Investigating the Utility of Eye-Tracking Information on Affect and Reasoning for User Modeling

K. Muldner[1], R. Christopherson[2], R. Atkinson[2], and W. Burleson[1]

Arizona State University, School of Computing and Informatics/Arts, Media and Engineering[1], Psychology of Education[2], Tempe, Arizona
{Katarzyna.Muldner,rmchris3,Robert.Atkinson,Winslow.Burleson}@asu.edu

Abstract. We investigate the utility of an eye tracker for providing information on users' affect and reasoning. To do so, we conducted a user study, results from which show that users' pupillary responses differ significantly between positive and negative affective states. As far as reasoning is concerned, while our analysis shows that larger pupil size is associated with more constructive reasoning events, it also suggests that to disambiguate between different kinds of reasoning, additional information may be needed. Our results show that pupillary response is a promising non-invasive avenue for increasing user model bandwidth.

1 Introduction

Increasing model *bandwidth*, i.e., the amount and quality of information available to a user model, without disrupting a user's interaction with an adaptive system is a key user modeling challenge [1]. Arguably, the higher the level of the information to be captured, the more complex a user model's construction becomes, because it may require sophisticated Artificial Intelligence (AI) techniques and innovative sensing devices. Thus, it is increasingly critical to show that (1) it is feasible to capture the necessary user states (*feasibility* requirement) and (2) the the increased model complexity improves system usability (*usability* requirement).

Here, we focus on the feasibility requirement, by investigating the utility of pupillary data provided by an eye tracker for informing a user model on high-level user states related to affect and reasoning style. Information on how a user is feeling and/or reasoning can be highly valuable, as it enables an adaptive system to respond appropriately to the user's needs and preferences. For instance, users engage in frustrating tasks on a computer significantly longer after an empathetic computational response (e.g., [2]); learning outcomes are improved when computational tutors provide tailored prompts to foster *meta-cognitive* skills, i.e., domain-independent reasoning abilities (e.g., [3]). However, information on high-level states is rarely observable and so challenging to obtain unobtrusively. A promising avenue corresponds to innovative sensing devices, which capture users' physiological responses that are a natural by-product of their interaction

G.-J. Houben et al. (Eds.): UMAP 2009, LNCS 5535, pp. 138–149, 2009.

with an adaptive system. For instance, D'Mello and Graessor [4] rely on machine learning to show that dialog and posture features can discriminate between affective states of boredom, confusion, flow and frustration. Burleson et al. [5] show that a learning companion, based on a model incorporating information from a pressure mouse, posture chair, video camera, and skin conductance bracelet, impacts students' motivation and attitudes towards the companion.

There is also work exploring how information on gaze patterns from an eye tracker can inform a user model, for instance to determine (1) attention shifts and/or focus [6,7]; (2) high-level reasoning via *self-explanation* [8], the process of explaining and clarifying instructional material to oneself [9]. Another branch of eye-tracking research focuses on pupil dilation. In tightly-controlled experimental settings, there is a clear link between mental effort and pupil dilation [10,11] and affect and pupil dilation [10,12], where affective responses and mental effort increase pupil size. However, these evaluations rely on an experimental protocol where the context is far removed from what a natural interaction with an adaptive system might entail. For instance, subjects categorize emotionally charged words [12], or listen to affect-induced audio at controlled time intervals [13]. When transferred to more realistic applications, there have been mixed results with respect to reliability of pupil information. Several attempts to find a link between reading difficulty and mental effort have failed (e.g., [14,15]), although Igbal et al. [14] did find that pupil size increased with more difficult file manipulation tasks. Conati et al. [8] failed to find a link between pupillary response and self-explanation, which is presumably associated with mental effort, and so pupil size. Clearly, more work is needed assessing the link between mental effort, affect and pupil response, and its utility for user modeling. Our research is a step in this direction.

As our test-bed application, we rely on the Example Analogy (EA)-Coach [16,17], an adaptive learning environment we developed that supports meta-cognition during example-based learning. Although a formal evaluation of the EA-Coach showed that in general, it effectively fosters meta-cognition [17], it also suggested that some students require more support than is currently provided by the system. Thus, we would like to extend the tutor with affective and meta-cognitive scaffolding, to help all students learn effectively from APS. Given that this scaffolding will be based on the EA-Coach user model, as the first step, we have been investigating ways to increase the model's bandwidth to provide adequate information on the relevant student states.

We begin with an introduction to the EA-Coach and its user model. We then describe the user study we conducted to evaluate whether affect and reasoning style impacts pupillary response. After we present our results, we conclude and provide suggestions for some future work.

2 The EA-Coach

The Example-Analogy (EA) Coach [16,17] is an adaptive learning environment that fosters meta-cognitive skills during *analogical problem solving* (APS),

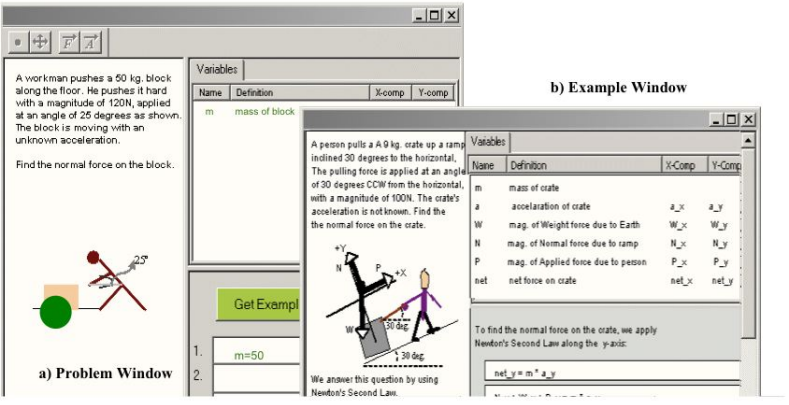

Fig. 1. The EA-Coach Interface: (a) problem window and (b) example window

i.e., using examples to aid problem solving, in the target domain of introductory Newtonian physics. Two meta-cognitive skills that are relevant to APS and therefore targeted by the EA-Coach include:

- *min-analogy*: solving the problem on ones own as much as possible instead of by copying from examples [18]
- *explanation-based learning of correctness (EBLC)*: a form of self-explanation that involves using ones existing common sense, overly general and/or domain knowledge to infer new rules that explain how a given example solution step is derived [19].

The EA-Coach includes an interface that students use to solve problems and refer to examples (see Fig. 1(a) and (b), respectively). To solve problems, students draw free-body diagrams and type equations in the problem window (see Fig. 1a). The EA-Coach does not constrain input of the problem solution, and students may enter the solution steps in any order and/or skip steps. The tutor provides immediate feedback for correctness on students' problem-solving entries, by coloring correct vs. incorrect entries red or green, respectively. It also informs students when it can not interpret their problem entries, but does not provide any other feedback or hints (e.g., related to physics).

While working on a problem, a student can ask for an example (via the 'Get Example' button, see Fig. 1a). In response, the EA-Coach adaptively selects the one from its example pool that has the best potential to help the student solve the problem and learn from doing so, and presents it in the example window (see Fig. 1b). Example selection is accomplished by a decision-theoretic process that we described in [17]; a key aspect of this process is EA-Coach user model. During selection, the model generates a *prediction* of how (1) student characteristics and (2) similarity between the problem and a candidate example will impact min-analogy and EBLC, and subsequent learning and problem solving outcomes. Once an example is presented to a student, the model relies

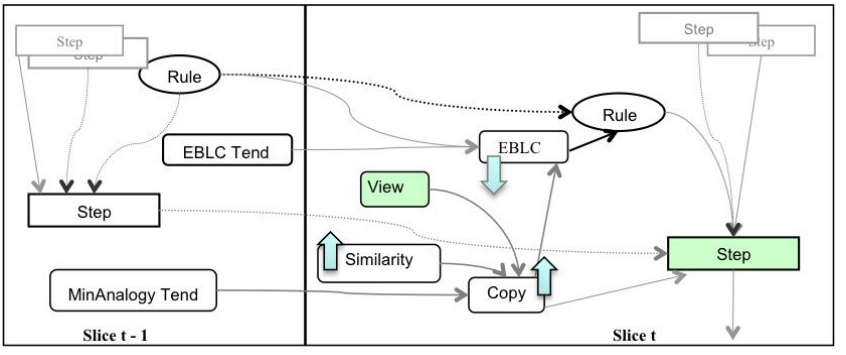

Fig. 2. Fragment of the EA-Coach User Model

on the same sources of information (problem/example similarity, student characteristics), as well as a student's interface actions, to update its *assessment* of the student. This assessment enables the EA-Coach to track how the student's knowledge and meta-cognition evolve as a result of interacting with the tutor. The same model structure is used during both modes (prediction, assessment).

2.1 The EA-Coach User Model

The EA-Coach user model [17,20] corresponds to a dynamic Bayesian network, a fragment of which is shown in Fig. 2. The network's backbone consists of nodes representing the solution steps for the problem the student is currently solving, and the domain rules deriving those steps (see *Step* and *Rule* nodes in Fig. 2), as well as two nodes to model meta-cognitive tendencies (see *EBLCTend* and *MinAnalogyTend* nodes in Fig. 2). For each problem-solving action being modeled, the network also includes nodes accounting for the impact of the example on the APS process (see Fig. 2, slice t), as follows: (1) *similarity* nodes, to capture the similarity between the target problem and example; (2) *copy* nodes, to capture the probability that a student generated the corresponding solution step by copying from the example; (3) *EBLC* nodes, to capture the probability that a student self-explained the corresponding rule from the example with EBLC; (4) *view* nodes representing whether a student viewed the corresponding example step[1]. When a student generates a solution step in the EA-Coach interface, the model enters this and example-viewing information as evidence (see shaded *Step* and *View* nodes in slice t in Fig. 2), and subsequently updates its belief in how the student reasoned (copied vs. self-explained through EBLC). For instance, in Fig. 2, slice t, a high problem/example step similarity increases the probability of copying, which decreases the probability of EBLC and so learning of

[1] View nodes are only included during assessment mode; the viewing information is provided by a *masking* interface that covers the example solution and is uncovered by moving the mouse over a region; this interface is not shown in Fig. 1 and was not used in the evaluation described in Section 3.

the corresponding rule. Note that the EA-Coach model has low bandwidth - for instance, the only explicit information on if and how a student self-explained with EBLC corresponds to whether the student viewed the related step in the example window and/or her subsequent problem-solving entry.

When we evaluated the EA-Coach, we found that in general, the tutor encouraged students to engage in the target meta-cognitive behaviors of min-analogy and EBLC [17]. However, the evaluation also showed that some students need more explicit scaffolding than what is currently provided by the system. Therefore, we have been working on designing this support. Since both affect and meta-cognition play a key role in the learning process, we are exploring incorporating affective support into the EA-Coach, as well as enriching its current level of meta-cognitive support. In order for this new scaffolding to be tailored to a student's needs, a challenge relates to how the model can obtain the necessary information, while at the same time preserving the free nature of the interaction with the EA-Coach.

3 Experiment: User Study

The aim of our study was to explore the utility of information derived from sensing devices for modeling high-level user states related to affect and reasoning style. Here, we focus our analysis on data coming from one sensor: an eye tracker. The study participants were 15 university students, who were either in the process of taking a first year university physics course, or had taken a physics course in high school, but had not taken any higher-level physics courses. This was the strategy used in the study methodology in [17], on which this study is directly based. The rationale behind this requirement was to include subjects who have had some exposure to physics, but who were not so expert as to find the physics problems trivial to solve, as we felt that this would provide less varied data. Subjects were either (1) payed for their participation (five subjects) or (2) given extra credit for a course they were enrolled in (ten subjects).

Each study session was conducted separately. During a session, a participant was introduced to the EA-Coach interface, calibrated an eye tracker, and used the EA-Coach to work on two Newton's Second Law problems of the type shown in Fig. 1[2]. For each problem that subjects solved with the EA-Coach, they were given the choice of accessing an example, which was provided by the EA-Coach. The similarity between the problem/example pairs was manipulated, so that for one of the problems, subjects received a more similar example with respect to the target problem than for the other problem (following the method described in [17]). By providing two different scenarios (high + low similarity), we hoped to maximize opportunities for subjects to express a wide range of affective and reasoning behaviors. The order of both the problems and the similarity type (low, high) was fully counterbalanced. Subjects were told that they had 60 minutes per problem, but that could stop before that if they wished.

[2] Prior to and following a session, participants were also asked to fill in questionnaires to assess their physics and self-regulation knowledge.

As subjects worked with the EA-Coach, a Tobii T60 eye tracker captured their gaze information. This eyetracker is a non-intrusive model that is fully integrated into a 17" monitor and so from a participant's perspective, it appears as a regular computer screen. To calibrate the eye tracker, participants were asked to focus on a series of 16 dots on the computer screen; this phase took approximately one minute. We also captured other physiological data using a set of non-invasive sensors, but this data analysis is in progress and is not reported here (the sensors included a bracelet to measure skin conductance, a pressure mouse and a pressure pad placed on subjects' chair, see [21]).

To obtain information on how subjects were reasoning and feeling during the study, we asked subjects to verbalize their thoughts and feelings via talk-aloud protocol [22], extended to include affect, as in [23]. The verbal data, along with subjects' eye gaze patterns and interface actions, was recorded via the Tobii system as video files; the EA-Coach logged all interface actions as text files.

3.1 Data Preparation: Coding the Transcripts

To investigate how users' affect and reasoning related to physiological responses, we needed data from our study on both kinds of events. To obtain this data, we first transcribed the video files, including subjects' actions, utterances and time stamps when they occurred. We then devised a coding scheme for identifying in the protocols instances of reasoning (e.g., self-explanation) and affect (e.g., happy).

The *reasoning* portion of the coding scheme (see Table 1, bottom) is based on one from a previous study we ran [17]. We coded utterances as *self-explanation* if subjects expressed a conclusion about a domain-specific principle related to physics [3]. We coded utterances as *analogy* if subjects expressed something about the relation between the problem and example and/or copied from an example (see Table 1 for examples), but did not provide indications of any other kind of reasoning beyond the analogy[4]. Finally, we included an *'other reasoning'* code because we wanted to capture instances when subjects expressed some reasoning, albeit too shallow to be classified as self-explanation, but that did involve more than just a straight comparison of problem/example constants via analogy (see Table 1 for examples). Note that while self-explanation is a highly constructive reasoning activity that correlates with positive learning outcomes (e.g., [9]), reasoning via analogy is associated with a lack of learning [18]; likewise, in our classification, *'other reasoning'* is a less constructive form of reasoning, as compared to *self-explanation*.

The *affect* portion of the coding scheme (see Table 1, top portion) is new and is based on several iterations through the data to solidify the codes. We originally planned on developing fine-grained categories of affect (e.g., 'happy', 'excited',

[3] We did not distinguish between different types of self-explanation (e.g., EBLC-based vs. other) because as a first step, we wanted to analyze in general if and how pupillary response relates to self-explanation.

[4] A simple comparison of problem/example constants is not a self-explanation, as it does not involve a conclusion about a domain-specific principle.

Table 1. Protocol Codes

Affective Codes:			
Code	**#**	**Description**	**Sample Verbalizations**
Positive	68	subject expresses positive affect related to happy or excited state	*" and i got it right and that makes me really happy", "oh that's exciting", "HOORAY", "now I feel good"*
Negative	69	subject expresses feeling negative affect related to frustration	*" now I'm mad", "oh my god this is irritating", "NO!!! not correct ", "Darn it!! "*
Shame	20	subject expresses feeling shame or remorse	*"I really do feel like such an idiot", "I fail ... sorry I took so long"*
Confusion	29	subject expresses confusion	*"I'm feeling confused", " maybe it wants me to draw the horizontal ... I can' understand"*
Reasoning Codes:			
Code	**#**	**Description**	**Sample Verbalizations**
Self-explanation	39	subject explains or clarifies a physics-related concept	*'since it is accelerating I know all the forces added together don't equal zero", "it would be zero because it is ... there is no x component"*
Analogy/ Copy	180	subject draws a comparison between problem and example and/or copies but provides no additional inference/reasoning	*"and their a is acceleration of block which is my mouse", "mag of the normal force... so this is e_y on mine"*
Other Reasoning	106	subject expresses some shallow reasoning that is not a self-explanation or pure analogy	*" well in this picture it is pulling it horizontally and then... 90 plus 40 ... 130?"*

'angry'). However, while subjects would sometimes clearly express a particular type of affect (e.g., "I feel happy" or "I'm irritated"), they would also at times express affect through a single phrase like "NO!!" or "HOORAY!". While in the latter case, the general direction of the affect, i.e,. positive or negative, was clear from the tone and the term used (e.g., "NO!" used to express negative affect), it was more difficult to unambiguously identify the precise emotion expressed. Therefore we broadened the affective categories so that *positive* codes included instances when subjects indicated feeling excited, happy, or generally good (see Table 1 for examples). The *negative* codes included instances when subjects explicitly expressed irritation or frustration, and/or expressed a negative utterance like "darn it!" that related to frustration (see Table 1 for examples).

The coding scheme described above was applied by the first author to classify the data in the verbal protocols, returning to the video files as needed. Overall, 186 instances of *affect* codes and 325 instances of *reasoning* codes were identified (see Table 1).

3.2 Results

As mentioned above, here we focus on data coming from the eye tracker, and in particular, on pupillary response. Given that there tends to be variability among subjects in terms of baseline pupil size, we used Z-scores to normalize

pupil sizes among participants (i.e., *normalized pupil value = (original pupil value - mean pupil size) / standard deviation*, as in [8]). We then associated each coded utterance in the transcripts with the normalized eye tracker data and the EA-Coach logs by standardizing the time stamps in the three sources of data (transcript files, EA-Coach logs, eye tracker logs).

To analyze the data, we originally intended to rely on repeated-measures analysis of variance and/or paired t-tests as appropriate, i.e., depending on the number of levels of the independent variable in question (method A, *within-subjects* analysis). An alternative technique involves using one-way ANOVA (method B, *between-subjects* analysis). Each approach suffers from a limitation. Method A can suffer from data sparseness, since not all subjects necessarily express all types of affect and/or reasoning. This reduces the sample size thereby decreasing power and increasing the chance of a type 2 error (i.e., failing to find an effect when one does in fact exist). The alternative is to use method B, as in for instance [8]. However, the set of data points associated with a given code are not independent, which increases the chance of a type 1 error (i.e., finding an effect when there in fact is none) if method B is used. Given these considerations, we decided to conduct both types of analyses, to triangulate across findings. We will now present the results, starting with findings pertaining to affect.

Results on Affect. To investigate the relationship between pupillary response and affect, we calculated the mean pupil size during the time period a subject expressed an affective response of the type we identified (see Table 1, top). We considered a five second time span, starting at the point when the utterance began (this threshold is similar to that used in related work, e.g., [24]).

We begin with the results from the within-subjects analysis. As anticipated, we found that each subject did not express every type of affective response, leaving missing data entries. When we included the *confusion* or *shame* affective codes in the analysis, we were left with only six subjects that expressed all four types of affect we identified in our analysis. Therefore, we decided to conduct the analysis on the *positive* and *negative* instances of affect only, since this was the only combination that left us with more than six data points. This analysis involved ten students; for each student, we calculated the mean pupil size associated with *positive* and *negative* events, respectively. A paired-samples t-test showed that *affect* had a significant effect on pupillary response($t(9)=2.294$, $p = 0.047$): on average, pupil size was smaller when subjects expressed negative affect, as compared to positive affect (0.0208 vs. 0.3876, respectively).

Recall that the EA-Coach provides immediate feedback for correctness by coloring subjects' entries red or green in the interface. Many of our subjects' affective responses related to entries they generated in the EA-Coach interface, and in particular were responses to an entry being correct or incorrect. Consequently, we wanted to investigate whether entry correctness (or lack of) was driving the affective results. To do so, we compared the mean pupil size five seconds after correct and incorrect entries. We did not find a significant impact of correctness (i.e., correct vs. incorrect entries) on pupillary response ($t(14)=0.508$, $p=0.620$).

As far as the between-subjects analysis is concerned, the ANOVA revealed a significant main effect of *affect* on pupillary response ($F(3,182) = 4.057$, $p = 0.008$). We then conducted Bonferroni post hoc pairwise comparisons to identify which affective responses differed significantly from one another. The only comparison that revealed a significant difference corresponded to the pair *positive-negative* affect ($p=0.006$), where mean pupil size was smaller for *negative* than *positive* (-0.0913 vs. 0.3214), thereby confirming the within-subjects analysis.

Results on Reasoning. To investigate the relationship between pupillary response and how subjects reasoned during the study, we calculated the mean pupil size during the time period a subject engaged in one of the three types of reasoning we identified in the transcripts (*self-explanation, analogy, 'other reasoning'*, see Table 1, bottom). For this analysis, we considered a 15 second time span, starting at the point when the utterance began (this threshold was found to disambiguate self-explanation and lack of in [8]).

We begin with the within-subjects results. As was the case with the affective data, each subject did not express each type of reasoning. Nine subjects did express all three types; for each student, we calculated the mean pupil size for each type of reasoning (*self-explanation*, *analogy* and *'other reasoning'* events). Since the *reasoning* variable has three levels, we conducted a repeated measures analysis of variance. The results revealed a significant main effect of *reasoning* on pupillary response ($F(2,8)=3.63$, $p=0.047$). Given that post-hoc tests are not recommended for within subjects analysis, we followed the method proposed in [25] and conducted pairwise comparisons to identify how the three types of *reasoning* varied from one another. We found that pupil size was significantly bigger for *self-explanation* than *'other reasoning'* (0.4074 vs. -0.0661, respectively; $t(9)= -2.382$, $p=0.04$). We also found that pupil size was bigger for *self-explanation* than *analogy*, but this did not reach significance (0.4074 vs. -.0210, respectively; $t(9)=1.744$, $p=0.115$). The difference between *'other reasoning'* than *analogy* was not significant (-0.0661 vs -0.0210, respectively, $t(9)=0.395$, $p=0.702$).

As was the case with the affect-related analysis, we wanted to investigate if our results were driven by subjects' problem-solving entries, and in particular the correctness (or lack thereof) of these entries. For this analysis, we also considered a 15 second window both prior to and following correct entries, and used paired samples t-tests to investigate differences in response between these two variables. We did not find a significant impact of correctness (or lack of) on pupillary response for either window (before, after).

As far as the between-subjects analysis is concerned, the ANOVA revealed a significant main effect of *reasoning* on pupillary response ($F(2, 322) = 6.454$, $p = 0.002$). We then conducted Bonferroni post hoc pairwise comparisons to identify which types of reasoning responses differed significantly from one another. These results showed that on average, (1) pupil size was significantly bigger for *self-explanation* than *'other reasoning'* (0.2311 vs. -0.0876, respectively; $p=0.008$) and (2) pupil size was significantly bigger for *analogy* than *'other reasoning'* (0.1195 vs. -0.0876, respectively; $p=0.006$). There was no significant difference in pupil size between *self-explanation* and *analogy*.

4 Discussion and Future Work

Our results show that pupillary response is a promising non-invasive avenue for increasing user model bandwidth. As far as affect is concerned, both the within and between subject analysis confirmed that subjects had significantly larger pupil size when they expressed positive affect, as compared to when they expressed negative affect. In contrast to tightly controlled experiments, our subjects were not induced to express affect, but rather expressed it as a natural by-product of the interaction with the EA-Coach. Their affective responses influenced pupil size, information that a user model could take into account when assessing affect, thereby allowing an adaptive application to tailor the interaction to a user's needs. Given that work in psychology shows pupil size increases for affective responses (e.g., [13]), our results indicate that subjects in our experiment experienced positive affect such as excitement more strongly than negative affect related to frustration. The context of our experiment, i.e., a pedagogical one, however, may have influenced particular affective responses, and so more investigation is needed to see how other, non-educational contexts impact pupillary responses. Another area in need of further research pertains to measuring affect. We found that talk-aloud protocol was not suited for performing fine-grained distinctions between affective states. In general, how to measure affect is a key challenge that is the subject of much research (e.g., see [26] for a review), but to date there is a lack of complete understanding related to this issue.

Our study also found support for the fact that how subjects reason impacts pupillary response. As we pointed out earlier, larger pupillary response has been associated with mental effort in tightly controlled experiments. We compared three types of reasoning: (1) *self-explanation*, a highly constructive reasoning activity, against (2) *analogy*, which included comparison of problem/example constants and/or copying from examples and which are not constructive activities, against (3) *other reasoning*. Since self-explanation is a more constructive type of reasoning than the other two, it should result in larger pupil size (as was for instance suggested in [8]). Both kinds of analyses we conducted did indeed confirm that self-explanation resulted in significantly larger pupil size than 'other reasoning'. However, we did not find a significant difference in pupil size between self-explanation and analogy episodes. In fact, our between-subjects analysis showed that analogy resulted in larger pupil size than 'other reasoning', something we did not expect, although this result was not confirmed by the within-subjects analysis. One reason why neither analysis found a difference between analogy and self-explanation is that analogy may actually require mental effort, *despite* the fact that it is a shallow reasoning style. We saw instances in the verbal protocols where subjects struggled aligning the problem/example constants (e.g., *"p underscore y... plus ... plus p [long pause] p is what p [another pause] applied by child applied force"* - a subject trying to substitute example-constant 'p' with one appropriate to her problem). These difficulties may have increased mental effort and thus pupil size. Our results suggest that the model may need additional information to disambiguate self-explanation and analogical reasoning. One way to do so could involve having the model analyze attention patterns

in the interface: since analogy requires the comparison of problem/example constants, but self-explanation does not, including gaze pattern information could disambiguate self-explanation from analogy.

As our next steps, we plan to conduct additional analysis related to investigating further the difference between positive and negative affect, and identifying the mitigating factors driving this difference. Another relevant avenue of investigation relates to exploring the interaction between affect and cognition. There is evidence that subjects process information better when they in a positive affective state [27], and so it would be interesting to analyze if and how this occurred in our study. We also plan to analyze other aspects of data provided by the eye tracker (fixations and saccadic eye movements) to explore how they may inform a user model. We plan to rely on our findings both from this experiment and subsequent analysis to extend the EA-Coach user model to take into account eye-tracker information, and design affect and additional meta-cognitive support based on the revised model. We will subsequently evaluate how this support impacts the tutor's pedagogical effectiveness and usability.

References

1. VanLehn, K.: Student modeling. Foundations of Intelligent Tutoring Systems, 55–78 (1988)
2. Klein, J., Moon, Y.: This computer responds to user frustration: Theory, design, results, and implications. Interacting with Computers 14, 119–140 (2000)
3. Aleven, V., Koedinger, R.: An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. Cognitive Science 26(2), 147–179 (2002)
4. D'Mello, S.K., Picard, R.W., Graesser, A.C.: Towards an affect-sensitive autotutor. IEEE Intelligent Systems 22(4), 53–61 (2007)
5. Burleson, W.: Affective Learning Companions: Strategies for Empathetic Agents with Real-Time Multimodal Affective Sensing to Foster Meta-Cognitive Approaches to Learning, Motivation, and Perseverance. Ph.D thesis, MIT (2006)
6. Gluck, K., Anderson, J.: Cognitive architectures play in intelligent tutoring systems? In: Cognition and Instruction: Twenty-Five Years of Progress, pp. 227–262 (2001)
7. Qu, L., Johnson, L.: Detecting the learner's motivational states in an interactive learning environment. In: 12th International Conference on Artificial Intelligence in Education, pp. 547–554 (2005)
8. Conati, C., Merten, C.: Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. Knowledge Based Systems 20(6), 557–574 (2007)
9. Chi, M., Basssok, M., Lewis, M., Reimann, P., Glaser, R.: Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science 13, 145–182 (1989)
10. Marshall, S.P.: Identifying cognitive state from eye metrics. Aviation, Space, and Environmental Medicine 78, 165–175 (2007)
11. Van Gerven, P.W.M., Paas, F., Van Merrinboer, J.J.G., Schmidt, H.G.: Memory load and the cognitive pupillary response in aging. Psychophysiology 41(2), 167–174 (2004)

12. Vo, M.L.H., Jacobs, A.M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., Hutzler, F.: The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. Psychophysiology 45(1), 130–140 (2008)
13. Partala, T., Surakka, V.: Pupil size variation as an indication of affective processing. Int. Journal of Human-Computer Studies 59(1-2), 185–198 (2003)
14. Iqbal, S., Zheng, X., Bailey, B.P.: Task-evoked pupillary response to mental workload in human-computer interaction. In: CHI 2004 extended abstracts on Human factors in computing systems, pp. 1477–1480 (2004)
15. Schultheis, H., Jameson, A.: Load in adaptive hypermedia systems: Physiological and behavioral methods. In: Adaptive hypermedia. Interacting with Computers, pp. 225–234 (2004)
16. Conati, C., Muldner, K., Carenini, G.: From example studying to problem solving via tailored computer-based meta-cognitive scaffolding: Hypotheses and design. Technology, Instruction, Cognition and Learning (TICL) 4(2), 139–190 (2006)
17. Muldner, K., Conati, C.: Evaluating a decision-theoretic approach to tailored example selection. In: IJCAI 2007, 20th International Joint Conference in Artificial Intelligence, pp. 483–488 (2007)
18. VanLehn, K.: Analogy events: How examples are used during problem solving. Cognitive Science 22(3), 347–388 (1998)
19. VanLehn, K.: Rule-learning events in the acquisition of a complex skill: An evaluation of cascade. The Journal of the Learning Sciences 1(8), 71–125 (1999)
20. Muldner, K.: Tailored Support for Analogical Problem Solving. Ph.D thesis, University of British Columbia (2007)
21. Dragon, T., Arroyo, I., Woolf, B.P., Burleson, W., el Kaliouby, R., Eydgahi, H.: Viewing Student Affect and Learning through Classroom Observation and Physical Sensors. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 29–39. Springer, Heidelberg (2008)
22. Ercisson, K., Simmon, H.: Verbal reports as data. Psychological Review 87(3), 215–250 (1980)
23. Craig, S., 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. Cognition and Emotion 22(5), 777–788 (2008)
24. Van Gerven, P., Paas, F., Van Merrienboer, J., Schmidt, H.: Memory load and the cognitive pupillary response in aging. Psychophysiology 41(2), 167–174 (2001)
25. Cardinal, R., Aitken, M.: ANOVA for the Behavioural Sciences Researcher. Routledge, London (2006)
26. Mauss, I., Robinson, M.: Measures of emotion: A review. Cognition & Emotion 23(2), 209–237 (in press)
27. Levens, S., Phelps, E.: Emotion processing effects on interference resolution in working memory. Emotion 8(2), 267–280 (2008)