Sequences of Frustration and Confusion, and Learning

Zhongxiu Liu
Visit Pataranutaporn
Jaclyn Ocumpaugh
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
{zhongxiuliu,vpatara,jocumpaugh}@wpi.edu

Ryan S.J.d. Baker
Teachers College, Columbia University
525 W. 120th St.
New York, NY 10027
+1 (212) 678-8329
baker2@exchange.tc.columbia.edu

ABSTRACT

In this paper, we use sensor-free affect detection [4] and a discovery with models approach to explore the relationship between affect occurring over varying durations and learning outcomes among students using Cognitive Tutor Algebra. Researchers have suggested that the affective state of confusion can have positive effects on learning as long as students are able to resolve their confusion [10, 22], and recent research seems to accord with this hypothesis [17]. However, there is some room for concern that some of this earlier work may have conflated frustration and confusion. We replicate these analyses using sensor-free automated detectors trained to distinguish these two affective states. Our analyses suggest that the effect may be stronger for frustration than confusion, but is strongest when these two affective states are taken together. Implications for these findings, including the role of confusion and frustration in online learning, are discussed.

Keywords

Affect, confusion, frustration, affect sequences, affect detection, learning outcomes, discovery with models, affective persistence

1. INTRODUCTION

Affect has become an area of considerable interest within research on interactive learning environments [1, 10, 11, 18, 23]. Though findings relating boredom and engaged concentration to learning have largely accorded to prior hypotheses, there have been surprising patterns of results for other affective states, with unstable effects for confusion between studies and often no effects for frustration [7, 21].

However, many of these early studies investigated overall proportions of affective states, rather than considering the potential differential impacts of affect manifesting in different ways. It may be important to consider the multiple ways a specific affective state can manifest, especially considering that there can be considerable variance in how long an affective state lasts [8], affect may be influenced by behavior and vice-versa [3, 5] and some affective states may not be unitary in nature (for example, [12] refers to "pleasurable frustration," which is presumably different in nature than the non-pleasurable frustration often discussed in the research literature).

This puzzle is of particular interest for the affective state referred to as confusion. While relationships between boredom and learning, and engaged concentration and learning, often follow hypothesized patterns [7, 21], confusion appears to manifest in unstable ways across studies. For example, while [7] and [9] find a positive relationship between confusion and learning, with an experimental intervention in the case of [9], [21] finds a negative relationship. Frustration, somewhat surprisingly, routinely does not appear to be correlated with differences in learning outcomes [7, 21].

One possibility is that these results — particularly the results for confusion — may be based on insufficient information. That is, the overall prevalence of an affective state may not accurately predict its impact; how it manifests matters. As [22] notes, students who become confused may either deliberate until they resolve their confusion or become hopelessly "stuck" in unresolved confusion; the former situation has been hypothesized to help learning while the latter undercuts student achievement [22]. As such, the duration of a student's state of confusion may be meaningful. Under this hypothesis, the longer a student remains confused, the less likely they are to resolve that confusion [22]. [10] suggests that confusion may have a dual nature when considered as an affective state: it is possible for it to trigger either persistence (engagement) or resistance to the learning process.

These hypotheses were investigated in Lee et al. [17], who analyzed students' affect over time as the students learned introductory computer programming. Lee and colleagues broke down students' compilation behaviors within this context into sequences of 8 compilations within the learning software, and used text replays [2] to label student behavior in terms of whether the student was thought to be confused. They then developed a data-mined model based on these labels, and distilled its outputs into sequences of two or three consecutive affective states (confused or not confused). They then correlated each student's proportion of these sequences with the student's mid-term exam scores. This test was given after the learning activity studied.

Lee et al. found evidence that short-term confusion that resolves seems to impact learning positively, whereas prolonged confusion affects learning negatively [17]. They found a fairly strong negative relationship between prolonged confusion (three measurements of confusion in a row) and learning (r=-0.337), while students who had brief periods of confusion followed by extended periods where the student was not confused had more positive learning (r=.233).

The results in [17] are intriguing, and show the benefits of this type of fine-grained analysis. However, there are some limitations to this study that may reduce confidence in its findings and therefore call for replication and clarification. (These limitations

were pointed out by the anonymous reviewers at the time of submission of [17]). One key potential limitation was that the operational definition of confusion used in [17] differs substantially from that used in prior research on affect and learning [3, 7, 21]. In [17], clips were coded as confused based on extended student difficulty, for example when a student failed to resolve an error on several consecutive programming compilations. It is not clear that these inferences capture confusion in the same sense that is traditionally described in the affect literature. In particular, this behavior and other aspects of the operational definition of confusion in [17] may have incorporated instances of frustration as well as confusion. This potential limitation was due to the approach used to label confusion; the human coders in [17] inferred and hand-labeled affect solely from a fairly limited subset of the information available in log files, as opposed to the field observations or video observations used in other work, each of which leverage more information to discriminate affect. While the text replay method has been shown to be reliable for inferring behaviors [2, 24], its use in affect labeling is relatively more experimental and may be more open to question.

Another limitation in this early work is that the measure of learning used (a mid-term exam) was not grounded in any measure of students' knowledge prior to the learning activity. As such, this work assumes that specific affective patterns led to student success, but it is equally possible that student prior knowledge led both to those affective patterns and to high scores on the mid-term.

In this paper, we build on this work, replicating it but extending it to address these concerns by incorporating models specifically tailored to distinguish confusion and frustration and by adding a pre-test. By doing so, we can better understand the relationship between duration of affect and student learning outcomes. In these analyses, we consider confusion and frustration taken independently, as well as the union of these two affective states (which in our current view may have been what was assessed in [17]).

2. METHODS

2.1 Tutor Studied

The learning system used in this study was Cognitive Tutor Algebra I, an interactive learning environment now used by approximately 500,000 students a year in the USA. The students in this study used a lesson on systems of algebraic equations as part of their regular mathematics curriculum. In Cognitive Tutors, students solve problems with exercises and feedback chosen based on a model of which skills the student possesses. Cognitive Tutor Algebra has been shown to significantly improve student performance on standardized exams and tests of problem-solving skill [14].

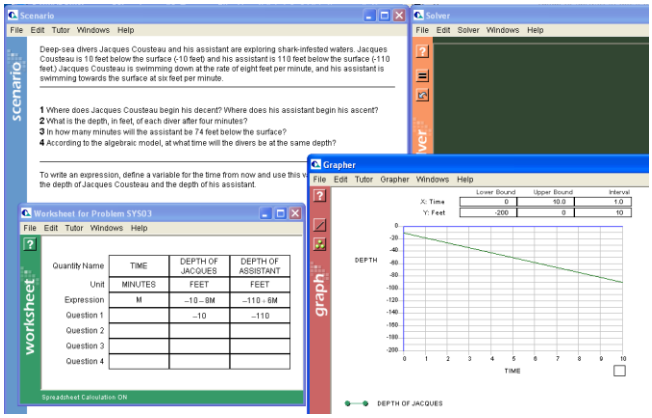

Figure 1: The Systems of Equations A lesson, from Cognitive Tutor Algebra I, used in this study.

2.2 Data Set

Data were collected from 89 students in rural Western Pennsylvania (the data presented here was also discussed in [4], where affect detectors were presented for this data; these affect detectors are in turn used in this paper, in "discovery with models" analyses). Compared with the state's average, students at this high school had a higher average on the PSSA standardized exam, were less likely to be a member of ethnic minority group, and were less likely to be eligible for free or reduced-price lunch. They were well-balanced for gender.

Each student in this study participated in a learning session using the Systems of Equations A lesson of Cognitive Tutor Algebra, which focuses on learning to graph and solve systems of equations. Each student used the tutor software for two class sessions. Tutor activities were preceded and followed by pre-test and post-test measures of learning. (Four students who did not complete all three of these activities were later excluded.) The average pre-test score was 75.2% (SD = 25.3%), and the average post-test score was 79.8% (SD = 23.5%).

During the learning session, two expert field observers coded students' affect following the protocol outlined in [19]. Within this protocol, holistic observations are conducted based on a combination of facial expression, posture, actions within the software, context, and other factors. Confusion and frustration are distinguished, with a key difference being that frustration involves negatively-valenced affect often combined with expressions of dissatisfaction or anger, whereas confusion is a less negative experience. Though the two states are relatively similar conceptually, typically they have not been challenging for observers to distinguish within this protocol; boredom and confusion have more often been the source of disagreement between coders [19]. Observations are conducted in a pre-determined order, with an approach designed to minimize observer effects and to sample evenly across students during the period of observation, both in terms of number of observations per student, and the time when observations occur.

After field observations were collected, they were synchronized with features distilled from interaction log data, and detectors were constructed and validated for several affect categories, two of which (confusion and frustration) will be used in this study. Complete detail on the automated detectors is given in [4]. In brief, the frustration detector was generated at using the REPTree algorithm, achieving a Kappa of 0.23 and an A' of 0.64, under student-level cross-validation. The confusion detector was produced using JRip, achieving a Kappa of 0.40 and an A' of 0.71, under student-level cross-validation. Note that the values of A' given here are lower than in [4]; these represent the exact same detectors, but the values of A' given in that earlier work were computed using the implementation in RapidMiner 4.6, which was afterwards discovered to be buggy. The values given here are re-computed using the Wilcoxon interpretation of A' rather than the AUC interpretation, using code at http://www.columbia.edu/~rsb2162/computeAPrime.zip.

In the study presented in the current paper, automated detectors were used in order to achieve repeated measurements of a student's affect over relatively brief periods of time, while avoiding observer effects (although the protocol in [19] is designed to be non-intrusive, and to reduce observer effects, continually observing a student over extended periods of time increases the probability that the student will notice that they are being observed and change their behavior). Labels were generated by automated detectors at the level of 20-second intervals of student behavior, termed clips. The grain-size of 20-seconds was selected because this matches the original length of the field observations used to create the detectors. Problem boundaries and other events were not considered when clips were created. While it could be argued that it is better to avoid allowing clips to extend across problem boundaries, affect may extend across these events, and avoiding these transitions may give a less representative picture of overall student affect. A total of 29,777 clips were generated across the students' use of the tutoring software.

Three applications of these detectors are studied. The first application uses only the confusion detector, labeling clips as either confused (C) or not (N), splitting students based on a 50% confidence cut-off. The second application uses only the frustration detector, labeling clips as either frustrated (F) or not (N), also splitting students based on a 50% confidence cut-off. The third applies both detectors simultaneously, and considers a clip as confused/frustrated (referred to as A for "Any" below) if either detector had confidence over 50%. This third application, in our view, may map best to the approach taken in [17].

Once clips were labeled, they were segmented into sequences of three consecutive states. These sequences were chosen to be comparable to the 3-step sequences in [17], but represent a finer level of granularity because of the shorter duration of clips in this work (20 seconds versus 8 compilations, which can take several minutes). Potential sequences for each application are included with their frequencies in Tables 1-3.

Table 1. Possible Sequences for Confusion, with Frequencies (%)

NNN	NNC	NCN	NCC	CNN	CNC	CCN	CCC
93.78	1.91	1.74	0.23	1.84	0.09	0.23	0.16

Table 2. Possible Sequences for Frustration, with Frequencies (%)

NNN	NNF	NFN	NFF	FNN	FNF	FFN	FFF
96.20	1.16	1.09	0.14	1.15	0.08	0.14	0.04

Table 3. Possible Sequences for "Any" (Unified Confusion/Frustration), with Frequencies (%)

NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
90.25	2.94	2.70	0.41	2.86	0.20	0.40	0.24

Once detectors were applied, the relative frequency of each sequence was compared to several learning measures, including pretest scores, posttest scores, and the difference between the two. Because the number of tests introduces the potential of spurious effects, the Benjamini & Hochberg (B&H) adjustment [6] is used as a post-hoc control. This method does not guarantee each test's significance, but it does guarantee a low overall proportion of false positives, while preventing the substantial over-conservativism found in methods such as the Bonferroni correction [cf. 20].

In this study, we consider two levels of baseline statistical significance (α=0.05 or 0.1) for the Benjamini & Hochberg adjustment. The 0.05 level reflects full statistical significance, whereas 0.1 reflects marginal significance. Within the B&H adjustment, each test retains its original statistical significance, but the α value cutoff for significance changes depending on the order of the test in significance among the tests run. For understandability, adjusted significance cutoffs are given in tables below for all tests run.

3. RESULTS

3.1 Duration of Affect and Learning Gains

In this section, we compare the relative frequency of sequences of confusion and frustration to assessments of gains in student learning over time. Learning gains are computed as post-pre; the alternate metric of (post-pre)/(1-pre) is difficult to interpret when some students obtain pre-test scores of 100%, which were seen in this data set. In order to understand the importance of individual patterns, we apply separate significance tests for each pattern (with post-hoc controls as discussed below), rather than building a unitary model to predict learning gains from a student's combination set of sequences.

Results for confusion diverged considerably from what might be predicted based on previous research. As shown in Table 4, only three of eight possible sequences showed marginal significance when correlated with confusion, and all of these effects disappeared after post-hoc controls were applied. That is, contrary to theoretical predictions [10, 22], and the interpretation of the findings in [17], differences in sequences of affective state of confusion do not appear to be associated with learning gains in this data.

Table 4. Confusion vs. Learning Gains (No results remain significant after post-hoc control)

3-step - diff	r	p	p cutoff (sig)	p cutoff (marginal)
NNC	0.21	0.054	0.00625	0.0125
CNC	0.198	0.070	0.0125	0.025
NNN	-0.181	0.097	0.01875	0.0375
NCN	0.179	0.101	0.025	0.05
CNN	0.157	0.151	0.03125	0.0625
NCC	0.149	0.173	0.0375	0.075
CCN	0.131	0.231	0.04375	0.0875
CCC	-0.049	0.654	0.05	0.1

By contrast, frustration (Table 5) shows several correlations with learning gains that remain marginally statistically significant after post hoc adjustments. Interestingly, the patterns for frustration match those reported for confusion in [17]. Namely, extended (3-step) periods of no frustration (NNN) are negatively correlated with learning gains. That is, 60 seconds without frustration

negatively impacts learning. Introducing one 20-second interval of frustration (as in NFN, NNF, FNN, and FNF) seems to improve learning outcomes (r=0.273, 0.25, 0.248, and 0.208, respectively), but this effect is reduced or eliminated if the sequence contains two intervals of frustration. Only one sequence with two intervals of frustration (FNF) remains marginally significant after post-hoc adjustment, but with a lower effect-size (r=0.208) than those with only one interval of frustration. These results accord with those for confusion in [17].

As such, one possible explanation is that the construct primarily being inferred in [17] was frustration. The findings seen here match well if that assumption is made; they do not match well, if the codes in [17] genuinely reflected the affective state of confusion. We will discuss this possibility further in section 3.3.

Table 5. Frustration vs. Learning Gains (Significant results are in dark gray; marginally significant results are in light gray)

3-step - diff	r	p	p cutoff (sig)	p cutoff (marginal)
NFN	0.273	0.011	0.00625	0.0125
NNN	-0.262	0.016	0.0125	0.025
NNF	0.25	0.021	0.01875	0.0375
FNN	0.248	0.022	0.025	0.05
FNF	0.208	0.056	0.03125	0.0625
FFF	0.174	0.111	0.0375	0.075
NFF	0.136	0.215	0.04375	0.0875
FFN	0.136	0.215	0.05	0.1

3.2 Duration of Affect and Pre-test/Post-test

In the previous section, we saw evidence that brief frustration is associated with positive learning gains, but that lengthier frustration is associated with poor learning gains. In this section, we break down the learning gain measure into its constituent parts, the student's pre-test score and post-test score. Results shown in Tables 6-7 show that pretest scores can predict the frequencies of both confusion and frustration during the learning session. Specifically, lower pretest scores are more likely to co-occur with sequences containing at least one instance of that particular affect (as in CNN, NCN, and NNC when only the confusion detector is applied in Table 6 or in FNN, NFN, or NNF when only the frustration detector is applied in Table 7). Similar effects are found for sequences where two instances of either affect have been detected (as in CCN and NCC, or FFN and NFF). Further, higher pretest scores correlate with higher frequencies of prolonged states of not-confused and not-frustrated (both of which are represented as NNN in Tables 6-7). All the significant r-values in Tables 6-7 remain significant or marginally significant after the post-hoc control.

Table 6. Confusion vs. Pretest Scores (Significant results are in dark gray; marginally significant results are in light gray)

3-step - pre	r	p	p cutoff (sig)	p cutoff (marginal)
NCC	-0.295	0.006	0.00625	0.0125
CCN	-0.283	0.009	0.0125	0.025
NNC	-0.26	0.016	0.01875	0.0375
NNN	0.255	0.018	0.025	0.05
CNN	-0.226	0.037	0.03125	0.0625
NCN	-0.195	0.074	0.0375	0.075
CNC	-0.161	0.141	0.04375	0.0875
CCC	-0.005	0.967	0.05	0.1

Table 7. Frustration vs. Pretest Scores (Significant results are in dark gray; marginally significant results are in light gray)

3-step - pre	r	p	p cutoff (sig)	p cutoff (marginal)
NNN	0.277	0.010	0.00625	0.0125
NNF	-0.273	0.011	0.0125	0.025
FNN	-0.27	0.012	0.01875	0.0375
NFN	-0.267	0.014	0.025	0.05
NFF	-0.231	0.033	0.03125	0.0625
FFN	-0.231	0.033	0.0375	0.075
FNF	-0.125	0.253	0.04375	0.0875
FFF	-0.02	0.854	0.05	0.1

Surprisingly, correlating the affective sequences to post-test scores shows essentially no relationships. As Tables 8-9 show, neither confusion nor frustration sequences are significantly correlated with posttest results. In other words, low pre-test results predict confusion and frustration will occur during the learning session, but presence of these affective states does not predict post-test performance. These results suggest either that the tutor was effective at bringing all students up to mastery, or that there was a ceiling effect in test performance. In other words, students who were confused or frustrated during the learning session because they began with low domain knowledge caught up to students who, because they began with high domain knowledge, experienced little confusion or frustration. However, it is notable that as was found when compared to learning gains and to pre-test results, confusion and frustration have the same pattern for post-test results.

Table 8. Confusion vs. Posttest Scores (No results remain significant after post-hoc control)

3-step - post	r	p	p cutoff (sig)	p cutoff (marginal)
CCN	-0.155	0.157	0.00625	0.0125
NCC	-0.147	0.180	0.0125	0.025
NNN	0.068	0.539	0.01875	0.0375
CNN	-0.064	0.561	0.025	0.05
CCC	-0.061	0.579	0.03125	0.0625
CNC	0.052	0.635	0.0375	0.075
NNC	-0.04	0.716	0.04375	0.0875
NCN	-0.005	0.966	0.05	0.1

Table 9. Frustration vs. Posttest Scores (No results remain significant after post-hoc control)

3-step - post	r	p	p cutoff (sig)	p cutoff (marginal)
FFF	0.177	0.106	0.00625	0.0125
FNF	0.102	0.351	0.0125	0.025
NFF	-0.093	0.396	0.01875	0.0375
FFN	-0.093	0.396	0.025	0.05
NFN	0.025	0.822	0.03125	0.0625
NNF	-0.009	0.937	0.0375	0.075
FNN	-0.008	0.946	0.04375	0.0875
NNN	0	1.000	0.05	0.1

3.3 Unifying Confusion and Frustration

Confusion and frustration have some theoretical similarities, although they are often considered separately in affective research. Both are affective states that occur when a student is struggling with difficult material and has not yet achieved understanding. As discussed earlier, one way to interpret the work in [17] is that their model of confusion may also have included instances of frustration. Hence it may be worth studying these two constructs in a unified fashion – treating them as if they are the same construct during analysis. Also, as discussed in previous sections, the relationships between confusion and learning, and frustration and learning, were qualitatively similar in our data set. They were of different magnitudes (frustration had higher correlations than confusion) but were generally pointing in the same direction. This trend also warrants a joint analysis of the two states.

In order to do so, we applied both detectors (which operate independently) to the data at the same time. Any instance that was labeled as either confused (C) or frustrated (F) in previous sections was now labeled as "any" (A), including the rare instances where a single clip was labeled by the detectors as indicating both confusion and frustration. Instances of A are contrasted with instances where neither (N) affect was detected. Table 10 shows the correlations between learning gains and 3-step any/neither (A/N) sequences.

Table 10. Correlations between 3-step "Any" sequences and Learning Gains. (Significant results are in dark gray; the marginally significant, in light gray.)

3-step - diff	r	p	p cutoff (sig)	p cutoff (marginal)
NNA	0.295	0.006	0.00625	0.0125
NAN	0.284	0.008	0.0125	0.025
NNN	-0.279	0.010	0.01875	0.0375
ANN	0.262	0.015	0.025	0.05
ANA	0.213	0.050	0.03125	0.0625
NAA	0.204	0.061	0.0375	0.075
AAN	0.19	0.081	0.04375	0.0875
AAA	0.01	0.931	0.05	0.1

Several findings from this analysis are similar to the findings presented earlier in this paper, but obtain higher correlations than are seen for confusion or frustration alone. Extended periods of "neither" (i.e., NNN) during the learning session are negatively correlated with learning gains (r = -0.279). All 3-step sequences of short term "any" (i.e., NNA, NAN, and ANN) are found to be positively correlated with learning gains, (r=0.295, 0.284, and 0.262, respectively). Moreover, ANA, NAA, and AAN are found to be positively correlated at a marginally significant level (r=0.213, 0.204, and 0.19, respectively).

Compared with the significant r-values of 3-step frustration and learning gains in Table 5, the r-values for "any" have larger magnitudes, meaning that combining confusion and frustration yields stronger correlations with learning gains than frustration does alone.

4. CONCLUSION AND DISCUSSION

In this paper, we discussed correlations between student test scores and sequences of two affective states—confusion and frustration—during learning with Cognitive Tutor Algebra. These affective states were studied both independently and in combination.

A decade ago, key theoretical models of confusion and frustration during learning and interaction hypothesized that confusion leads to frustration [16] as part of a process where students fail to learn. In line with this theory, researchers suggested that identifying and responding to frustration was essential [13, 15]. However, research looking at overall proportions of student affect (e.g., confusion or frustration) found inconsistent patterns for confusion and null results for frustration (e.g., [7, 21], leading one paper to argue that frustration is significantly less important to learning than other affective states such as boredom [3]).

Research that followed this suggested that the dynamics of affect over time might play an important role in learning outcomes. Confusion that led to frustration, for example, was hypothesized to lead to poorer learning outcomes than confusion that resolved [10, 22].

In this paper, we find a pattern that accords broadly with [17], where confusion and frustration are associated positively with learning for brief episodes and negatively for lengthy episodes. Somewhat contrary to expectations (but consistent with the work in [17]), this effect is strongest if the two affective states are considered together, and weakest if confusion is considered alone

(with frustration in the middle). This finding is not inconsistent with the prior literature (differing relations between frustration and learning based on the length of frustration are quite consistent with overall null effects) but does reinterpret it somewhat.

One important limitation to the research presented here is that the length of the affective sequences differs from that found in [17], complicating comparisons between the two. It is known that different affective states often have different durations [8]. However, these durations are likely to be determined by the population and learning context as well. In other words, brief frustration in one context may be lengthy frustration in another. (This possibility may explain the similarity in results between this paper and [17]. Although the time per affective observation was different, the times used in each environment may have matched the general time for a student to make progress in the different environments, as computer programming is a more time-consuming activity than completing highly scaffolded mathematics problems.) Understanding what the "tipping point" is between brief and lengthy confusion or frustration, in different contexts, may be a valuable step for future research.

Overall, this paper's results suggest that attempting to understand overall relationships between affective states and learning is prone to conflating multiple phenomena. Affective states are not unitary; it matters at minimum how long they are, it matters what follows them [23], and probably other factors matter as well (such as culture, for instance). Researchers have also considered the possibility of multiple types of frustration (for instance, [12] speaks of "pleasurable frustration"). Our results show temporal effects for frustration that are highly similar to those hypothesized for confusion, results that deserve more careful consideration in future research. Though a student's overall degree of frustration has often been associated with null effects [e.g., 7, 21], it appears that frustration is associated with differences in learning when considered in a finer-grained fashion. It may be that the conditions that lead to both frustration and confusion (the struggle associated with learning material that is not immediately apparent) are necessary components of the learning process, and both frustration and confusion only become detrimental if a student is unable to reach resolution in an adequate time frame. It is also possible that frustration may be simply an outcome of the cognitive processes underlying these phenomenon, or even just a result of confusion being resolved or not resolved (e.g., different types or intensities or durations of confusion might trigger persistence or resistance, while varying lengths of frustration merely reflect these differences). The similar patterns between confusion and frustration raise questions about whether the best theoretical split is even between confusion and frustration, or whether we should instead move to comparing brief-confrustion, extended-confrustion, and perhaps pleasurable-confrustion (alternate terms for the affective state combining confusion and frustration are welcome). Work to understand and model these affective states in their full complexity will be an essential area of future research. These endeavors will be supported by the advent of data-mined models, such as the ones used here, that can identify affect in a fashion that is both fine-grained and scalable.

5. ACKNOWLEDGMENTS

This research was supported by grant "Toward a Decade of PSLC Research: Investigating Instructional, Social, and Learner Factors in Robust Learning through Data-Driven Analysis and Modeling," National Science Foundation award #SBE-0836012. We would like to thank the anonymous reviewers of [17] for very helpful comments and suggestions.

6. REFERENCES

[1] Arroyo, I., Cooper, D., Burleson, W., Woolf, B. 2010. Bayesian Networks and Linear Regression Models of Students' Goals, Moods, and Emotions. *Handbook of educational data mining* (Oct. 2010). Taylor and Francis Group, London, UK, 323.

[2] Baker, R.S.J.d., Corbett, A.T., Wagner, A.Z. 2006. Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring System* (Jhongli, Taiwan, June 26-30, 2006), 29-36.

[3] Baker, R.S.J.d., D'Mello, S., Rodrigo, M., Graesser, A. 2010. Better to be frustrated than bored: The incidence and persistence of affect during interactions with three different computer-based learning environments. *International Journal of Human-computer Studies*, 68, 4 (Dec. 2010). Elsevier B.V., Oxford, UK, 223-241.

[4] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. 2012. Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. *Proceedings of the 5th International Conference on Educational Data Mining* (Chania, Greece, June 19-21, 2012), 126-133.

[5] Baker, R.S.J.d., Moore, G., Wagner, A., Kalka, J., Karabinos, M., Ashe, C., Yaron, D. 2011. The Dynamics Between Student Affect and Behavior Occuring Outside of Educational Software. *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*. (Memphis, TN, Oct 9-16, 2011).

[6] Benjamini, Y.,Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Author(s). *Journal of the Royal Statistical Society*, Series B, 58, 1 (1995), London, UK, 289-300.

[7] Craig, S., Graesser, A., Sullins, J., Gholson, B. 2004. Affect and Learning: an Exploratory Look into the Role of Affect in Learning with AutoTutor. *Journal of Educational Media*, 29, 3 (Oct. 2004). Taylor & Francis, London, UK, 241-250.

[8] D'Mello, S.K., Graesser, A.C. 2011. The Half-Life of Cognitive-Affective States during Complex Learning. *Cognition and Emotion*, 25, 7 (2011). Taylor and Francis Group, London, UK, 1299-1308.

[9] D'Mello, S.K.., Lehman, B., Pekrun, R., Graesser, A.T. In Press. Confusion Can Be Beneficial For Learning. *Learning and Instruction*. Elsevier B.V., Oxford, UK.

[10] D'Mello, S.K., Person, N., Lehman, B.A. 2009. Antecedent-Consequent Relationships and Cyclical Patterns between Affective States and Problem Solving Outcomes. *Proceedings of 14th International Conference on Artificial Intelligence in Education* (Brighton, UK, July 6-10, 2009), 57-64.

[11] Forbes-Riley, K., Litman D. 2009. Adapting to Student Uncertainty Improves Tutoring Dialogues. *In Proceeding of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (Brighton, UK, July 6-10, 2009), 33-40.

[12] Gee, J.P. 2007. *Good video games+ good learning: Collected essays on video games, learning, and literacy*

(Mar. 2007). Peter Lang Pub Incorporated, Bern, Switzerland.

[13] Hone, K. 2006. Empathic Agents to Reduce User Frustration: The Effects of Varying Agent Characteristics. *Interacting with Computers*, 18, 2 (Mar. 2006). Elsevier B.V., Oxford, UK, 227-245.

[14] Koedinger, K.R., Corbett, A.T. 2006. Cognitive Tutors: Technology bringing learning science to the classroom. *The Cambridge Handbook of the Learning Sciences* (Apr. 2006). Cambridge University Press, Cambridge, UK, 61-78.

[15] Klein, J., Moon, Y., Picard, R. 2002. This computer responds to user frustration – Theory, design, and results. *Interacting with Computers*, 14, 2 (Feb. 2002). Elsevier B.V., Oxford, UK, 119-140.

[16] Kort, B., Reilly, R., Picard R. 2001. An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy—Building a Learning Companion. *Proceedings of the 1st IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges* (Madison, WI, Aug 06-08, 2001), 43-48.

[17] Lee, D.M., Rodrigo, M.M., Baker, R.S.J.d., Sugay, J., Coronel, A. 2011. Exploring the Relationship between Novice Programmer Confusion and Achievement. *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction* (Memphis, TN, Oct 9-12, 2011).

[18] McQuiggan, S.W., Robison, J.L., Lester, J.C. 2010. Affective Transitions in Narrative-centered Learning Environments. *Educational Technology & Society*, 13, 1(Jan. 2010). International Forum of Educational Technology & Society, 40-53.

[19] Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. 2012. Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

[20] Perneger, T.V. 1998. What's Wrong with Bonferroni Adjustments. *British Medical Journal*, 316 (Apr. 1998). BMJ Publishing Group, London, UK, 1236-1238.

[21] Rodrigo, M.M.T., Baker, R.S.J.d., Jadud, M.C., Amarra, A.C.M., Dy, T., Espejo-Lahoz, M.B.V., Lim, S.A.L., Pascua, S.A.M.S., Sugay, J.O., Tabanao, E.S. 2009. Affective and Behavioral Predictors of Novice Programmer Achievement. *Proceedings of the 14th ACM-SIGCSE Annual Conference on Innovation and Technology in Computer Science Education* (Paris, France, July 06-09, 2009), 156-160.

[22] Rodrigo, M.M.T., Baker, R.S.J.d., Nabos, J.Q. 2010. The Relationships between Sequences of Affective States and Learner Achievements. *Proceedings of the 18th International Conference on Computers in Education* (Putrajaya, Malaysia, Nov 29 - Dec 3, 2010).

[23] Sabourin, J., Rowe, J., Mott, B., Lester, J. 2011. When Off-Task is On-Task: the Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Artificial Intelligence in Education*, 21 (2011). Springer, Berlin/Heidelbery, Germany, 534-536.

[24] Sao Pedro, M. A., Baker, R.S.J.d., Montalvo, O., Nakama, A., Gobert, J.D. 2010. Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. *Proceedings of the 3rd International Conference on Educational Data Mining*, 181-190.