ARTICLE IN PRESS

Learning and Instruction xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system?

Michelle Taub[a,*], Roger Azevedo[a], Ramkumar Rajendran[b], Elizabeth B. Cloude[a], Gautam Biswas[c], Megan J. Price[a]

[a] Department of Learning Sciences and Educational Research, University of Central Florida, USA
[b] Interdisciplinary Program in Educational Technology, Indian Institute of Technology Bombay, Mumbai, India
[c] Department of Electrical Engineering and Computer Science, Vanderbilt University, USA

ARTICLE INFO

*An earlier version of this study was presented at the 14th International Conference on Intelligent Tutoring Systems (ITS 2018) in Montreal, Quebec, Canada and published in R. Nkambou, R. Azevedo, J. Vassileva (Eds.), *Proceedings of the* 14th *International Conference on Intelligent Tutoring Systems (ITS 2018)* (pp. 233–242). Amsterdam, The Netherlands: Springer.

Keywords:
Cognitive strategies
Emotions
Intelligent tutoring systems
Metacognitive processes
Self-regulated learning

ABSTRACT

The goal of this study was to investigate 65 students' evidence scores of emotions while they engaged in cognitive and metacognitive self-regulated learning processes as they learned about the circulatory system with MetaTutor, a hypermedia-based intelligent tutoring system. We coded for the accuracy of detecting students' cognitive and metacognitive processes, and examined how the computed scores related to mean evidence scores of emotions and overall learning. Results indicated that mean evidence score of surprise negatively predicted the accuracy of making a metacognitive judgment, and mean evidence score of frustration positively predicted the accuracy of taking notes, a cognitive learning strategy. These results have implications for understanding the beneficial role of negative emotions during learning with advanced learning technologies. Future directions include providing students with feedback about the benefits of both positive and negative emotions during learning and how to regulate specific emotions to ensure the most effective learning experience with advanced learning technologies.

1. Introduction

Self-regulated learning (SRL) is a complex educational construct that describes students as active learners (Winne, 2018). SRL has many dimensions, including cognitive, affective, and metacognitive processes. Cognitive strategies include taking notes, summarizing, and making inferences used during thinking, comprehending, and problem-solving, while metacognitive processes involve making judgments of learning and feelings of knowing while monitoring the products of strategy-use (Winne, 2018). Affective processes involve students' emotions and how they regulate those emotions to enhance learning. Research has shown positive emotions can have a more positive influence on learning compared to negative emotions (D'Mello, Kappas, & Gratch, 2017; Pekrun, 2006). However, evidence demonstrates students have difficulty deploying effective SRL processes (Azevedo, Taub, & Mudrick, 2018). To address this, advanced learning technologies (ALTs) have been designed to teach SRL strategy use as students learn about various domains, e.g., science (Azevedo et al., 2018; Biswas, Segedy, & Buchongchit, 2016). Research on SRL with ALTs (e.g., intelligent tutoring systems, hypermedia, open-ended learning environments) predominantly focuses on one single process (e.g., metacognitive monitoring or emotions), but not both, which limits our understanding of the interplay between these processes, which have all been found to play an important role during learning (Azevedo et al., 2018). Additionally, studies often do not focus on micro-level SRL processes. For example, studies typically investigate macro-level processes, such as planning or monitoring (e.g., Bannert, Sonnenberg, Mengelkamp, & Pieger, 2015), as opposed to micro-level processes, (e.g., content evaluation; Greene & Azevedo, 2009), which are specific types of planning, monitoring, or strategies. See Greene and Azevedo (2009) for a detailed description of macro- and micro-level SRL processes. Studies investigating emotions often focus on positive and negative emotions that

* Corresponding author. University of Central Florida, Department of Learning Sciences and Educational Research, College of Community Innovation and Education, PO Box 161250, Orlando, FL, 32816-1250, USA.

E-mail addresses: michelle.taub@ucf.edu (M. Taub), roger.azevedo@ucf.edu (R. Azevedo), ramkumar.rajendran@iitb.ac.in (R. Rajendran), elizabeth.cloude@knights.ucf.edu (E.B. Cloude), gautam.biswas@vanderbilt.edu (G. Biswas), megan.price@knights.ucf.edu (M.J. Price).

https://doi.org/10.1016/j.learninstruc.2019.04.001
Received 1 July 2018; Received in revised form 23 April 2019; Accepted 26 April 2019
0959-4752/

cluster emotions together (e.g., Jarrell, Harley, Lajoie, & Naismith, 2017), as opposed to studying specific emotions and their impact on learning. The goal of this study was to examine participants' emotions (joy, anger, surprise, contempt, confusion, and frustration) while they engaged in cognitive (taking notes, making summaries) and metacognitive (judgment of learning, feeling of knowing) processes during learning with a hypermedia-based intelligent tutoring system (ITS).

1.1. Theoretical frameworks

We used two theoretical frameworks: the information processing theory of SRL because it views SRL as temporally unfolding events, but does not include emotions, and the model of affective dynamics because it describes emotions experienced during learning with ALTs.

Winne's (2018) Information Processing Theory explains how learners process information as events that unfold over time, and accounts for factors such as the contextual demands of the task. They outline four phases of learning students go through as they complete a task: 1) task definition, 2) goals and planning, 3) enactments (e.g., strategies students use) and 4) adaptation (e.g., adapting methods previously chosen to achieve their goal). During each phase, students monitor multiple features: conditions, operations, products, evaluations, and standards (COPES). Conditions are resources available to students, such as prior domain knowledge. Operations involve cognitive strategies and metacognitive processes used to complete the task. Products comprise new information created as students monitor, search, and assemble information. Standards describe the quality of the products and allow students to conduct evaluations, which help them assess whether the products are a good fit based on the standards and products.

D'Mello & Graesser's Model of Affective Dynamics (2012) seeks to disentangle the complex nature of students' emotions during learning using cognitive-affective states. The model emphasizes the role of cognitive disequilibrium, which occurs when students encounter an obstacle to their learning goal (e.g., contradictions, interruptions). This cognitive disequilibrium plays a vital role in students' deep learning when students are able to return to a state of equilibrium. A student's affective state of confusion is a key indicator of this disequilibrium, and if left unresolved can transition to frustration and then boredom, leading to disengagement from learning. However, if the student is able to resolve their confusion, they return to a state of cognitive equilibrium and continue to stay engaged in their learning goals. Both of these models are appropriate for our study because we examined emotions during SRL.

1.2. Literature review

Many studies have investigated the use of cognitive and metacognitive processes with ALTs (Bannert et al., 2015; Kinnebrew, Segedy, & Biswas, 2017; Taub & Azevedo, 2019). One study found that students in an experimental condition who developed self-directed prompts had higher immediate and delayed learning outcomes after establishing self-directed metacognitive prompts compared to students in a control condition who did not create prompts (Bannert et al., 2015). Another study revealed differences between high and low performers working on a science learning task with an ALT (Kinnebrew et al., 2017) by combining learning analytics and sequence mining methods. The researchers identified metacognitive strategies like an informed guess-and-check approach to model building, and cognitive strategies like a systematic reading and note-taking approach that differentiated high and low performers. In another study, students with high levels of prior knowledge engaged in more sequences containing accurate levels of cognitive and metacognitive processes than students with low prior knowledge while students with low prior knowledge engaged in more sequences containing inaccurate metacognitive processes than students with high prior knowledge (Taub & Azevedo, 2019). These studies show how students have used cognitive and metacognitive processes during learning with ALTs.

Studies have also examined students' emotions and how they impact learning. One study classified students into clusters and found the positive cluster to have the highest overall learning, the negative activating (i.e., high intensity, such as frustration) emotions cluster to be associated with the lowest learning outcomes, and the negative deactivating (i.e., low intensity, such as boredom) emotions cluster to be associated with learning outcomes in-between the two, during a diagnostic reasoning task (Jarrell et al., 2017). In another study, participants in a condition with the induction of negative emotions had higher performance than participants in a condition with the induction of positive emotions during multimedia learning (Knörzer, Brünken, & Park, 2016). A third study that induced emotions (D'Mello, Lehman, Pekrun, & Graesser, 2014) found that performance was higher on the post-test for participants in a condition where confusion was induced by including contradictory information. Sabourin and Lester (2014) found confusion and boredom to be negatively correlated with learning outcomes during game-based learning, as did Andres et al. (2015) who found sequences of engaging in inefficient gameplay behaviors to be positively correlated with confusion and boredom. Thus, findings are mixed regarding the impact of emotions, with the majority of findings indicating a negative impact of negative emotions on overall learning.

There are thus many research studies that have investigated the use of cognitive and metacognitive processes, or the use of affective and cognitive or metacognitive processes during learning with ALTs. However, there is limited research that has focused on *all three processes together*. These studies have especially not examined specific emotions, specific cognitive and metacognitive processes, and the quality of those processes during learning with ALTs. We started doing this research by examining cognitive processes and emotions during learning with a hypermedia-based intelligent tutoring system (Taub et al., 2018). In this previous work, we examined 38 students' emotions expressed while note taking and summarizing, and how this related to the accuracy of those notes and summaries as well as overall learning. Results revealed that confusion was significantly correlated with the accuracy of summaries, however no emotion was significantly correlated with overall learning. We expanded this work by examining cognitive processes (notes and summaries), but by also examining metacognitive processes, and by increasing our sample size.

1.3. Current study

The goal of this study was to extend our previous research (Taub et al., 2018) and examine students' emotions while engaging in cognitive and metacognitive SRL processes during learning with MetaTutor, a hypermedia-based intelligent tutoring system. It was important to do so because we previously only examined cognitive processes. For this study, we still examined facial expressions of joy, anger, surprise, contempt, confusion, and frustration, as they have been found to be prominent during learning (D'Mello et al., 2014). To extend the previous study, we examined these emotions during four cognitive and metacognitive processes: taking notes and summarizing (cognitive), and judgment of learning and feeling of knowing (metacognitive). These are commonly used cognitive learning strategies (Bonner & Holliday, 2006), and metacognitive monitoring processes (Greene & Azevedo, 2009) that help students ensure their understanding of complex materials. We assessed the accuracy of these processes and how they were associated with facial expressions of emotions and overall learning of the circulatory system with MetaTutor. The accuracy (i.e., correctness) of cognitive processes was defined using latent semantic analysis (LSA) scores and the accuracy of metacognitive processes was scored based on their judgments, the accuracy calibration of those judgments, and a 3-item multiple choice quiz. Overall learning was scored using proportional learning gain (PLG), which measured gain in post-test score (content test out of 30) while taking pre-test score (counterbalanced 30-item test) into account. Thus, in doing this study, this allowed us to examine the relationship between cognitive, metacognitive, *and* affective SRL processes, allowing for more generalizable findings we can use toward developing adaptive learning environments.

We posed the following research questions, which involved investigating emotions during these different processes.

RQ1: *Is there a change in mean evidence score for different emotions while engaging in cognitive and metacognitive processes?*
RQ2: *What is the relationship between evidence scores for different emotions, the accuracy of cognitive and metacognitive processes, and PLG?*
RQ3: *Do mean evidence scores of emotions predict the accuracy of cognitive and metacognitive processes?*

We hypothesized that: H1: There will be significantly higher evidence scores of joy and surprise, and significantly lower evidence scores of anger, contempt, confusion, and frustration when students use cognitive and metacognitive processes. H2: Evidence scores of joy and surprise will be significantly positively correlated with accuracy scores and PLG, and evidence scores of anger, contempt, confusion, and frustration will be significantly negatively correlated with accuracy scores and PLG. H3: Mean evidence scores of joy and surprise will significantly positively predict accuracy, and mean evidence scores of anger, contempt, confusion, and frustration will significantly negatively predict accuracy.

2. Methods

2.1. Participants and materials

Participants were 65 undergraduate and graduate students (83% female) majoring in Education from a large North American university. They were 18–32 years old ($M_{\text{age}} = 21.8$, $SD = 3.2$) and were compensated $10/hour for their participation. Out of all 65 participants with usable face data, 48 made judgments of learning, 32 made feelings of knowing, 33 took notes, and 32 made summaries.

The pre-test and post-test were 30-item multiple-choice tests on the circulatory system developed by a graduate student in Biology with high expertise of the circulatory system. The pre- and post-tests were counterbalanced (i.e., versions A and B), such that about half of participants ($n = 30$) completed pre-test A and post-test B, and the others ($n = 35$) completed pre-test B and post-test A. Thus, all participants took both versions of the test. Cronbach's αs reveal acceptable-good internal consistency for test A ($\alpha = .71$) and test B ($\alpha = .79$). Pre-test scores ranged from 9 (30%) to 28 (93%) out of 30 ($M = 19.2$ (64%), $SD = 4.5$). Post-test scores ranged from 9 (30%) to 29 (97%) out of 30 ($M = 22.4$ (75%), $SD = 4.03$).

2.2. MetaTutor: a hypermedia-based intelligent tutoring system

The MetaTutor system is a learning environment where participants learn about the circulatory system by navigating 47 pages of text content with static images (Azevedo et al., 2018). The interface (see Fig. 1) was strategically developed to foster the use of cognitive and metacognitive processes, such as a timer (top left), table of contents (left), sub-goal progress bars and overall learning goal (top center), text content and image (center), pedagogical agent (top right), and SRL palette (right). The SRL palette allows participants to engage in specific cognitive strategies and metacognitive processes. For example, participants could take notes or summarize (cognitive strategies). They could also judge their understanding of the content on a page (judgment of learning [JOL]) or assess how familiar they already were with the page content (feeling of knowing [FOK]) (metacognitive processes). MetaTutor has four embedded pedagogical agents who each assist with a specific component of SRL: Pam for *planning*, Sam for *strategizing*, Mary for *monitoring*, and Gavin for *guiding*. Participants had 90 min to read the content pages and inspect images and engage in SRL processes to accomplish their overall goal of learning everything they could about the human circulatory system.

2.3. Experimental procedure

This was a 2-day study, where on day 1, participants completed the consent form, a demographics questionnaire, questionnaires on emotions and motivation, and the 30-item pre-test.

On day 2, participants interacted with MetaTutor for the 90-min session to learn about the circulatory system. The timer did stop when engaging in SRL strategies and processes, and so the session could have lasted up to 3 h. First, we calibrated participants' video recordings, where they were asked to sit with a neutral expression for 6 s (to collect a baseline). Next, they were provided with some introductory videos about navigating the system and setting sub-goals. Participants then started the sub-goal setting phase where they set 2 sub-goals with the assistance of Pam the Planner. MetaTutor has a total of 7 pre-determined sub-goals that participants were guided to set (path of blood flow, heartbeat, heart components, blood vessels, blood components, purposes of the circulatory system, and malfunctions of the circulatory system). Once set, participants were provided more videos about using SRL strategies and processes. The next step was to begin the learning session where participants aimed at completing their sub-goals and overall learning goal. Once complete, participants were given the 30-item post-test, followed by a series of self-report questionnaires about their emotions and feedback for the system. Participants were then debriefed, paid, and thanked for their time.

Prior to learning, participants were randomly assigned to either a *prompt and feedback* or *control* condition. These conditions differed in the amount of assistance provided by the pedagogical agents. They provided prompts to engage in SRL processes in the *prompt and feedback* condition. In the *control* condition, the agents did not prompt them or provide them feedback on their performance. For this study, we did not examine condition because we wanted to determine the accuracy in performance on the SRL processes and strategies, and participants could engage in these processes in both conditions. Additionally, preliminary Spearman correlations revealed no significant associations between condition and evidence scores of all 6 emotions, proportional learning gain, or accuracy scores of cognitive and metacognitive processes ($p > .05$).

During learning, we collected multichannel process data. The system recorded log files of all interactions, including mouse clicks (e.g., clicks on the SRL palette, quiz scores) and keyboard entries (e.g., note or summary entries), at the millisecond-level. We also collected video recordings, which we ran through facial recognition software to detect facial expressions of emotions, at a sampling rate of approximately 30 Hz (iMotions, 2018). We used both log files and videos of facial expressions of emotions for our dependent and independent variables for our analyses.

2.4. Data coding and scoring

We created several dependent and independent variables by extracting the data and post-processing them. We scored the pre- and post-test data, use of the SRL palette, and evidence score values provided by the facial recognition software.

2.4.1. Proportional learning gain

Proportional learning gain (PLG) is a measure of learning outcome that examines one's post-test score while taking their pre-test score into account. Thus, we not only examine their actual score at post-test, but also how many points participants gained. We used the following equation (Witherspoon, Azevedo, & D'Mello, 2008), using ratio scores (i.e., score out of 30):

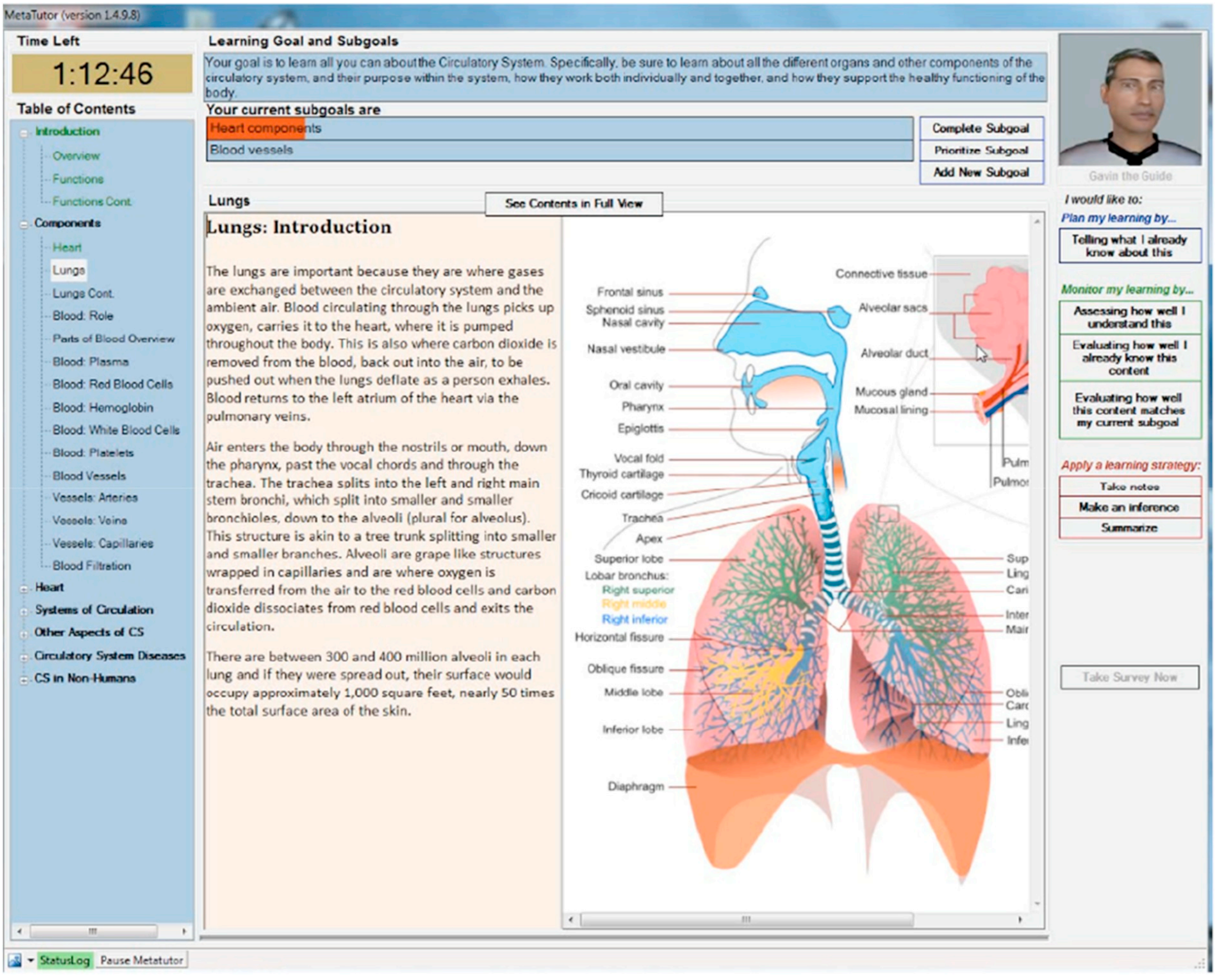

Fig. 1. Screenshot of the MetaTutor interface.

$$\frac{PostTestRatio - PreTestRatio}{1 - PreTestRatio}$$

PLG scores ranged from −2 to .86 (M_{PLG} = .23, *SD* = .45), where on average, participants gained 23% from pre to post (however participants greatly varied in their scores based on the high *SD*). A negative score indicates the participant obtained a lower score on the post-test than on the pre-test, a score of 0 indicates the participant obtained the same score at both pre- and post-test, and a positive score indicates an increase in score from pre to post.

2.4.2. Accuracy of cognitive and metacognitive processes

We examined when participants engaged in cognitive and metacognitive SRL processes during learning, and how accurate they were at engaging in these processes. The cognitive learning strategies we examined were taking notes and making summaries, as they are commonly used strategies and are important for describing content in one's own words to ensure understanding of the material (Bonner & Holliday, 2006). We scored the accuracy of participants' notes and summaries using Latent Semantic Analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007), an application from http://lsa.colorado.edu/. We used a one-to-many approach, which captured the overlap between the page content (i.e., one text) and the content in each of participants' notes or summaries (i.e., many texts). The semantic output generated 2 scores: one score for the overlap of the notes with high school biology content, and one score for the overlap of the notes with general reading ability up to first year college (indicating the student's college level vocabulary of the circulatory system; Landauer, Foltz, & Laham, 1998). This generated a value between 0 and 1, where 1 was an exact match between the notes or summaries and the text content. Thus, the higher the value, the more accurate the notes or summaries are (i.e., the closer they are to the text). For this study, scores on high school biology ranged from .31 to .90 (*M* = .59, *SD* = .14) for taking notes and .14 to .95 (*M* = .60, *SD* = .17) for making summaries and scores for general reading up to first year college ranged from .55 to .94 (*M* = .75, *SD* = .095) for taking notes and .30 to .99 (*M* = .79, *SD* = .13) for making summaries. A score of 1 would indicate exact copying from the text, however we did not find this value in our dataset, indicating that no one copied the text verbatim.

The metacognitive monitoring processes we examined were judgments of learning (JOL) and feelings of knowing (FOK), as they are both key processes that have been shown to positively impact learning and understanding of complex materials (Azevedo et al., 2018; Greene & Azevedo, 2009). A JOL involves a student answering *how well do you feel you understand the content on this page* (on a scale from 1 = *I feel I strongly do not understand* to 6 = *I feel strongly I understand*). Once they make this judgment, they take a 3-item page quiz to check if their judgment is calibrated with their actual understanding. An FOK follows the same steps, however, participants respond to the question, *How strongly do you feel you already know the content on this page* (where 1 = *I do not feel at all strong,* and 6 = *I feel very strong*). We then coded the accuracy of making these metacognitive judgments by allotting points for three different aspects of the processes, for a standardized total of 1. First, we allotted 50% of points to the accuracy of the judgment. A judgment was deemed accurate if the participant judged their correct level of understanding or correct level of familiarity with the content. For example, if a participant judged they did not understand the content at all (a rating of 1) and they obtained a 0% on the quiz, their judgment was accurate. In this case, the participant would get full points for accuracy (.5). Second, we allotted 25% of points to participants' actual quiz score. For example, if the participant obtained 0% on the quiz (a 0 out of 3), they would receive 0 points. Third, we allotted 25% to their response itself out of 6. Using the same example, the participant judged not understanding the content at all (a score of 1 out of 6, or 17%), they would earn .0425 points. Thus, if we refer back to our abovementioned example, the participant would get .5 + 0 + .0425 points, for a total of .5425 out of 1. In this case, the participant did not obtain a high score on the quiz, however since they were aware of that, they obtained points for the correctness of their self-awareness. We chose to develop this coding scheme because it rewards participants for being accurate in their metacognitive monitoring as well as for obtaining high quiz results, as compared to participants who would achieve high quiz results, but were less accurate in their metacognitive monitoring. Participants received an accuracy score for each JOL and FOK they made, where scores for JOL ranged from .25 to 1 (M = .64, SD = .22) and scores for FOK ranged from .21 to .93 (M = .57, SD = .23).

2.4.3. Mean evidence scores

To determine participants' emotions from their facial expressions during learning, we ran their face videos through FACET, automatic facial recognition detection software developed by iMotions (2018), validated by empirical testing (Dente, Küster, Skora, & Krumhuber, 2017). FACET uses a Support Vector Machine (SVM) algorithm to extract facial features to predict action units (i.e., features associated with different facial areas, such as inner brow raiser) and emotions, which are comprised of different action units. The software generates an evidence score value for each emotion and action unit, where an evidence score is defined as the likelihood of a human coder coding for that emotion or action unit on a log (base10) likelihood. For example, a score of 1, 2, or 3 indicates the likelihood of 10, 100, or 1,000 human coders coding for that emotion, respectively. We investigated joy, anger, surprise, contempt, confusion, and frustration for our analyses because they relate closely to learning with intelligent tutoring systems (D'Mello et al., 2014).

The video input to the iMotions software contained 25 frames (photos) per second. The classifier analyzed these frames and predicted the value of a learner's emotions for each frame. Hence, the raw output file from iMotions contained 25 evidence score values for the 6 emotions listed above per second. We preprocessed the raw output files to aggregate emotion values per second and transformed the evidence score to a representation that could be easily used for our analysis. These preprocessing steps were used to rescale and smoothen the noise in the raw data, and therefore do not alter the values obtained from iMotions. Our preprocessing step is briefly described below:

1. Based on our consultation with experts from iMotions, we replaced the negative values in the raw data to zero.
2. The standard feature rescaling process was applied to convert the evidence values of each emotion to scale the range in [0–1]. The formula used for rescaling was:

$$rescaled(x) = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

3. To remove the noise in the data represented as a sudden spike of evidence score for an emotion, we applied a median filtering process, using a window size of 11 observations. The center evidence value in the window was replaced with the median of all the evidence values in the window. The window size was selected after several iterations with varied window size.
4. To remove noise and to smoothen the data we applied a standard mean filter, with a window size of 25. This window covered a 1-s interval. The mean filtering replaces the center evidence value in the window with the mean score of all the evidence values in the window.
5. The evidence values of emotions per second were computed as the mean of all evidence value observations over a period of 1 s.

3. Results

3.1. Research question 1: Is there a change in mean evidence score for different emotions while engaging in cognitive and metacognitive processes?

For this research question, we examined if there were significant differences between participants' observed emotions while they engaged in SRL processes. We conducted four separate one-way repeated-measures ANOVAs, one for each SRL process. We conducted separate analyses because participants engaged in different numbers of each of these processes, which would result in unequal data cells per participant. For all tests, Mauchly's test of Sphericity was violated ($p < .05$), and so we used Greenhouse Geisser corrected values.

The first repeated-measures ANOVA revealed there were significant differences in evidence scores for different emotions when making JOLs; $F(2.788, 131.014) = 7.027$, $p < .001$, $\eta_p^2 = .13$. Pairwise comparisons (see Table 1) revealed that joy, anger, confusion, and frustration had significantly higher mean evidence scores than contempt, and anger had a significantly higher mean evidence score than frustration when participants made JOLs.

Our second repeated-measures ANOVA revealed that participants' mean evidence scores were significantly different for different emotions when they made FOKs; $F(2.301, 71.333) = 5.32$, $p < .01$, $\eta_p^2 = .15$. Pairwise comparisons (see Table 1) revealed that joy and anger had significantly higher mean evidence scores than contempt, and anger had a significantly higher mean evidence score than frustration when

Table 1
Descriptive statistics for RM-ANOVAs with emotions during cognitive and metacognitive processes.

	Evidence Score $M(SD)$						RM-ANOVA	
	Joy	Anger	Surpr.	Cont.	Conf.	Frust.	F	Comparison
JOL (n = 48)	.08 (.094)	.14 (.12)	.088 (.083)	.048 (.045)	.11 (.093)	.084 (.073)	7.027***	4 < 1 = 2 = 5 = 6, 6 < 2
FOK (n = 32)	.085 (.075)	.14 (.12)	.078 (.075)	.054 (.054)	.12 (.10)	.096 (.086)	5.315**	4 < 1 = 2, 6 < 2
TN (n = 33)	.088 (.083)	.12 (.090)	.087 (.065)	.049 (.046)	.092 (.089)	.068 (.057)	4.209*	4 < 1 = 2, 6 < 2
SUMM (n = 32)	.093 (.082)	.14 (.12)	.088 (.063)	.056 (.043)	.10 (.089)	.088 (.076)	4.832**	4 < 1 = 2, 6 < 2

*$p < .05$, **$p < .01$, ***$p < .001$.
Note. JOL = judgment of learning, FOK = feeling of knowing, TN = take notes, SUMM = summarize, Surpr. = surprise, Cont. = contempt, Conf. = confusion, Frust. = frustration. All F values are reporting using Greenhouse-Geisser corrections. Comparisons are only shown for evidence scores with significant differences. For comparisons, 1 = joy, 2 = anger, 3 = surprise, 4 = contempt, 5 = confusion, 6 = frustration.

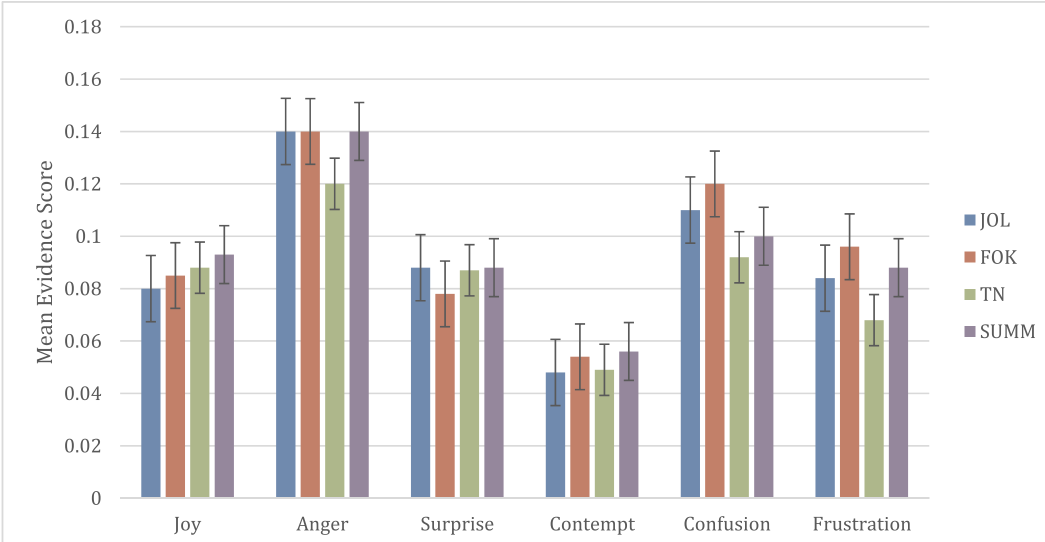

Fig. 2. Mean Evidence Scores and Standard Error Bars for SRL Processes.
Note. JOL = judgment of learning, FOK = feeling of knowing, TN = take notes, SUMM = summarizing.

participants engaged in FOKs.

Results from our third repeated-measures ANOVA revealed that participants' mean evidence scores were significantly different for different emotions during note taking; $F(2.629, 84.129) = 4.21$, $p < .05$, $\eta_p^2 = .12$. Pairwise comparisons (see Table 1) revealed that when participants were taking notes, mean evidence scores of joy and anger were significantly higher than contempt, and mean evidence scores of anger were significantly higher than frustration.

Our final repeated-measures ANOVA revealed that mean evidence scores were significantly different for different emotions during summarizing; $F(2.710, 84.013) = 4.83$, $p < .01$, $\eta_p^2 = .14$. Pairwise comparisons (see Table 1) revealed that joy and anger had significantly higher mean evidence scores than contempt, and anger had significantly higher mean evidence scores than frustration while participants were summarizing. See Fig. 2 for a graphical representation of all four ANOVA results. Although we were not able to examine this statistically, this figure demonstrates that overall, mean evidence scores were fairly similar for all emotions for all four processes, with contempt having the lowest scores, and anger having the highest scores for all processes.

3.2. Research question 2: what is the relationship between evidence scores for different emotions, the accuracy of cognitive and metacognitive processes, and PLG?

For this research question, we ran 4 separate correlations, one for each cognitive or metacognitive process. Each correlation matrix is represented in Tables 2–5 for JOLs, FOKs, note taking, and summarizing, respectively. See section 2.4.2 for a description of how cognitive and metacognitive processes were coded.

Table 2
Correlation matrix for mean evidence scores, accuracy score of JOLs, and PLG. (n = 48).

	1	2	3	4	5	6	7	8
1 – Joy	–							
2 – Anger	.091	–						
3 - Surprise	.10	.31*	–					
4 – Contempt	.77***	.07	.04	–				
5 – Confusion	-.13	.66***	-.15	-.14	–			
6 – Frustration	.21	.75***	.057	.24	.67***	–		
7 – PLG	.17	.12	.006	.056	.26	.28$^+$	–	
8 - AccuracyScore	-.020	.19	.067	-.071	-.11	-.062	-.040	–

*$p < .05$, **$p < .01$, ***$p < .001$, $^+p < .06$.
Note. PLG = proportional learning gain.

Table 3
Correlation matrix for mean evidence scores, accuracy score of FOKs, and PLG. (n = 32).

	1	2	3	4	5	6	7	8
1 – Joy	–							
2 – Anger	.20	–						
3 - Surprise	.36*	.25	–					
4 – Contempt	.69***	.19	.69***	–				
5 – Confusion	-.092	.60***	-.30	-.19	–			
6 – Frustration	.22	.82***	-.030	.044	.73***	–		
7 – PLG	.13	-.016	.044	.23	.21	.11	–	
8 - AccuracyScore	-.059	-.028	-.37*	-.34$^+$	.080	-.090	-.26	–

*$p < .05$, **$p < .01$, ***$p < .001$, $^+p < .06$.
Note. PLG = proportional learning gain.

Overall, results revealed significant correlations between evidence scores of different emotions. For example, for all four processes, there were significant correlations between joy and contempt ($p < .001$), anger and confusion ($p < .05$), anger and frustration ($p < .001$), and confusion and frustration ($p < .05$). In addition, other mean evidence scores of emotions were correlated with each other for specific processes. During JOLs, anger and surprise were correlated ($r(46) = .31$, $p < .05$). During FOKs, surprise and contempt were correlated ($r(30) = .69$, $p < .001$). Joy and surprise were correlated during FOKs ($r(30) = .36$, $p < .05$) and note taking ($r(31) = .39$, $p < .05$). During note taking, joy and confusion were negatively correlated ($r(31) = -.37$, $p < .05$) and contempt and confusion were negatively correlated ($r(31) = -.36$, $p < .05$). During summarizing, joy and frustration were correlated ($r(30) = .37$, $p < .05$).

Furthermore, we found some correlations between mean evidence scores of emotions and accuracy scores of cognitive and metacognitive processes. Specifically, surprise and FOK score accuracy were negatively correlated ($r(30) = -.37$, $p < .05$) and frustration and LSA score of notes for general reading up to first year college were positively correlated ($r(31) = .44$, $p < .05$). These correlations indicate that when participants had higher mean evidence scores of surprise, they had lower FOK accuracy scores, and when participants had higher mean evidence scores of frustration, their LSA scores of general reading (i.e., use of college level vocabulary of the circulatory system) during note taking were higher.

Table 4
Correlation matrix for mean evidence scores, LSA scores for notes, and PLG. (n = 33).

	1	2	3	4	5	6	7	8	9
1 – Joy	–								
2 - Anger	.15	–							
3 – Surprise	.39*	.33	–						
4 – Contempt	.79***	.041	.22	–					
5 – Confusion	-.37*	.44*	-.33$^+$	-.36*	–				
6 – Frustration	.19	.65***	-.076	.17	.37*	–			
7 – PLG	.19	-.27	-.023	.23	-.14	.028	–		
8 – LSA (HSBio)	.33$^+$	.10	.26	.30	-.24	.042	-.049	–	
9 – LSA (GenRead)	.14	.19	.021	.14	.10	.44*	-.029	.063	–

*p < .05, **p < .01, ***p < .001.
Note. PLG = proportional learning gain, LSA = latent semantic analysis, HSBio = high school biology, GenRead = general reading up to first year college.

Table 5
Correlation matrix for mean evidence scores, LSA scores for summaries, and PLG. (n = 32).

	1	2	3	4	5	6	7	8	9
1 – Joy	–								
2 - Anger	.10	–							
3 – Surprise	.10	.17	–						
4 – Contempt	.71***	-.12	.080	–					
5 – Confusion	-.025	.57**	-.28	-.28	–				
6 – Frustration	.37*	.76***	-.14	.14	.67***	–			
7 – PLG	.12	-.14	-.27	.12	.019	.19	–		
8 – LSA (HSBio)	.25	.087	.28	.20	.029	.095	.098	–	
9 – LSA (GenRead)	.020	.093	.20	.004	.12	.12	-.17	.83***	–

*p < .05, **p < .01, ***p < .001.
Note. PLG = proportional learning gain, LSA = latent semantic analysis, HSBio = high school biology, GenRead = general reading up to first year college.

3.3. *Research question 3: Do mean evidence scores of emotions predict the accuracy of cognitive and metacognitive processes?*

For this research question, we ran two linear regressions based on the significant correlations between surprise and FOK accuracy score and frustration and LSA score for general reading. We did not include PLG in our models because there were no significant correlations between PLG and other variables.

Our first simple regression revealed that mean evidence score of surprise was a significant predictor of FOK accuracy score; $F(1, 30) = 4.73$, $p < .05$, $R^2 = .14$. FOK accuracy score decreased 1.114 points for each mean evidence score point increase in surprise. Our second simple regression revealed that mean evidence score of frustration was a significant predictor of LSA score (general reading) for notes; $F(1, 31) = 7.219$, $p < .05$, $R^2 = .19$. LSA score for notes increased by 3.46 points for each mean evidence score point increase in frustration.

4. Discussion

The goal of this study was to examine the relationship between cognitive, metacognitive, and affective self-regulated learning processes during learning with an intelligent tutoring system. We examined participants' evidence scores of different emotions (joy, surprise, anger, contempt, confusion, and frustration), accuracy scores of cognitive (notes and summaries) and metacognitive (JOLs and FOKs) processes, and proportional learning gain. Overall, we found that these processes are related to each other, demonstrating the important, and not necessarily disadvantages, of expressing frustration during learning of complex topics with advanced learning technologies (D'Mello et al., 2014).

Our first research question revealed that mean evidence scores of emotions were significantly different during all cognitive and metacognitive SRL processes. Specifically, joy and anger had significantly higher mean evidence scores than contempt, and anger had a significantly higher mean evidence score than frustration. This partially confirms H1 because we predicted higher levels of joy and surprise and we did find higher scores for joy compared to contempt, however we also predicted lower levels of anger, contempt, confusion, and frustration, and found higher scores of anger. Theoretically, this may indicate that higher levels of joy reveal that at some points during the learning session participants were able to resolve any confusion they had. However, at other times, participants were not able to resolve the impasse, resulting in higher levels of anger. The model of affective dynamics (D'Mello & Graesser, 2012) does not focus on anger, but perhaps anger and frustration are closely related (which we did find in RQ2 with a high correlation) and should be added to the model. Thus, from our results, we see evidence for moments of both resolving and not resolving a state of confusion.

Our second research question revealed an association between cognitive, metacognitive, and affective processes from our significant correlations between mean evidence scores of emotions and accuracy of cognitive and metacognitive processes. Specifically, surprise was negatively correlated with FOK accuracy score, and frustration was positively correlated with LSA score of notes on general reading up to first year college. This did not confirm H2 because we predicted the opposite: for joy and surprise to be positively correlated with accuracy and anger, contempt, confusion, and frustration to be negatively correlated with accuracy. We also did not find any significant correlations between emotions and PLG, which does not support H2 either. The model of affective dynamics does emphasize the importance of confusion (D'Mello & Graesser, 2012; D'Mello et al., 2014), however frustration is seen as a result of not resolving that confusion and moving towards boredom and disengagement. In our study, frustration was positively correlated with accuracy of notes, thereby demonstrating a possible benefit for frustration as well (Munshi et al., 2018). From an SRL perspective, participants might have been surprised from reading facts about the circulatory system because it misaligned with their perceived prior knowledge of the topic, which negatively impacted their ability to engage in accurate FOKs. In addition, it is possible that participants were feeling frustrated because they did not understand the content, pushing them to engage in adaptations to ensure their notes were accurate. These results do not align with previous research that found a negative impact of negative emotions on overall learning (Jarrell et al., 2017; Sabourin & Lester, 2014). However, studies that focused on specific negative emotions were mixed, such that some emphasized the negative impact of confusion and boredom (Andres et al., 2015), some found a positive impact of confusion (D'Mello et al., 2014), but we found a positive impact of frustration. Thus, more studies are needed to investigate the impact of frustration on learning.

Results from our third research question revealed that not only did we find significant correlations between surprise and frustration with accuracy of FOKs and notes, respectively, but mean evidence scores of these emotions predicted these accuracy scores as well. Again, lower scores of surprise predicted higher accuracy of FOKs, and higher scores of frustration predicted higher accuracy of notes. This did not confirm H3 because we predicted that joy and surprise would positively predict accuracy, and anger, contempt, confusion, and frustration would negatively predict accuracy. Theoretically, we again question the positive role of frustration on accuracy of notes, such that higher evidence scores of frustration might not indicate the participant is on their way towards disengagement. Once again, frustration could have caused participants to persist to take more accurate notes, whereas surprise could have impeded the ability to engage in accurate metacognitive monitoring. These results do not align with previous studies that found a positive impact of positive emotions and a negative impact of negative emotions on overall learning (Jarrell et al., 2017; Sabourin & Lester, 2014). Surprise, a positive emotion was negatively associated with accuracy of FOK score, and frustration, a negative emotion was positively associated with accuracy of notes. These previous studies, however, did not investigate the impacts of specific emotions on learning, and so the direct influence of these emotions remains unclear.

4.1. Implications for learning and instruction

These results have implications for learning as it leads us to question our current approaches for examining overall learning. For example, many studies investigate how using an ALT impacts an overall learning outcome score and often fail to find a significant learning gain effect (e.g., Taub, Azevedo, Bradbury, Millar, & Lester, 2018). However, if we include examining processes *during* learning, this can provide more specific details of *how* students are progressing through a learning session. In this study, we did not find an association between evidence scores of emotions and proportional learning gain, however we did find significant correlations between evidence scores of emotions and the accuracy of some cognitive and metacognitive processes, and surprise and frustration significantly predicted metacognitive and cognitive process accuracy, respectively. This can be informative for us to provide students with scaffolding while learning complex topics with ALTs to address their needs in real-time.

Results contribute to theories of SRL because we focus on the relationship between SRL and emotions, which is not typically addressed. For example, it is important to understand how emotions can impact engaging in different learning strategies during phase three of the information processing theory (Winne, 2018). Specifically, we were able to understand how different emotions related to different micro-level cognitive and metacognitive SRL processes (Greene & Azevedo, 2009) students used during learning. This allows us to further understand the complex nature of SRL, and how it relates to different learner characteristics that can impact how they self-regulate.

Results also demonstrate the important role of frustration on SRL. The model of affective dynamics, as well as other studies, have demonstrated that confusion can positively impact learning (D'Mello et al., 2014). Our findings reveal that frustration can be a beneficial emotion as well, positively predicting note-taking accuracy. In addition, surprise, a positive emotion, is not always beneficial. Surprise negatively predicted FOK accuracy, and so experiencing some positive emotions might not be as beneficial as they seem.

Results demonstrate the strengths of using video data of facial expressions to investigate emotions instead of relying on self-report or observational data, which can be subject to biases (e.g., experimenter bias). This seems to be especially influential because previous studies mentioned that examined emotions during learning with different types of ALTs used either self-report (D'Mello et al., 2014; Jarrell et al., 2017; Sabourin & Lester, 2014) or observational data (Andres et al., 2015). These studies found opposite results from the current study, perhaps due to the nature of data collection. This leads us to question the use of obtrusive vs. unobtrusive measures to detect emotions during learning because they could yield contradictory results.

Using multimodal data, such as videos of facial expressions with log-file data helps us capture the temporal nature of SRL as events that unfold over time (Azevedo et al., 2018; D'Mello et al., 2017). These data channels provide us information about students' use of different cognitive and metacognitive processes during learning with an ALT as well as which emotions they were expressing as they were engaging in these processes. Thus, using multimodal multichannel data allows us to investigate how emotions change during the use of different SRL processes.

4.2. Limitations

Although our results yielded interesting findings, there are some limitations, which we must address. First, we investigated emotions and SRL processes at the aggregated level, and not at each instance of engaging in a cognitive or metacognitive process (i.e., each participant's instances were averaged into one score instead of having many scores for each specific event). In doing so we were unable to examine how both emotions and SRL processes changed over instances. Furthermore, in aggregating the data, we cannot be sure about the directionality or causality of the variables. For example, were participants surprised because they obtained low scores, or were they surprised first, causing low scores? In addition, our sample size might not have been large enough to examine the relationships between these processes. Our sample decreased even more because some participants' video data were not adequate for the recognition software. Thus, we should ensure that in future studies, all video data are acceptable for these types of software. Finally, as we are unaware of the algorithms that iMotions uses to detect emotions, we cannot claim this is the only valid tool that can be used to measure emotions.

4.3. Future directions

Findings from this study pave the way for many future directions towards teaching students how to use cognitive and metacognitive SRL processes during learning, as well as being aware of their emotions and motivations. This would include analyzing the interplay between these processes and how they change over time (e.g., Cloude, Taub, & Azevedo, 2018).

Another direction for future research could include adding more data channels. In this study, we used log files and videos of facial expression data, however future studies can include participants' eye-tracking data and electrodermal activity, which have both been used to examine emotions during learning with ALTs (e.g., Harley, Bouchet, Hussain, Azevedo, & Calvo, 2015; Jaques, Conati, Harley, & Azevedo, 2014).

Furthermore, future research can expand from only examining emotions to making participants *aware* of their emotions so we can understand an appropriate, beneficial amount of an emotion (see Harley, Lajoie, Frasson, & Hall, 2017). For example, is a certain amount of frustration ideal for learning? Is too much surprise negatively related to an inability to accurately metacognitively monitor one's learning? If we can determine the most effective levels of emotions, we can use this information to teach students how to regulate their emotions (Azevedo et al., 2017) to ensure their levels of emotions play a beneficial role in learning.

The goal of this research is to assess the most optimal uses of cognitive, affective, metacognitive, and motivational SRL processes so we can develop adaptive ALTs that cater to each student's individual learning needs in real-time. This would ensure that all students are gaining a positive learning experience, learning about complex phenomena that will help them succeed in school and in the future.

Acknowledgments

This research was supported by funding from the National Science Foundation (DRL#1431552; DRL#1660878, DRL#1661202) and the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006). The authors would like to thank the members from the SMART Lab at UCF for their assistance.

References

Andres, J. M. L., Rodrigo, M. M. T., Baker, R. S., Paquette, L., Shute, V. J., & Ventura, M. (2015). Analyzing student action sequences and affect while playing Physics Playground. *Paper presented at the international workshop on affect, meta-affect, data and learning (AMADL 2015) at the 17th international conference on artificial intelligence in education (AIED 2015), Madrid, Spain*.

Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. H. Schunk, & J. A. Greene (Eds.). *Handbook of self-regulation of learning and performance* (pp. 254–270). (2nd ed.). New York, NY: Routledge.

Azevedo, R., Taub, M., Mudrick, N. V., Millar, G. C., Bradbury, A. E., & Price, M. J. (2017). Using data visualizations to foster emotion regulation during self-regulated learning with advanced learning technologies. In J. Buder, & F. Hesse (Eds.). *Informational environments: Effects of use and effective designs* (pp. 225–248). Amsterdam, The Netherlands: Springer.

Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short- and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior, 52*, 293–306.

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice—a learning by teaching system: Betty's brain. *International Journal of Artificial Intelligence in Education, 26*, 350–364.

Bonner, J. M., & Holliday, W. G. (2006). How college science students engage in note-taking strategies. *Journal of Research in Science Teaching, 43*, 786–818.

Cloude, E. B., Taub, M., & Azevedo, R. (2018). Investigating the role of goal orientation: Metacognitive and cognitive strategy use and learning with intelligent tutoring systems. In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.). *Proceedings of the 14th international conference on intelligent tutoring systems (ITS 2018)* (pp. 44–53). Amsterdam, The Netherlands: Springer.

Dente, P., Küster, D., Skora, L., & Krumhuber, E. G. (2017). Measures and metrics for automatic emotion classification via FACET. In J. Bryson, M. De Vos, & J. Padget (Eds.). *Proceedings of the conference on the study of artificial intelligence and simulation of behaviour (AISB)* (pp. 160–163). Red Hook, NY: Curran Associates.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145–157.

D'Mello, S., Kappas, A., & Gratch, J. (2017). The affective computing approach to affect measurement. *Emotion Review, 10*, 174–183.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29*, 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003.

Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of sophisticated mental models. *Contemporary Educational Psychology, 34*, 18–29.

Harley, J. M., Bouchet, F., Hussain, S., Azevedo, R., & Calvo, R. (2015). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior, 48*, 615–625.

Harley, J. M., Lajoie, S. P., Frasson, C., & Hall, N. (2017). Developing emotion-aware, advanced learning technologies: A taxonomy of approaches and features. *International Journal of Artificial Intelligence in Education, 27*, 268–297.

iMotions (2018). *Attention tool.* Boston, MA: iMotions Inc. [Computer software] Version 7.1.

Jaques, N., Conati, C., Harley, J., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.). *Proceedings of the 12th international Conference on intelligent tutoring systems—lecture Notes in computer science 8474* (pp. 29–38). Amsterdam, The Netherlands: Springer.

Jarrell, A., Harley, J. M., Lajoie, S., & Naismith, L. (2017). Success, failure and emotions: Examining the relationship between performance feedback and emotions in diagnostic reasoning. *Educational Technology Research & Development, 65*, 1263–1284.

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2017). Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. *IEEE Transactions on Learning Technologies, 10*, 140–153.

Knörzer, L., Brünken, R., & Park, B. (2016). Emotions and multimedia learning: The moderating role of learner characteristics. *Journal of Computer Assisted Learning, 32*, 618–631.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In J. Zhang, T. Mitrovic, D. Chin, & L. Chen (Eds.). *Proceedings of the 26th Conference on user modeling, adaptation, and personalization* (pp. 131–138). New York, NY: ACM.

Pekrun, R. (2006). The cintrol-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315–341.

Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing, 5*, 45–56.

Taub, M., & Azevedo, R. (2019). Investigating students' cognitive and metacognitive self-regulated learning during learning with a hypermedia-learning environment. *International Journal of Artificial Intelligence in Education, 29*, 1–28.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction, 54*, 93–103.

Taub, M., Mudrick, N. V., Rajendran, R., Dong, Y., Biswas, G., & Azevedo, R. (2018). How are students' emotions associated with the accuracy of their note taking and summarizing during learning with ITSs? In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.). *Proceedings of the 14th international conference on intelligent tutoring systems (ITS 2018)* (pp. 233–242). Amsterdam, The Netherlands: Springer.

Winne, P. H. (2018). Cognition and metacognition within self-regulated learning. In D. H. Schunk, & J. A. Greene (Eds.). *Handbook of self-regulation of learning and performance* (pp. 36–48). (2nd ed.). New York, NY: Routledge.

Witherspoon, A., Azevedo, R., & D'Mello, S. (2008). The dynamics of self-regulatory processes within self- and externally-regulated learning episodes. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.). *Proceedings of the international conference on intelligent tutoring systems—lecture notes in computer science 5091* (pp. 260–269). Berlin: Springer.