Blinded by Science?: Exploring Affective Meaning in Students' Own Words

Sarah E. Schultz[1](✉), Naomi Wixon[1], Danielle Allessio[2], Kasia Muldner[3], Winslow Burleson[4], Beverly Woolf[2], and Ivon Arroyo[1]

[1] Worcester Polytechnic Institute, Worcester, MA, USA
{seschultz,mwixon,iarroyo}@wpi.edu
[2] University of Massachusetts, Amherst, MA, USA
allessio@educ.umass.edu, bev@cs.umass.edu
[3] Carleton University, Ottawa, ON, Canada
kasia.muldner@carleton.ca
[4] New York University, New York, NY, USA
wb50@nyu.edu

Abstract. This work addresses students' open responses on causal attributions of their self-reported affective states. We use qualitative thematic data analysis techniques to develop a coding scheme by identifying common themes in students' self-reported attributions. We then applied this scheme to a larger set of student reports. Analysis shows that students' reasons for reporting a certain affect do not always align with researchers' expectations. In particular, we discovered that a sizable group of students externalize their affect, attributing perceived difficulty of the problem and their own negativity as lying outside of themselves.

1 Introduction

When an adaptive tutor, MathSpring, asked a student to explain the reason for her self-reported affect of high interest, she typed in "because i think i will learn a lot of new information in this website about math. *[sic]*" This student was explaining that her interest stems from anticipating progress in her learning. This type of open-response data, however, is not typically collected in studies on student affect. Instead, as described in [4], existing methods in the Intelligent Tutoring Systems (ITS) community generally focus on classifying affect categorically, such as "interest," "frustration," "excitement," and so on. While this information is valuable, the way that students define these terms may not be the same as what researchers believe the terms mean [3], so it will be valuable to find out *why* students report feeling a certain way.

To be scalable, the classification of student open-ended responses would need to be done automatically by relying on natural language processing (NLP) techniques. However, before investing in the design of such technologies, it is necessary to determine how much of an advantage examining these reports truly offers.

In this work, we examine data from two populations of students interacting with the MathSpring system (formerly Wayang Outpost). The system periodically asked the

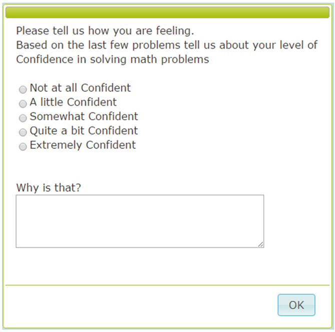
A. Micarelli et al. (Eds.): ITS 2016, LNCS 9684, pp. 314–319, 2016.
DOI: 10.1007/978-3-319-39583-8_35

students to (1) self-report on their affect using a Likert scale and (2) explain why they were feeling that way using an open-ended response dialog box (Fig. 1). We use a qualitative approach to create and apply a coding scheme to students' self-reported explanations and then use data mining techniques to explore what type of student responses different prompts and reports of affect were likely to elicit. Additionally, we seek to determine whether students' reports of reasons for feeling a certain way align with our expectations, as researchers, of what types of attributions occur with different types of affect.

2 Methods

This research was conducted using data from two studies using the MathSpring ITS (formerly Wayang Outpost) [1]. The studies were run in 2011 with 7th and 8th grade students (N = 123), and 2015 with 7th grade students (N = 209), in Massachusetts and California respectively.

To obtain in situ information about student affect, MathSpring prompted reports every five minutes or every eight problems, whichever came first without interrupting a problem. Students were asked to report on a target emotion (e.g., interest, excitement) via a 5-point Likert scale, and to explain why they were feeling that way (Fig. 1).



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?

Fig. 1. Student self-report of affect. Open prompt (bottom) asked students to explain why they felt that way.

We use two types of qualitative thematic data analysis; open coding (phase 1), where coders independently code student report data with little direction, and axial coding [2], where core categories are developed based on coders' open coding schemes (phase 2). Our third phase consists of validating those sub-categories (which we use as our tags) through inter-rater reliability as measured with Cohen's kappa.

Phase 1: Open Coding. Our first step was "open coding," [2] wherein coders parse and reflect on data with the goal of naming and categorizing phenomena that occur within. Here, a set of 450 randomly selected open responses was gathered from a dataset collected in 2011 and given to the five coders (the first four authors, and one additional coder). The coders were told how the data was collected, but were not given a coding scheme apart from the directive to independently arrive at a set of approximately 10 categories that would encompass approximately 70 % of the responses. Coders were instructed that they could tag a response with multiple tags if they felt they were applicable.

Phase 2: Axial Coding. Once all five coders created their schemes and tagged the data, we entered the "axial coding" phase [2], in which the 2nd, 3rd, and 4th authors reviewed the independently devised coding schemes to determine simple commonalities among them. Where similar categories were created by most coders, they were merged into a single category. Ten final categories were determined, as follows, with examples of responses that were tagged with each attribution and abbreviations in parentheses:

- IDK (idk) – doesn't understand why they feel the way they feel (thus "IDK" for "I don't know") or doesn't want to tell us why. (e.g., "???????????????????", "meh").
- Boring (bor) – describes something as boring (e.g., "math is boring").
- Easy (easy) – says the material is easy (e.g., "too easy," "its simple" [*sic*]).
- Hard (hard) – says the material is hard or difficult (e.g., "it is a little confusing and hard", "it gives me a good challenge").
- Internal (int) – attributes their feelings to internal causes (e.g., "i am smart" [*sic*]).
- External (ext) – attributes their feelings to external causes (e.g., "It is kind of fun").
- Positive (pos) – the valence is positive (e.g., "I like this program").
- Negative (neg) – the valence is negative (e.g., "I hate math").
- Supportive (sup) – feels supported (e.g., "It is fun but it also helps me learn a lot.").
- Unsupportive (unsup) – does not feel supported (e.g., "is not helpful").

Phase 3: Application and Validation of Tags. The coding scheme was applied by the first four authors to the 2015 dataset, coding each response. After tagging student responses, inter-rater reliability was determined by Cohen's kappa. The highest agreement between any two authors is displayed in Table 1. Overall, the second author had highest agreement with other authors, so their tags were used for analyses.

Table 1. Best inter-rater reliability kappa values

Code	idk	bor	easy	hard	int	ext	pos	neg	sup	unsup
Kappa	0.80	0.79	0.9	0.78	0.84	0.69	0.84	0.78	N/a	0.55

3 Results

First we report the frequency of each tag in all self-reports (Fig. 2). Then we examine the frequency of each tag given a particular self-reported affect state (e.g., "high frustration"). Forced choice Likert response of <3 was categorized as low, while >3 was categorized as high. Neutral responses (3 on the Likert scale) were not included in this analysis. Finally, we examine frequent combinations of tags over all affect reports.

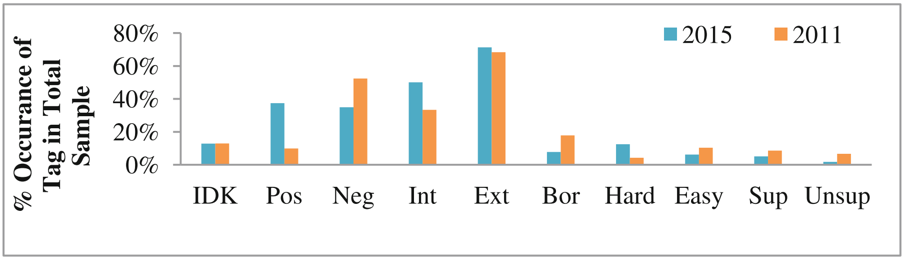


Fig. 2. Frequency of each code out of a total sample group (2015 N = 449; 2011 N = 464) (Color figure online)

The second dataset, from 2015, was collected after several improvements were implemented in MathSpring, e.g., a screen showing student progress. Perhaps in part due to the improvements in MathSpring, the 2015 participants' reports (on average) were more often positive and less often negative in valence. Additionally, their reports were more frequently attributed to internal causes. They were less likely to report boredom, and were more likely to describe their work as hard or challenging. Figure 2 shows how frequently each tag occurred in each dataset.

Table 2 shows the frequency of each tag for each affect report (note: this includes only reports for which students included both a non-neutral affect rating and a text response). In 2015, only prompts on excitement and interest were given. Table 3 shows how often each tag occurred for each type of report, similar to Table 2.

Table 2. Percentage of tagged reports of affect containing each tag 2011

	IDK	Pos	Neg	Int	Ext	Sup	Unsup	Easy	Hard	Bor
Low Exc.	3.1	1.9	29.0	14.5	30.9	3.8	3.8	1.9	0.8	10.3
High Exc.	0.0	18.9	5.4	18.9	37.8	5.4	0.0	5.4	8.1	0.0
Low Int.	7.2	1.3	23.8	11.5	31.5	3.4	3.4	5.5	1.7	10.6
High Int.	8.7	21.7	8.7	21.7	26.0	8.7	0.0	0.0	4.3	0.0
Low Conf.	8.9	0.9	28.6	18.8	25.9	4.5	3.6	0.9	1.8	6.3
High Conf.	10.5	12.1	8.9	15.3	29.0	1.6	0.0	13.7	2.4	6.5
Low Frust.	4.9	12.3	9.9	19.8	29.6	2.5	0.0	12.3	1.2	7.4
High Frust.	3.6	0.0	33.1	13.0	31.4	5.3	5.3	0.0	2.4	5.9

Many of the tags appear frequently where expected, for example the "positive" tag appears most frequently in reports of high interest and high excitement. Some others are more surprising, such as the "external" tag, which is frequent in all reports and most frequent in reports of high excitement in 2011 and low interest in 2015.

Additionally, the modular nature of attributions for emotions can be combined to provide more complex meanings. For example a reason such as "Math is easy!" would get the tags "ext, easy" or external attribution and easy. Table 4 shows the most frequent tag combinations for each dataset. Reports where the student chose the neutral

Table 3. Percentage of tagged reports of affect containing each tag 2015

	IDK	Pos	Neg	Int	Ext	Sup	Unsup	Easy	Hard	Bor
Low Exc.	4.8	5.4	23.1	20.4	29.4	4.2	1.5	1.8	5.7	3.6
High Exc.	5.2	31.4	2.9	22.9	29	2.4	0.0	2.9	3.3	0.0
Low Int.	5.0	5.9	20.8	17.1	29.5	3.7	0.9	2.2	7.8	7.1
High Int.	6.4	27.9	3.0	23.2	28.3	5.2	0.0	3.9	2.1	0.0

Table 4. Most frequent tag combinations and frequency of occurrence for 2015 and 2011 datasets

Code combination	pos int ext	neg int ext	idk	ext easy	neg ext bor	neg ext	pos ext	pos int
Freq 2015	18.1	12.5	11.4	4.2	3.1	2.9	7.3	6.6
Freq 2011	3.3	11.4	13.7	4.1	2.6	14.4		

Code combination	ext hard	No code	ext bor	neg ext unsup	neg int	idk neg	neg int ext hard	Total (except no code)
Freq 2015	4.7						3.1	73.8
Freq 2011		5.7	4.5	3.1	2.5	1.4		62.8

"3," or did not choose a rating of affect, but did include an open response are included here. If a code combination appears in only one dataset, its frequency is left blank in the other column.

While there are a total of 511 possible combinations, the given complex tag combinations were able to cover 62.8 % of all possible instances for 2011, and 73.8 % for 2015.

Some of the most common instances were "pos, int, ext" and "neg, int, ext" which often included a reference to a relationship between self and an external entity with an associated valence (e.g., "I like this program" or "I hate math").

One interesting observation is that while "negative" co-occurs with "external" in many of the above combinations, including those two alone, it never co-occurs with "internal" unless "external" is also part of the combination. This indicates that negative attributions are more often blamed on external reasons, such as the software or the domain, rather than on one's self (unless it is a relationship between self and other).

Finally, the other very frequent category was "idk" or "I don't know." When this appeared, it was most often the only tag given to the statement (e.g., "because I'm just not"), but it also sometimes co-occurs with "negative;" in these cases, students may offer no explanation in a manner that is hostile towards the system (e.g., "None of your business!").

4 Discussion and Future Work

In general, hand coding and analyzing student text about specific affect self-reports has enabled us us to explore reasons/attributions that further describe particular affective states. We have found similarities across disparate datasets for students that seemed to have had somewhat different experiences; not only in expected areas such as valence, but also with regard to intrinsic vs. extrinsic attributions and difficulty. For example, we were surprised that in one dataset students were more likely to report that the material was "hard" when reporting high interest or excitement than when reporting low interest or excitement, while in the other dataset the opposite was true.

One of the most overwhelming findings of this work is the prevalence of students who externalize their affect, especially negative valence emotions. It is important that we recognize the existence of this sizable group of students. In the future, we plan to look more in-depth at each of these areas in order to understand each of these groups of students better, as well as the relationship between their emotions and their reasons for them.

An advantage to our coding scheme is that it has also prompted us to think about possible new affective constructs. We can attempt to build models predicting these reason tags using highly contextualized features, instead of looking at emotions labels. This would imply inspecting the relationship between attributions and students' contextualized performance and behaviors to see if these attributions may be responsible for different behaviors, and vice versa.

Acknowledgements. Thanks to Jaclyn Ocumpaugh for all her support and help and to Samantha Tapia for helping as an additional coder. This research was funded by the National Science Foundation, #1324385, Cyberlearning DIP, Impact of Adaptive Interventions on Student Affect, Performance, and Learning; Burleson, Arroyo and Woolf (PIs). Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Arroyo, I., Beal, C.R., Murray, T., Walles, R., Park Woolf, B.: Web-based intelligent multimedia tutoring for high stakes achievement tests. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 468–477. Springer, Heidelberg (2004)
2. Corbin, J.M., Strauss, A.: Grounded theory research: procedures, canons, and evaluative criteria. Qual. Sociol. **13**(1), 3–21 (1990)
3. Wixon, D.A., Ocumpaugh, J., Woolf, B., Burleson, W., Arroyo, I.: La Mort du Chercheur: how well do students' subjective understandings of affective representations used in self-report align with one another's, and researchers'? In: International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015), p. 34 (2015)
4. Porayska-Pomsta, K., Mavrikis, M., Pain, H.: Diagnosing and acting on student affect: the tutor's perspective. User Model. User-Adap. Interact. **18**(1–2), 125–173 (2008)