Engaging with the Scenario: Affect and Facial Patterns from a Scenario-Based Intelligent Tutoring System

Benjamin D. Nye[1(✉)], Shamya Karumbaiah[2], S. Tugba Tokel[3], Mark G. Core[1], Giota Stratou[4], Daniel Auerbach[1], and Kallirroi Georgila[1]

[1] Institute for Creative Technologies, University of Southern California, Los Angeles, USA
{nye,core,auerbach,kgeorgila}@ict.usc.edu
[2] Penn Center for Learning Analytics, University of Pennsylvania, Philadelphia, USA
shamya16@gmail.com
[3] Department of Computer Education and Instructional Technology, METU, Ankara, Turkey
stugba@metu.edu.tr
[4] Keysight Technologies, Atlanta, USA
giotastr@gmail.com

Abstract. Facial expression trackers output measures for facial action units (AUs), and are increasingly being used in learning technologies. In this paper, we compile patterns of AUs seen in related work as well as use factor analysis to search for categories implicit in our corpus. Although there was some overlap between the factors in our data and previous work, we also identified factors seen in the broader literature but not previously reported in the context of learning environments. In a correlational analysis, we found evidence for relationships between factors and self-reported traits such as academic effort, study habits, and interest in the subject. In addition, we saw differences in average levels of factors between a video watching activity, and a decision making activity. However, in this analysis, we were not able to isolate any facial expressions having a significant positive or negative relationship with either learning gain, or performance once question difficulty and related factors were also considered. Given the overall low levels of facial affect in the corpus, further research will explore different populations and learning tasks to test the possible hypothesis that learners may have been in a pattern of "Over-Flow" in which they were engaged with the system, but not deeply thinking about the content or their errors.

1 Introduction

Engagement, confusion, frustration, boredom, and related states have been demonstrated to impact learning gains on many traditional learning tasks, such as

C. Penstein Rosé et al. (Eds.): AIED 2018, LNAI 10947, pp. 352–366, 2018.
https://doi.org/10.1007/978-3-319-93843-1_26

math problems, reading text, and generating short-answers to questions [2,6,7]. There are many hardware/software systems available to detect learner emotions through physical signs. Facial expression trackers such as the Computer Expression Recognition Toolbox (CERT) [19] output measures for facial action units (AUs [10]), and based on the AU values, also output measures for general emotion categories (e.g., neutral, confusion, frustration). AUs are numeric codes representing the muscular movements that produce facial appearance changes. In this paper, we explore the utility of bottom-up information such as facial AUs which are fine-grained but not tied directly to cognitive-affective states (e.g., boredom). We use factor analysis to search for categories implicit in patterns of facial AUs. We also explore the use of top-down information such as the CERT emotion categories which are more coarse-grained and not designed for learning environments, where many affective states (e.g., fear, joy, disgust) are less relevant [6].

To explore the insights provided by learner facial cues during computer-based learning scenarios, we instrumented an intelligent tutoring system (ITS) for leadership training called the Emergent Leader Immersive Training Environment (ELITE)-Lite [16] to unobtrusively collect facial expressions from users and align this data to behavioral log files of system context and user behavior. This data allowed us to analyze the relationship of facial cues to other components of the experience (e.g., type of learning activity, correctness of responses) and to look for opportunities where it may be effective to leverage such cues to improve learning. The bottom-up factors enable a different perspective on the facial data, which is hypothesized to help identify patterns that might not be discovered by a top-down approach. In the following sections, we describe the theoretical background, data sample, and analysis of results, and discuss the findings.

2 Theoretical Background

A growing body of literature has studied emotions during computer-based learning and interaction, with affect measured using techniques such as self-report, human observation, text analysis, facial expression cues, speech audio analysis, physical sensors (e.g., pressure, conductance), and inferences from patterns of learner task behavior [2,9,15]. More recently, there has also been a shift toward multimodal affect detection such as through systematic analysis of combinations of tutor-student dialogue, facial affect, and task behavior [14]. A significant number of adaptive and non-adaptive learning tasks have been studied, which range from passive tasks (e.g., reading text and watching videos) to active tasks such as procedural problems (e.g., solving equations) or generative responses (e.g., deep reasoning questions, programming).

Within the space of learning environments that have been studied, some consensus has emerged about the utility of four key cognitive-affective states: engagement/flow, confusion/disequillibrium, frustration, and disengagement/boredom [6]. By comparison, traditional emotion categories such as disgust, fear and sadness have not been relevant to most learning tasks studied.

Among the four key emotions that occur during learning, engagement/flow has generally been shown to be positive and indicating greater attention and processing, and more recently evidence has also been found for confusion as a predictor of learning [7]. In the area of scenario-based learning, researchers studying the Crystal Island ITS found engagement to be associated with better learning outcomes [23]. In contrast, disengagement and boredom are generally understood to hinder learning [2]. The impact of frustration is not as well established - possibly because some students like extreme challenges while others prefer steady difficulty, as seen in video game players [17].

Overall, the majority of studies on cognitive and affective cues have relied on a combination of self-report and human raters. This introduces two limitations. First, human annotation may not provide results that are actionable in a real-time system. Second, since affect taxonomies are determined prior to observing the data, common facial cues or patterns that predict user outcomes might be missed because they did not fit into an expected category. This is particularly relevant because certain cognitive-affective cues are not necessarily social in nature and do not fit directly into traditional taxonomies of affect (e.g., looking "lost in thought"). Thus, human emotion annotation may not be able to be automated and might also miss important signals. A subset of literature has specifically studied how facial action units (AUs [10]) interact with behavior and learning with computers [14]. However, although AUs are fine-grained enough to avoid missing information, they need to be aggregated into metrics that can be interpreted.

To look for patterns in prior literature, we reviewed studies of AUs of learning technology users. Eleven such studies were identified based on data sets from users in seven different systems [1,3,8,12,26–28], which are summarized in Table 1. Each row indicates the AUs that were reported as associated with learning behaviors or outcomes in that study. AUs that were not reported in any of these studies are excluded from the table. In most cases, AUs were studied individually for effects on learning, though studies on AutoTutor and BrainSkills aligned AUs to engagement, confusion, frustration, boredom, and delight [5,8,21,27]. These alignments are noted with subscripts. Across these studies, AUs relevant to learning in at least three studies were associated with the eyebrows (e.g., raised eyebrows AU1 + AU2, wrinkled/furrowed brow A4), lip corners, dimples, and tightening (AU12, AU14, AU23), and eye lids (tightened lids AU7). Some of these cues also tended to co-occur across studies. For example, raised inner eyebrows (AU1) were reported with raised outer eyebrows (AU2) in five out of seven studies in which either AU1 or AU2 was highlighted.

As such, there is some evidence that certain combinations of AUs indicate events or cognitive states that influence learning. For example, Grafsgaard et al. [14] reported that their best model of learning gains includes facial AUs in addition to data from textual dialogue and behavioral logs. While systematic exploration of these factors will require study of many systems and scenarios over time, this framing underpins certain decisions made for the data collection and analysis presented next.

Table 1. Relevant Action Units (AUs) in different learning systems

Study	Platform	*N*	Action Units																			
			1	2	4	5	6	7	10	12	14	15	17	18	20	23	24	25	26	28	43	45
[13]	JavaTutor	65	X	X	X	-	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-
[14]	JavaTutor	63	-	X	X	-	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-
[12]	JavaTutor	65	-	-	X	-	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-
[26]	JavaTutor	67	-	-	X	X	-	-	-	-	-	X	-	-	-	X	-	-	-	-	-	-
[28]	Crystal Island	65	-	-	-	X	-	-	-	-	-	-	-	X	-	X	X	-	-	-	-	-
[3]	Physics Playground	137	X	X	-	-	X	-	X	X	X	X	X	X	X	X	-	X	-	X	-	-
[8]	AutoTutor	30	X_f	X_f	X_c	-	-	X_c	-	X_c	X_f	-	-	-	-	-	-	-	-	-	X_b	-
[21]	AutoTutor	28	-	-	X_c	-	-	$X_{c,d}$	-	$X_{f,d}$	-	-	-	-	-	-	-	-	X_d	X_d	-	-
[5]	AutoTutor	5	X_f	X_f	X_c	-	-	X_c	-	X_c	X_f	-	-	-	-	-	-	-	-	-	X_b	-
[27]	BrainSkills	34	X_e	-	-	-	-	-	X_e	-	-	-	-	-	-	-	-	-	-	-	-	X_e
[1]	TeachTown Basics	7	X	X	X	-	X	X	-	X	-	-	-	-	-	-	-	-	-	-	-	X
		Count	6	6	8	2	2	4	2	5	6	2	1	2	1	3	1	1	1	2	2	2

f = frustration, c = confusion, e = engagement, d = delight, b = boredom

3 Data Collection and Methodology

We summarize the data collection below; for more details see [4]. Data was collected on learners using the ELITE-Lite system, which was instrumented to collect a corpus of video logs via laptop web cameras. ELITE-Lite is a scenario-based ITS which uses multiple-choice-based role-playing interactions to allow learners to practice basic counseling skills, while a virtual coach proactively provides hints and feedback [16]. Each video log is a 30–60 min clip of participants interacting with ELITE-Lite.

The experiment was conducted at a competitive private university in California in Fall 2015. Data was collected across two randomly-assigned conditions that varied by the prevalence of guidance for partially-correct (mixed) answers, with one condition always giving hints/feedback for mixed answers and the other never giving hints/feedback for mixed answers. Correct answers never resulted in textual coach guidance (only graphical flag feedback), while incorrect answers always resulted in textual guidance. Users in the two conditions had only very subtle differences in their experiences, making both conditions relatively equivalent from the standpoint of this analysis.

A total of 80 students participated in the study, but only 39 had complete and usable data for affect analysis. Recordings were collected in an open lab with potential distractions, which in some cases added noise to the audio/video log. All students wore headphones for sound, and some minor confounds are that some wore glasses or had facial hair. As noted in prior log file analysis [4], 6 participants were omitted because their game logs were incomplete. A large number of videos were omitted from this analysis because the video/audio recording was corrupted, the participant's face was consistently obscured or cropped, or automated analysis identified problems (e.g., poor results due to issues with lighting or missing frames). Of the 39 participants with complete data (29 male and 10 female), 33 participants

identified themselves as Asian/Pacific Islander, 3 as White, 2 as Black/African American and 1 preferred not to respond. Most of the participants came from technical majors such as computer science and thus expressed high comfort with computers. No significant differences were observed in learning gains between participants omitted versus those included.

The experimental procedure consisted of (1) one pre-survey, (2) one pre-test, (3) one ELITE-Lite Introductory Video, (4) two sessions of the same scenario (Being Heard) with a virtual coach providing hints and feedback, (5) one session with a second scenario (Bearing Down) without coach support, (6) one post-survey, and (7) one post-test. Excluding surveys and tests, participants spent close to an hour in ELITE-Lite. In each scenario, learners play the role of a military supervisor helping a subordinate with a problem. Each scenario is relevant to a broad audience: Being Heard addresses a request to transfer due to sexual harassment and Bearing Down addresses a fight between two subordinates. The first scenario was repeated to allow the participants to identify and correct their mistakes.

Learning gain was measured using two types of tests - a shallow knowledge test (e.g., definitions of skills like *active listening*) and a deeper Situational Judgment Test (SJT) which required rating possible actions. Both tests may be considered transfer tasks, in that they test skills under substantially different conditions than the training scenarios. The pre-survey collected self-reported traits: help seeking, growth mindset, interest, lack of anxiety, organization, confidence, and experience. In addition to the web camera recordings, game logs recorded participant responses and interactions with the system. Based on the game logs, each response event was coded by Question Difficulty, whether the question had been seen before (Repeat), Correctness, Time Taken to Answer, if Hint and/or Feedback was presented, the Phase (e.g., session), and a Unique Question ID. Game logs were also aligned to video annotation, so that the system phase was known for each video frame.

Video analysis was conducted using a commercial version of the Computer Expression Recognition Toolbox (CERT) [19], which performs real-time facial expression recognition. CERT reports emotion estimates as well as the activation of 20 action units (AUs), following the well-validated Facial Action Coding System (FACS). CERT estimates are trained primarily on posed facial expressions aligned to Ekman's taxonomy (Baseline, Joy, Anger, Surprise, Fear, Contempt, Sadness, Disgust) [11,19,20]. The commercial version extends this to estimate Confusion and Frustration. This tool processed the video logs outputting evidence levels for each AU and emotion measure for each frame.

Human annotators labeled a sample of the data to assist in interpretation of the data. CERT output was compared with the annotations from two humans for a sample of 6 videos, examining block sizes of 3 s. This interval length has been used by other research [26], and was appropriate for this system as events (e.g., decision points, feedback) did not typically occur longer than this rate. A sweep examining other block sizes (1 s to 5 s) was briefly explored, but 3 s remained the most interpretable and consistent between raters. One concern was whether

participants looked off-screen, a sign of boredom, but this turned out to be rare, and so this was not analyzed further.

The human annotators found large amounts of Baseline facial expressions that appeared to range from deliberate engagement to mild boredom, and a true lack of pronounced facial expression. Some Confusion was seen, but the other emotion categories including Frustration were so infrequent that we created an Other category from the maximum of the CERT evidence levels across all other remaining emotions. These "top-down" emotion categories were analyzed based on their evidence levels for the overall experiment (Overall), for specific learning tasks (Phases), and for the 3 s windows before and after each participant decision since decision points are highly likely to be relevant to learning.

In addition, for these 3 s windows, we also considered the AU evidence levels for each learner. In general, evidence levels in CERT can be positive, suggesting the feature is present, negative, suggesting the feature is absent, or zero, indicating uncertainty. AU levels below zero were treated as zero, so that we only account for variation during activation (similar to [25]). A factor analysis was applied to identify linear combinations of AUs that co-occurred in patterns in our data sample. These bottom-up factors were then calculated and analyzed similarly to the top-down emotion categories, to identify new insights with respect to learning events.

4 Results and Analysis

For overall learning gains, the impact of hints, and student traits as well as a preliminary exploration of coarse-grained affect, see [4,22]. The current analysis concerns these questions: (1) What were the distributions of top-down affect detected overall and during different phases of using the system?, (2) What bottom-up patterns of facial cues occur and how do these relate to phases of system use?, (3) How does student affect relate to responses (e.g., correctness, before/after submitting an answer) and to learning gains?, and (4) What self-reported student traits were associated with differences in affect detected?

4.1 Affect Results (Top-Down Categories)

Table 2 summarizes the descriptive statistics of the three top-down emotion categories reported by CERT for the whole session (Overall), for system phases, and for an average over each user's windows of ±3 s around submitting a decision (N = 39 users). The system phases reported are the introduction video (Intro), then one scenario twice in a row (Being Heard: Scen 1 and Scen 2) before continuing to a second scenario (Bearing Down: Scen 3). For each affective state, the table shows the average evidence level across all recorded frames. Across subjects, there was substantial variability, as seen in the standard deviations. As would be expected, Pearson's correlations showed significant pairwise correlations between each emotion overall ($p < .01$ for all). Baseline was negatively

Table 2. CERT evidence level means and standard deviations for overall, system phases, and around decisions

Condition	Baseline	Confusion	Other
Overall	0.92 ± 0.52	−0.72 ± 0.68	0.25 ± 0.44
Intro	0.83 ± 0.70	−0.81 ± 0.87	0.36 ± 0.40
Scen 1	0.84 ± 0.65	−0.50 ± 0.87	0.27 ± 0.49
Scen 2	0.58 ± 0.64	−0.47 ± 0.93	0.46 ± 0.40
Scen 3	0.73 ± 0.57	−0.50 ± 0.99	0.48 ± 0.52
Decisions	0.64 ± 0.57	−0.41 ± 0.70	0.20 ± 0.44

correlated with Confusion (r = −.35) and Other (r = −.88), with Confusion positively correlated with Other (r = .36).

Since each measure varies over time, different expressions dominate at different times. If it were assumed that only one affective state could be active at a time (e.g., discretized to the one with maximum evidence), then the prevalence of each would be 69% Baseline, 5% Confusion, and 26% Other. Overall, there are relatively low evidence levels of Confusion and Other emotions. Among emotions included in Other, the average values were all less than zero (−.38 to −2.51; −.96 for Frustration). The highest estimated emotion in Other was Contempt (−.42 ± .43), which was still quite uncommon. However, Contempt did show a positive but non-significant increase between Intro and Scen 3 (−.43 to −.34; p = .08), which might indicate it could be a useful indicator of decreasing engagement or study fatigue over the course of about an hour. Human annotation indicated similar patterns over a set of 2316 tags for 3 s intervals, with a relatively neutral but engaged expression (Baseline) dominating the experience (92.9%), while clear signs of Confusion (3.5%) or strong evidence of Other affect (6.8%) were rare.

Different system phases of learning activities (Phase) influenced these emotion estimates. Analyses showed a significant main effect for Phase, $F(9,324) = 2.78$, $p < .00$. Baseline and Other emotions were not statistically significant for different Phases overall ($p > .06$ and $p > .41$, respectively). However, a statistically significant effect was found for Confusion $F(3,108) = 3.37$, $p < .02$, partial eta-squared = .09. Furthermore, within-subjects analysis showed that there was a significant linear trend for Confusion over Phases, $F(1,36) = 5.67$, $p < .02$, partial eta-squared = .14. Overall, confusion was very low during the introductory video and rose for the first two sessions to a steady-state for the final session.

Compared to Overall affect, affect around decisions showed lower levels of Baseline and higher Confusion, with Other remaining similar, as shown in Table 2. Discretizing to only consider the maximal state, the average prevalence over 3 s before a decision to 3 s after was 69% Baseline, 7% Confusion, and 24% Other. As such, Baseline was still dominant.

4.2 AU Factor Results (Bottom-Up)

A factor analysis was applied to the non-negative AUs for all 3 s periods before and after user decisions, for each frame recorded. These periods were chosen because they were anticipated to have the highest volatility of affective reactions, since they covered both the decision-making process, the delivery of hints/feedback after a decision, and the beginning of the virtual agent's response to the learner. Direct Oblimin (Oblique) rotation and structure matrix coefficients were used, since it is reasonable that the same AU could appear in multiple factors. A loading cutoff of .5 was chosen based on reviewing the scree plot, which resulted in seven distinct factors that explained 69.5% of the variance. This cutoff resulted in the factors presented in Table 3. To facilitate interpretation, these factors are mapped to related reference citations that presented similar combinations of factors and a summary label is given to each factor next to it.

A few factors showed significant and non-trivial correlations (N = 1954 decision events). Factor 1 correlated moderately with Factor 5 ($r = .45$, $p < .05$) and weakly with Factor 4 ($r = .16$, $p < .05$). Factor 3 correlated with Factors 5, 6, and 7 ($r = .16$, $r = .14$, and $r = .14$; $p < .05$ for all). Factor 4 correlated negatively with Factor 6 ($r = -.11$; $p < .05$).

Table 3. Factor AU Loadings and related work

Factor: Summary Label	AUs (loadings)	Related Refs: AUs
Factor 1: Mouth Tightened	14, 17, 23, 24, 28 (0.86, 0.62, 0.89, 0.90, 0.62)	[25]: 14, 17, 23
Factor 2: Surprise/Mouth Covered	5, 25, 26 (0.77, 0.92, 0.95)	[11]: 1, 2, 5, 25, 26 [25]: 20, 25, 26
Factor 3: Eyebrows Raised	1, 2 (0.93, 0.92)	[25]: 1, 2
Factor 4: Happy/Smile	6, 12 (0.78, 0.78)	[11]: 6, 7 [1]: 12
Factor 5: Frown/Pursed lips	9, 10, 15, 17 (0.58, 0.77, 0.54, 0.50)	[25]: 10, 15, 17
Factor 6: Thinking/Uncertain	4, 18, 43 (0.74, 0.81, 0.56)	[8], [21], [5]: 4, 7, 12
Factor 7: Lips Stretched	20 (0.54)	[18]

It is important to note that these factors represent relatively subtle differences in facial expression. Figure 1 shows examples of Factor 1, 3, 5, and 6. Three images are of a single user to demonstrate that while differences within a participant can be identified through careful inspection, they are not immediately obvious. The final image for Factor 1 is included because it indicates a second issue: users tend to move their hands as they think and use computers in a natural manner. In general, this appeared to be either consistent (e.g., their face was so consistently covered that their data was excluded from processing)

or random (e.g., unrelated to detected cues). However, there is some possibility that Factor 2 represents both mouth opening or a hand placed near the mouth area, since frames with high levels of this factor appeared more likely to have a hand near the chin or mouth. All other factors appear to be related exclusively to facial AUs.

These factors map reasonably to prior literature on facial affect patterns, though not all factors map to traditional emotion labels. For example, Factor 2 and Factor 4 share action units with Eckman's Surprise and Happy/Smile categories, respectively [10]. Factor 6 has some similarity with D'Mello and Graesser's confusion [6], in that the shared AU4 represents a furrowed brow. However, Factor 1 (Mouth Tightening), Factor 3 (Eyebrows Raised), and Factor 5 (Frown) appear to be facial cues related to cognition that do not necessarily map neatly to a traditional emotion category. These factors are instead similar to bottom-up factor patterns that have been found when studying users in other cognitively-demanding tasks such as negotiating with a computer agent [25]. Finally, AU20 has been associated with embarrassment [18] but Factor 7 might not represent a significant pattern. It contains only a single action unit and the scree plot indicates that Factor 7 is the first point of the relatively flat tail. As a group, the factors found in the data may offer indicators related to engagement, which could not be measured directly in the earlier analyses using the top-down emotion categories (i.e., Baseline was used as a noisy proxy instead).

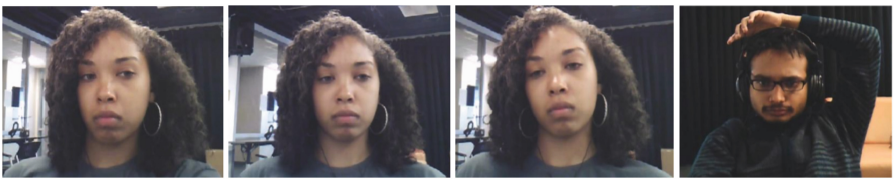

Fig. 1. Example factors (from left to right) - Factor 3, Factor 6, Factor 5 and Factor 1

To explore individual differences in the prevalence of factors, an analysis was conducted that counted the number of decision points where each learner was at least one standard deviation above the mean value for that factor across all participants. This approach found that strong presentations of factors were reasonably spread across participants, as shown in Fig. 2. Factor 4 was a notable exception, with nearly 50% of strong presentations by one subject. The most pronounced pattern however was that the facial affect detection showed much higher evidence of emotions for some subjects versus others. This pattern was also observed for the top-down emotions. This may indicate that certain learners were more reactive, or that the affect detection software shows systematic biases toward certain kinds of faces.

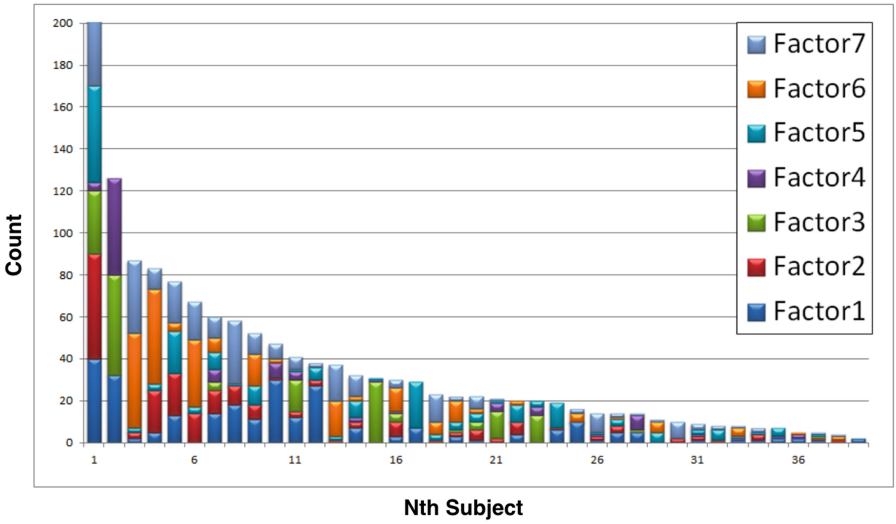

Fig. 2. Counts of factors 1σ above mean, by participant

To consider if the factors differed qualitatively from the top-down categories, we reviewed the most-correlated top-down emotion for each bottom-up factor. No factors are direct analogs to CERT categories, but some are at least moderately correlated while others appear to capture different dimensions of cognitive-affective states. Correlations were calculated based on the set of 3 s-window averages before decisions ($N = 1655$ complete cases; $p < .001$ for all reported). Factors which might be considered fairly similar are Factor 6 with Confusion ($r = .46$) and Factor 4 with Joy ($r = .42$). Factor 1 correlates fairly evenly with a number of emotions (Frustration $r = .35$; Contempt $r = .30$; negatively with Baseline $r = -.32$). Factor 3 has small correlations with a mixed group of more passive emotions (Sadness $r = .41$; Surprise $r = .29$; Fear $r = .28$). Factor 5 has only small correlations with top-down categories, but may relate to an active negative reaction such as a response to negative feedback (Disgust $r = .29$; Sadness $r = .26$). Factor 7 shows no correlations above .2 magnitude with any top-down emotion. As such, the bottom-up factors offer a different interpretation on the AUs than the top-down categories.

Table 4 shows the means for these factors overall and for each experimental phase. Each factor was normalized to fall in a range of $[0, 1]$ by dividing by its maximum possible weight. Two values for Factor 2 were so small ($< .001$) that they were rounded to zero for presentation. Due to space constraints, standard deviations are not shown but they were roughly on the same order of magnitude as each factor mean.

Among factors that showed changes between phases, within-subjects paired two-tailed t-tests were applied to test significance of differences between each successive phase. Supporting some of the intuitions from the correlation data,

Factor 1 was highest around active decision-making and lowest during passive videos, while Factor 3 showed the reverse relationship. Between the Intro and Scen 1, Factor 1 increased ($p < .05$), Factor 2 decreased to nearly zero ($p < .01$) and Factor 4 also decreased ($p < .01$). Between Scen 1 and Scen 2, Factor 4 increased ($p < .001$). Between Scen 2 and Scen 3 Factor 3 increased ($p < .01$). When comparing Decision points against the Overall average, Factor 1 was higher around decisions ($p < .001$) as was Factor 5 ($p < .02$), and Factor 7 was lower ($p < .05$).

Table 4. Factor overall and phase means of subjects (N = 39)

Phase	F1	F2	F3	F4	F5	F6	F7
Overall	0.11	0.003	0.15	0.03	0.12	0.27	0.01
Intro	0.08	0.004	0.18	0.04	0.09	0.34	0.01
Scen 1	0.11	˜0	0.14	0.01	0.08	0.37	0.004
Scen 2	0.12	0.01	0.14	0.03	0.08	0.34	0.002
Scen 3	0.10	˜0	0.18	0.02	0.08	0.35	0.01
Decisions	0.14	0.003	0.13	0.03	0.13	0.27	0.003

4.3 Relationship of Learner Outcomes to Facial Cues

To look at the relationship between detected affect prior to a response and its correctness (i.e., correct, rather than mixed or incorrect), two mixed models were evaluated using the *afex* R package [24]. These models were selected by including the affective cues studied in this work, then adding a limited set of factors known to affect correctness from prior exploratory analysis [4]. First, a mixed model was built that examined the predictive value of top-down emotions, in the form: *IsCorrect ˜ Confusion + Other + Repeat + TimeTaken + QuestionDifficulty + Hint + (1 + Confusion + Other — Subject.Id)* where Repeat refers to whether the question had been seen before, and Hint refers to whether a hint was given. Second, a similarly-structured mixed model was evaluated for factors (i.e., each factor having a fixed effect and a random effect conditional on the subject, to address potential systematic differences in the prevalence of factors between subjects). In both cases, statistically significant models were produced (emotions marginal $R^2 = .16$ and conditional $R^2 = .18$; factors marginal $R^2 = .16$ and conditional $R^2 = .23$). However, for both models, only QuestionDifficulty, Repeat, and TimeTaken were significant at $p < .05$ with QuestionDifficulty explaining the majority of the variance.

The next analysis attempted to estimate learning gains between the pre- and post-tests. This analysis was complicated by the fact that overall learning gains were relatively modest. Across the sample of all participants with log files (N = 74), the learning gain for participants was 0.08 (from 0.57 to 0.65). While this represents a fairly small variance, there was higher gain for SJT items (0.11 out of 1) than for Knowledge (0.04 out of 1) [4]. The subset of 39 with full data for

this analysis was not significantly different in terms of gains (0.10 for SJT, 0.05 for Knowledge). A model was fit for each type of facial measure (e.g., the set of top-down or bottom-up measures) averaged across all decision events. For both sets of metrics these models did not reach significance.

4.4 Relationship of Self-reported Traits to Facial Indicators

As noted earlier, students were pre-surveyed for traits related to experience/interest in the domain, learning strategies, growth mindset, and a limited subset of academic emotions. Under Pearson's correlations with Bonferroni adjustment for repeated tests, few self-reported traits showed statistically significant results for this sample size, with only Experience and Anxiety (e.g., test anxiety) notable. Experience was positively correlated with Baseline prior to decisions ($r = .40$, $p < .01$) and negatively with Other ($r = -.40$, $p < .01$), which captured strong affect other than Confusion. Lack of anxiety was negatively correlated with Confusion ($r = -.34$, $p < .05$), with students who reported more academic anxiety also showing more confusion evidence. Post-surveys, which focused primarily on impressions of the system, showed no significant correlations with affect for this sample size.

Bottom-up factors correlated with indicators of interest and effort, unlike the correlations for the top-down emotion categories which correlated with confidence and anxiety. However, the majority of these correlations were still small to moderate (.2–.4). Factor 2 showed moderate negative correlations with self-reported academic effort ($r = -.39$; $p < .05$) and with organized study habits ($r = -.40$; $p < .05$). Two factors were near significance for positive correlations with effort as well, Factor 3 ($r = .29$; $p = .08$) and Factor 5 ($r = .28$; $p = .08$). Factor 4 correlated negatively with interest in the subject ($r = -.43$; $p < .05$). Factors 2 and 4 may indicate taking the learning experience less seriously, while Factors 3 and 5 might be related to greater engagement or deliberate practice.

5 Conclusions and Future Directions

Facial expression trackers such as the Computer Expression Recognition Toolbox (CERT) [19] output measures for facial action units (AUs [10]) as well as emotion categories from traditional taxonomies. Using factor analysis to search for categories implicit in AU patterns, we explored the utility of this bottom-up information which is fine-grained but not tied directly to cognitive-affective states (e.g., boredom).

Bottom-up factors show the potential to add value beyond the traditional emotion taxonomies. In our review of related work (Sect. 2), we compiled a unified map of AU patterns studied by learning science researchers. Comparing this map to the factors we found in our data (Table 3), we see that only one (Factor 6) has been previously seen in a learning context while the other factors have been observed in other cognitive software tasks [25] and mentioned in Ekman and Friesen's book on facial expressions [11]. These tables provide a template

for building an ongoing record as new studies report results, and meta-analyses compile them.

Results suggest that the factors we identified may be relevant to engagement. Correlations with self-reported traits show negative relationships between Factor 2 and academic effort and organized study habits as well as a negative relationship between Factor 4 and interest in the subject. It could be the case that these traits correspond with a flow state in which Factor 2 and Factor 4 are less active. We also compared average evidence for each factor across the four phases of the experiment as well as the 3 s before and after decisions (Table 4). Factor 1 was highest in these decision making windows and lowest during the video watching phase while Factor 3 was the reverse. Learners might have been more likely to be in the flow state during the video, but humor and graphics in the video may have also triggered emotional reactions. However, these factors must be interpreted with care, since they are based on a limited set of tasks and a particular subject population: factor analysis from other systems is necessary to find common patterns.

Engagement/flow has generally been shown to be positive (e.g., [23]) and indicating greater attention and processing. However, this research did not isolate any facial expressions having a significant positive or negative relationship with either learning gain, or performance once question difficulty was also considered. These results indicate that the detected facial cues give insight into the learner's mental state, but as potential cues to predict learning did not offer a consistent signal.

It is important to note that learners on average showed limited signs of facial affect, either for top-down emotion categories (e.g., Confusion) or the bottom-up factors. For example, the high level of baseline and lack of confusion (69% and 5% when discretized) differs substantially from some previous research such as D'Mello and Graesser [6], which reported only 42% in Flow + Neutral and 19% in Confusion. The scenario-based learning context may have damped academic emotions such as confusion due to sense of flow in the scenario. Considering the low levels of confusion, frustration, and no signs of overt disengagement, all supporting evidence indicates that learners were in an engaged/equilibrium state as per D'Mello and Graesser's model [6]. This state might indicate a pattern of "Over-Flow" where learners are engaged in the experience and content, but float past their failures and potential impasses (e.g., insufficient confusion). Alternatively, this sample of learners (who were primarily computer science students) might have relatively low presentation of affect. A broader sample might include more variance in facial affect and offer more insight into performance.

To explore these issues, research is underway to examine these contributing factors. First, we have changed recruiting methods to attract a more diverse pool of learners. Second, a new activity (an interactive after-action review) has been added to promote impasses by prompting learners to diagnose or retry decisions where they made a mistake. By studying the same system with a new task and different subject population, it may be possible to disentangle if lower facial affect was due to insufficient impasses during scenario flow (i.e., over-flow).

This should contribute to the broader discussion on how to balance scenario flow against creating impasses that promote learning.

Acknowledgments. The effort described here is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Ahmed, A.A., Goodwin, M.S.: Automated detection of facial expressions during computer-assisted instruction in individuals on the autism spectrum. In: CHI Conference on Human Factors in Computing Systems (2017)
2. Baker, R.S., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: 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. Int. J. Hum. Comput. Stud. **68**(4), 223–241 (2010)
3. Bosch, N., D'Mello, S.K., Ocumpaugh, J., Baker, R.S., Shute, V.: Using video to automatically detect learner affect in computer-enabled classrooms. ACM Trans. Interac. Intell. Syst. (TiiS) **6**, 17 (2016)
4. Core, M.G., Georgila, K., Nye, B.D., Auerbach, D., Liu, Z.F., DiNinni, R.: Learning, adaptive support, student traits, and engagement in scenario-based learning. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (2016)
5. Craig, S.D., D'Mello, S., Witherspoon, A., Graesser, A.: Emote aloud during learning with AutoTutor: applying the facial action coding system to cognitive-affective states during learning. Cogn. Emot. **22**(5), 777–788 (2008)
6. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learn. Instr. **22**(2), 145–157 (2012)
7. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. Learn. Instr. **29**, 153–170 (2014)
8. D'Mello, S.K., Craig, S.D., Gholson, B., Franklin, S., Picard, R., Graesser, A.C.: Integrating affect sensors in an intelligent tutoring system. In: Affective Interactions: The Computer in the Affective Loop Workshop, pp. 7–13 (2005)
9. D'Mello, S.K., Craig, S.D., Sullins, J., Graesser, A.C.: Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. Int. J. Artif. Intell. Educ. **16**(1), 3–28 (2006)
10. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Stanford University, Palo Alto (1977)
11. Ekman, P., Friesen, W.V.: Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues. Malor Books, Cambridge (2003)
12. Grafsgaard, J., Wiggins, J., Boyer, K.E., Wiebe, E., Lester, J.: Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In: Educational Data Mining (2014)
13. Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.: Automatically recognizing facial expression: predicting engagement and frustration. In: Educational Data Mining (2013)

14. Grafsgaard, J.F., Wiggins, J.B., Vail, A.K., Boyer, K.E., Wiebe, E.N., Lester, J.C.: The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In: International Conference on Multimodal Interaction (ICMI), pp. 42–49. ACM (2014)
15. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: an empirical study of MOOC videos. In: Learning at Scale Conference, pp. 41–50. ACM (2014)
16. Hays, M.J., Campbell, J.C., Trimmer, M.A., Poore, J.C., Webb, A.K., King, T.K.: Can role-play with virtual humans teach interpersonal skills? In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (2012)
17. Juul, J.: Fear of failing? The many meanings of difficulty in video games. Video Game Theor. Read. **2**, 237–252 (2009)
18. Keltner, D.: Signs of appeasement: evidence for the distinct displays of embarrassment, amusement, and shame. J. Pers. Soc. Psychol. **68**(3), 441–454 (1995)
19. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (CERT). In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 298–305 (2011)
20. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
21. McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., Graesser, A.: Facial features for affective state detection in learning environments. In: Annual Meeting of the Cognitive Science Society, pp. 467–472 (2007)
22. Nye, B., Karumbaiah, S., Tokel, S.T., Core, M.G., Stratou, G., Auerbach, D., Georgila, K.: Analyzing learner affect in a scenario-based intelligent tutoring system. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS, vol. 10331, pp. 544–547. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_60
23. Rowe, J.P., Shores, L.R., Mott, B.W., Lester, J.C.: Integrating learning, problem solving, and engagement in narrative-centered learning environments. Int. J. Artif. Intell. Educ. **21**(1–2), 115–133 (2011)
24. Singmann, H., Bolker, B., Westfall, J., Aust, F.: afex: Analysis of Factorial Experiments (2018). https://CRAN.R-project.org/package=afex. r package version 0.19-1
25. Stratou, G., Morency, L.P.: Multisense - context-aware nonverbal behavior analysis framework: a psychological distress use case. IEEE Trans. Affect. Comput. **8**(2), 190–203 (2017)
26. Vail, A.K., Grafsgaard, J.F., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Predicting learning from student affective response to tutor questions. In: International Conference on Intelligent Tutoring Systems, pp. 154–164 (2016)
27. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans. Affec. Comput. **5**(1), 86–98 (2014)
28. Xu, Z., Woodruff, E.: Person-centered approach to explore learner's emotionality in learning within a 3D narrative game. In: Learning Analytics & Knowledge Conference, pp. 439–443 (2017)