway.

Coder N: Cool!

Coder D: I kind of broke it out, like I don’t know why.

Coder N: Do you feel that the “IDK?” could fall under-

Coder D: It’s just “Off task”, my “IDK” is just a straight “I do not know”, right. Instead of doing an “IDK” in an “off task” way that I should have I... I don’t know why. I did a short cut and an “IDK” with a question mark.

Coder N: No, that makes total sense!

Coder D: But anyway I did put it on one of my sheets somewhere, but I just wanted to re... say it again. Just in case. So anyway I do... I mean. Yeah. So that’s why it looks like it might be a typo that some of my IDKs have a question mark with them but I’m just sort of separating between “off task” and “I don’t know”

Coder N: Well I appreciate that. I may have overlooked it or pulled it out because I thought you were uncertain on it, but I was mistaken and it’s kind of neat that we seem to have a fairly similar coding scheme, because you’re doing an off-task for that. One other thing I want to address here because you

have a really common sort of internal and external... I like those. In some ways I feel that they do fall under the software or content or domain being sort of an external thing in many cases... no... hmm

Coder D: I think math... yeah the content. I'm sorry I didn't hear you say "the content". Yeah both the content and the system are external whereas my own personal skill level my confidence my ability are obviously internal. And a lot of these kids... well it was off task... but they were talking about how cold or hungry or tired they were and obviously that's all internal. ... Oh! Well no! Sometimes... well anyway

Coder N: I'm not being critical... well I have to be honest, I am being a little critical.

Coder D: I don't care.

Coder N: I know you don't.

Both laugh

Coder N: The reason I'm going with these instead of the internal/external is it feels like A) it's more into what the student is describing at face value and B) it's kind of specific to a particular instance. That external it's not something that doesn't fall into the category of domain or system... it's usually in reference to those particular things.

Coder D: Sure, however...

Coder N: Feel free to keep discussing.

Coder D: What makes sense to me doesn't need to make sense to you.

Coder N: So I've got IDK, I've content, I've got negative, I've got positive... is that just positive? But I also feel like there's a performance or proficiency or success thing as well... or confidence maybe. Do people have a preference for how I describe positive performance? Actually maybe just performance, because you could have negative performance or positive performance. Then you've got system, hard, easy...

Coder D: If you're going to do syst... oh I'm sorry... no forget it it's already there.

Coder N: Alright you've got one for learning. It looks like you've got one for sort of nonsense... is kind of it or DTG like what I did for my masters stuff.

Coder D: Isn't nonsense off task? Oh okay. I coded nonsense as off task.

Coder N: Well that actually makes sense because they're very close together. If someone says "my cat's breath smells like cat food" vs another person who says "bbbbbbbbbbbbbbbb" both of those are those equivalent things? Or are they not equivalent? One of them can be self-stimming just like random rambling, and the other one could be "I am really depressed that my dad moved out" or something like that. Those are two things that are not directly related to task, but they're kind of different.

Coder D: Right, they're both off task, and one is nonsense. I don't know what that tells you for your research, that you need to know the sub group of off task. Maybe you need to know the sub group of off task messages how much of them are nonsense.

Coder N: Yeah... I'm only up to 12 codes right now which isn't too bad-

Coder D: what I'm trying to say is they both could be coded either/or... to me one's a sub code to another, because all of them could be coded off task but some of those off task codes could be nonsense... so I don't know which way you want to think about it.

Coder N: Yes, I see what you're saying and I think you're correct. Nonsense is off task but off task isn't necessarily nonsense.

Coder D: I'm sorry are you putting 12 codes here that you think that are... that... are.... An utterance would be coded both off task and nonsense or...

Coder N: No... but... I'm not making a clear statement, I'm just trying to make sure I understand your point which seems to be that: If someone says something is nonsense that means it's also a subset of off task, but you can have off task things that are very sensible and not nonsense. Describing it as a subclass I'm just kind of repeating that back.

Coder D: Right!

Coder R: Yeah, but I think the point she was trying to make is that those two categories: they don't inform you a lot as to the research question. Like whether or not it's off task or nonsense, it doesn't really matter whether the student put "ABCD" or "the sky is blue", both of those, they're so interchangeable they don't have a broader effect on the other 9 codes you're trying to do. So even if you have one code for off task, or one code that is nonsense you could just group everything into that. I don't think. I think that she's trying to say that knowing the specifics... that doesn't really inform you of the answer.

Coder D: Well it is though, It is for Coder N. Is that a thing that you want to know about?

Coder N: Well, the thing is. And this is why I have a bit of a soft spot for it, my masters work was on without thinking fastidiously or WTF behaviors where students would just kind of button mash or run in circles or things like that. I've looked at it as a construct before so I find it to be potentially interesting. But I also feel, and this is the thing where I'm indecisive. I think like both of you that there's a super-heading of off-task that this kind of relates to. We distinguished between off task and WTF behaviors in the past because off-task is completely disengaged from the task whereas button mashing can be somewhat task related.

Coder D: You're saying "what the fuck" could be what the fuck about this math problem or this system bugging up... so it's specific to the behaviors.

Coder N: Or just like opening and closing a window rapidly one time after another. I feel like the sort of WTF you're describing is more like an "anger" or "negative at system". But if you're just opening and closing Jane repeatedly really fast, that's a similar thing to this sort of "nonsense" or "uninterpretable" behavior. Where you're maybe bored that's one possibility for it and you just start tapping buttons because it's the thing to do. Which I see as a little bit different than an off-task description of it.

Coder D: So you've answered my question. I didn't realize this was of interest to you. I think it's great. I think that it's interesting, especially through this lens. I didn't realize you were looking at your data this way. So let's move on.

Coder N: I appreciate we talked about this. That's why I do this. I only have my opinion, and I never trust it so that's why I want to bring in this discussion.

I think I'll keep the "nonsense" one, it doesn't hurt to have it.

Open response, most common "Agency" I was at a loss for how to describe this because the term "Agency" because this is a question of "what do you wish you could do to improve this class right now?" So rather than it being a "why do you feel the way you do" I wanted it to be a "what is your hope, your expectation, your plan?" but the problem is that also is a bit of a question of "is it in their power or not within their power?" You could say "I wish that I were better at math" and they may view that as within their power or not within their power. So I put this as this thing that could be either way: "what do you wish you could do to improve this class right now?" So the responses we got here: I think we're getting further and further, it's harder and harder for students to think about this. A lot of students would say that they're unsure "IDK" again because they don't know what to do to improve it. Some students pointed to the content, which could mean that want more challenges or they want to ask for more help or they want different questions, things like that so the content, the material, is what they would like to change about the system. Some of them I said, "disengage", Coder R's "not relevant", Coder C's "Personal", Coder S did "Leave" or "nonsense", and Coder SH did "Quit". These are students who when asked "What could we do to improve the system" said "Well if I weren't in math class right now that would really improve my day. If I could just go home and not do this." We go that a lot. And I think that was also kind of useful, because we're getting a measure of disengagement if what they want is to not be doing it. There's a separate thing here from the content. I used "design" or "learning companion", Coder S mentioned "Display", and Coder SH mentioned "Aesthetics". That includes things like "music" or things like that where students are focusing on how the system is being presented to them. And saying that they would prefer it to be more attractive in some way, or to be more smoothly designed. So they're not referring to the content. They're referring to how the content is presented. Similar to that there was also the call for "I really wish this were more fun" Many of us found this to be a thing which isn't as specific as a particular design concern of the system. Just like "couldn't there be more games in here" "I would really enjoy it if this were more of a fun experience for me". This is where we got down to doing Coder D's internal, because a lot of students had an aspirational "What they would like to change about the system. If they could change anything" to one that was "internal" or about self-betterment. Which I thought that was really kind of neat. That there are some students who rather than saying "I wish the content were presented differently." They're saying "I wish doing better at math, I need to keep working hard". I think that's interesting to note that when they're faced with a challenge,

their idea of what to change is an internal rather than external one. We have a little more system design, but rather than being “oh wouldn’t it be nice if you could design the system differently?” they’re pointing out a particular error or bug in the system. So not that we intentionally designed the system badly, they’re saying “well you have a mistake here” or “this part isn’t working”. So that’s subtly distinct from the idea of “I would like better music” or “I would like a more attractive interface”. To say that “this question is broken.”

And finally we’ve got a whole lot of “the student left this problem blank”. I think that’s everything we’ve got. And you can see there’s fewer codes for this one... only 8 codes is what I’ve got. Am I missing anything and do these seem like apt descriptions of what’s going on here?

Coder D: Looks good to me.

Coder N: ::fixing shared screen:: I should have done this earlier.

Coder R: No I think this covers most of them. One that we both had in common, all three of us actually, the internal aspect of the student wanting to improve their own self rather than having to improve the functionality of the system. There’s a lot of common themes across multiple people. So I think this is pretty good.

Coder N: Coder D, do you feel ok about it? Do you think that the stuff that we’ve got here is OK to work for external. The content and the aesthetics and all that stuff works?

Coder D: Yes.

Appendix H: Python Kappa Program

```
from __future__ import division #makes it so that / is always floating point ("normal") division, and // is
integer division

from datetime import date

from collections import defaultdict, deque

import math

import csv

import itertools

import pandas as pd

import numpy as np

import re

np.set_printoptions(threshold=np.nan)

import re, math, os, sys, pickle #these are default library modules

tagInput = pd.read_csv('Coder NVCoder CForced2FAKE.csv')

print(tagInput)

for x in range (1, len(list(tagInput))):

    tagList = {}

    globals()[list(tagInput)[x]] = tagList

    for y in range (0, tagInput.shape[0]):

        #may need to use actual global name [list(tagInput)[x]] rather than taglist

        if tagInput.iloc[y][list(tagInput)[x]] in tagList:
```

```

tagList[tagInput.iloc[y][list(tagInput)[x]]] += 1
else:
tagList[tagInput.iloc[y][list(tagInput)[x]]] = 1

#print(tagList)
newTagList = {}
sepTagList = {}
delTagList = {}
for key in tagList:
if isinstance(key, str):
key = key.rstrip()

words = ((re.split(re.compile("|".join([" & ", " & ", " & ", " & " ])), key)))
if len(words)>1:
delTagList[key] = tagList[key]
for z in range (0, len(words)):
word = words[z]
#print(key)
#print(tagList[key])
#print(word)
if word in tagList:
#print("***")
#print(tagList[word])
tagList[word] += tagList[key]
#print(word)

```

```
#print(tagList[word])

#print("888")

elif word in sepTagList:

    #print("***")

    #print(sepTagList[word])

    sepTagList[word] += tagList[key]

    #print(word)

    #print(sepTagList[word])

    #print("888")

else:

    sepTagList[word] = tagList[key]

    #print("new tag: ", word)

#print(sepTagList)

newTagList = {**tagList, **sepTagList}

else:

    newTagList = tagList

#print("***")
```

```
#print(newTagList)

#print("%%")

#print(delTagList)

#print("888")

for key in delTagList:

    del newTagList[key]

#print(newTagList)

globals()[list(tagInput)[x]] = {key: newTagList[key] for key in newTagList if not pd.isnull(key)}

#globals()[list(tagInput)[x]] = newTagList

#print("WINWINWINWINWIN")
```

```
#print(tagsA.split('_')[0]+'X'+tagsB.split('_')[0]+'_'+tagsA.split('_')[1])
```

```
def namestr(obj, namespace):
```

```
    return [name for name in namespace if namespace[name] is obj]
```

```
def findMax(comboTagsArray):
```

```
    return sorted([(index, row.index(np.argmax(comboTagsArray)), np.var(row)) for index, row in enumerate(comboTagsArray)])
```

```
if np.amax(comboTagsArray) in row]], key=lambda tup: tup[2], reverse=True)
```

```
def calcKappa(agreeTrue, agreeFalse, aTruebFalse, aFalsebTrue):
```

```
    totalRate = agreeTrue + aTruebFalse + agreeFalse+ aFalsebTrue
```

```
    chanceAgree = (((agreeTrue+aTruebFalse)*(agreeTrue+aFalsebTrue))/ totalRate) +  
    (((agreeFalse+aFalsebTrue)*(agreeFalse+aTruebFalse))/ totalRate)
```

```
    totalAgree = agreeTrue + agreeFalse
```

```
    kappa = ((totalAgree-chanceAgree)/(totalRate-chanceAgree))
```

```
    return kappa
```

```
def findKappas(tagsA1, tagsB1):
```

```
    #Initially the coders' sets of tags are ordered and placed in an array where
```

```
    #the smaller set of possible tags (i.e. "tag lexicon" is set as columns
```

```
    print(tagsA1)
```

```
if(min(len(tagsA1), len(tagsB1)) == len(tagsA1)):
```

```
    #global tagsA
```

```
    tagsA = tagsA1
```

```
    #global tagsB
```

```
    tagsB = tagsB1
```

```
else:
```

```
#global tagsB
tagsB = tagsA1
#global tagsA
tagsA = tagsB1

#print(tagsA1)
#print(tagsB1)
comboTags = [[0 for x in range(len(tagsA))] for y in range(len(tagsB))]

#tagsA = {key: tagsA[key] for key in tagsA if not pd.isnull(key)}
#tagsB = {key: tagsB[key] for key in tagsB if not pd.isnull(key)}
#print(type(tagsA))

#tagsA=sorted(tagsA)
#tagsB=sorted(tagsB)

#print(comboTags)

bTagArray = [[0 for x in range(2) ] for y in range(len(tagsB))]
#print(bTagArray)

#print(tagsA)
#print(tagsB)
```



```
for row in range(0, len(tagInput)):
```

```
    for tagX in range (0, len(tagsA)):
```

```
        for tagY in range (0, len(tagsB)):
```

```
            if(str(tagInput[namestr(tagsB, globals())[0]][row]).find(str(list(tagsB)[tagY])) != -1 and  
               str(tagInput[namestr(tagsA, globals())[0]][row]).find(str(list(tagsA)[tagX])) != -1):
```

```
                bTagArray[tagY][0] = str(list(tagsB)[tagY])
```

```
                bTagArray[tagY][1] += 1
```

```
bTagDict = dict(bTagArray)
```

```
comboTags = [[0 for x in range(len(tagsA))] for y in range(len(bTagDict))]
```

```
#print(bTagDict)
```

```
#The following for loop progresses through each row of the codes in the total coded file then it progresses through
```

```
#each possible tag in each of the two coders' "tag lexicon" dictionaries. It searches the ENTIRE input for any of the
```

```
#tags found in each lexicon and then increments the comboTags array by 1 for each set of paired codes found.
```

```
#this fills the comboTags matrix where the columns are the tags of the smaller "tag lexicon" and the rows are the tags of
```

```
#the larger "tag lexicon" with the raw count of how many times each of those tags co-occur in the total coded file
```

```
for row in range(0, len(tagInput)):
```

```
    for tagX in range (0, len(tagsA)):
```

```
        for tagY in range (0, len(bTagDict)):
```

```
            #print("1", str(tagInput[namestr(tagsB, globals())[0]][row]))
```

```
            #print("2", str(tagInput[namestr(bTagDict, globals())[0]][row]))
```

```
            if(str(tagInput[namestr(tagsB, globals())[0]][row]).find(str(list(bTagDict)[tagY])) != -1 and
```

```
                str(tagInput[namestr(tagsA, globals())[0]][row]).find(str(list(tagsA)[tagX])) != -1):
```

```
comboTags[tagY][tagX] += 1
```

```
#print(comboTags)
```

```
#print(bTagArray)
```

```
outputArray = []
```

```
comboRedux = np.asarray(comboTags)
```

```
tagsARedux = tagsA
```

```
tagsBRedux = bTagDict
```

#The following finds the maximum number in an array of each coder's tags. In cases where there are two maximums

#it selects the maximum from the row with the highest variance because this is the most "distinct" of codes.

#Then it creates a new set of dicts for each coder, and alters the comboTags array to remove the column &

#row which were attributed in the prior row.

```
for tagX in range (0, min(len(tagsARedux), len(tagsBRedux))):
```

```
    firstMax = findMax(np.ndarray.tolist(comboRedux))
```

```
    #print(comboRedux)
```

```
    #print(tagsA[list(tagsA)[firstMax[0][1]]])
```

```

#print(tagsB[list(tagsB)[firstMax[0][0]])

#print(comboRedux[firstMax[0][0]][firstMax[0][1]])

#print(np.sum(comboTags))

#print(bTagDict[list(bTagDict)[firstMax[0][0]])

    kappa = calcKappa(comboRedux[firstMax[0][0]][firstMax[0][1]], np.sum(comboTags) -
tagsA[list(tagsA)[firstMax[0][1]]] - bTagDict[list(bTagDict)[firstMax[0][0]]] +
comboRedux[firstMax[0][0]][firstMax[0][1]],

        tagsA[list(tagsA)[firstMax[0][1]]] - comboRedux[firstMax[0][0]][firstMax[0][1]],
bTagDict[list(bTagDict)[firstMax[0][0]]] - comboRedux[firstMax[0][0]][firstMax[0][1]])

#print(list(tagsBRedux)[firstMax[0][0]], list(tagsARedux)[firstMax[0][1]],
comboRedux[firstMax[0][0]][firstMax[0][1]], kappa)

#print(np.sum(comboTags))

#print(np.sum(comboTags) - tagsA[list(tagsA)[firstMax[0][1]]] -
bTagDict[list(bTagDict)[firstMax[0][0]]] + comboRedux[firstMax[0][0]][firstMax[0][1]])

#print(tagsA[list(tagsA)[firstMax[0][1]]] - comboRedux[firstMax[0][0]][firstMax[0][1]])

#print(list(bTagDict)[firstMax[0][0]])

#print(bTagDict[list(bTagDict)[firstMax[0][0]])

#print(bTagDict[list(bTagDict)[firstMax[0][0]]] - comboRedux[firstMax[0][0]][firstMax[0][1]])

```

```

if(tagX == 0):

    outputArray = [[namestr(tagsB, globals())[0], namestr(tagsA, globals())[0], "Co-
Occurrences", "Kappa"], [list(tagsBRedux)[firstMax[0][0]], list(tagsARedux)[firstMax[0][1]],
comboRedux[firstMax[0][0]][firstMax[0][1]], kappa]]

    #if(tagX == 0):

        #outputArray = [[namestr(tagsBRedux, )[0], namestr(tagsARedux, )[0], "Co-
Occurrences", "Kappa"], [list(tagsBRedux)[firstMax[0][0]], list(tagsARedux)[firstMax[0][1]],
comboRedux[firstMax[0][0]][firstMax[0][1]], kappa]]

elif(tagX > 0): outputArray = np.append(outputArray, [[list(tagsBRedux)[firstMax[0][0]],
list(tagsARedux)[firstMax[0][1]], comboRedux[firstMax[0][0]][firstMax[0][1]], kappa]], axis = 0)

#print(outputArray)

comboRedux1 = np.delete(np.delete(comboRedux, firstMax[0][0], 0), firstMax[0][1], 1)
comboRedux = comboRedux1

del tagsARedux[list(tagsARedux)[firstMax[0][1]]]
del tagsBRedux[list(tagsBRedux)[firstMax[0][0]]]

for x in range (1, len(list(tagInput))):

    tagList = {}

    globals()[list(tagInput)[x]] = tagList

for y in range (0, tagInput.shape[0]):

    #may need to use actual global name [list(tagInput)[x]] rather than taglist

```

```

if tagInput.iloc[y][list(tagInput)[x]] in tagList:
    tagList[tagInput.iloc[y][list(tagInput)[x]]] += 1
else:
    tagList[tagInput.iloc[y][list(tagInput)[x]]] = 1

#print(tagList)

newTagList = {}
sepTagList = {}
delTagList = {}
for key in tagList:
    if isinstance(key, str):

        words = ((re.split(re.compile("|".join([" & ", " & ", " & ", " & " ])), key)))

        if len(words)>1:

            delTagList[key] = tagList[key]

            for z in range (0, len(words)):

                word = words[z]

                #print(key)

                #print(tagList[key])

                #print(word)

                if word in tagList:

                    #print("****")

                    #print(tagList[word])

                    tagList[word] += tagList[key]

```

```
#print(word)
```

```
#print(tagList[word])
```

```
#print("888")
```

```
elif word in sepTagList:
```

```
#print("***")
```

```
#print(sepTagList[word])
```

```
    sepTagList[word] += tagList[key]
```

```
#print(word)
```

```
#print(sepTagList[word])
```

```
#print("888")
```

```
else:
```

```
    sepTagList[word] = tagList[key]
```

```
#print("new tag: ", word)
```

```
#print(sepTagList)
```

```
    newTagList = {**tagList, **sepTagList}
```

```
else:
```

```
    newTagList = tagList
```

```
#print("****")
#print(newTagList)
#print("%%")
#print(delTagList)
#print("888")
    for key in delTagList:
        del newTagList[key]
#print(newTagList)
    globals()[list(tagInput)[x]] = {key: newTagList[key] for key in newTagList if not pd.isnull(key)}
    #globals()[list(tagInput)[x]] = newTagList

return(outputArray)
```

```
#codeArray = findKappas(Coder S_Attributions, Coder T_Attributions)
```

```
codeArray = findKappas(Coder N_Forced_Attributions, Coder C_Forced_Attributions)
```



```
#codeArray = np.append(codeArray, findKappas(Coder N_Forced_Attributions, Coder  
C_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder N_Forced_Attributions, Coder  
S_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder N_Forced_Attributions, Coder  
SH_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder N_Forced_Attributions, Coder  
T_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder N_Forced_Attributions, Coder  
D_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder R_Forced_Attributions, Coder  
C_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder R_Forced_Attributions, Coder  
S_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder R_Forced_Attributions, Coder  
SH_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder R_Forced_Attributions, Coder  
T_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder R_Forced_Attributions, Coder  
D_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder C_Forced_Attributions, Coder  
S_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder C_Forced_Attributions, Coder  
SH_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder C_Forced_Attributions, Coder  
T_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder C_Forced_Attributions, Coder  
D_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder S_Forced_Attributions, Coder  
SH_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder S_Forced_Attributions, Coder  
T_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder S_Forced_Attributions, Coder  
D_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder SH_Forced_Attributions, Coder  
T_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder SH_Forced_Attributions, Coder  
D_Forced_Attributions), axis=0)
```

```
#codeArray = np.append(codeArray, findKappas(Coder T_Forced_Attributions, Coder  
D_Forced_Attributions), axis=0)
```

```
print(codeArray)
```

```
#firstMax = findMax(comboRedux)
```

```
#comboRedux = np.delete(np.delete(comboRedux, firstMax[0][0], 0), firstMax[0][1], 1)
```

```
#del tagsARedux[list(tagsARedux)[firstMax[0][1]]]
```

```
#del tagsBRedux[list(tagsBRedux)[firstMax[0][0]]]
```

```
#print(comboRedux)
```

```
#print(tagsARedux)
```

```
#print(tagsBRedux)
```

```
#print(firstMax)
```

```
#print(firstMax[0][0], firstMax[0][1])
```

```
#print(list(tagsB)[firstMax[0][0]], list(tagsA)[firstMax[0][1]], comboTags[firstMax[0][0]][firstMax[0][1]])
```

Appendix I: Cognitive Causal Attributions for Annoyance/Frustration

Broken Website/Repeat Questions (N=14)

- this website isnt good. and they dont have the right answers
- you keep asking the same questions
- because i am being asked repeated questions
- because i am getting repeated questions
- because i would put in a right answer and it was wrong
- i am getting questions that dont make sense or repeated questions
- i had a question about a negative multiplied by a negative and its supposed to equal a positive but instead it equaled another negative
- i just got these questions
- i keep getting the same problems over and over
- I'm being asked the same questions
- nothings coming up
- This is filled with bugs
- this is really slow

Dislike of Website (N=8)

- This website is aggy
- this website is annoying sometimes
- because i dont like this website i like it WAY better
- cause this site sucks. I think this site needs some special help
- AKA PLATANOS
- because this websites retarded and makes me triggered!!!
- Jane won't go away
- Because this is repetitive and boring

Dislike of Topic/Math (N=5)

- i dont like fractions at all .
- I DONT LIKE THIS TOPIC
- i dislike fractions , decimals , and percentages
- i hate math
- i strongly hate fractions

Too Hard (N=5)

- because i am not able to understand the problem
- because i keep getting questions wrong
- ITS HARDDDD!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
- please stop giving me this hard es muy malo and it hurts my head ill offer you free cheerios if you stop giving me this stuff..... - Yours Truly Jordan Jeeveruthnam Moodley :) ps: if you dont stop giving me these problems im gonna lose my mind
- this is too hard and too spooky 4 me plz give easier problems thx ill give you free cheerios'