Metacognition and Learning (2023) 18:465–494
https://doi.org/10.1007/s11409-023-09336-z

A multi-level growth modeling approach to measuring learner attention with metacognitive pedagogical agents

Megan Wiedbusch[1] · James Lester[2] · Roger Azevedo[1]

Received: 6 January 2022 / Accepted: 24 January 2023 / Published online: 3 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Pedagogical agents have been designed to support the significant challenges that learners face when self-regulating in advanced learning environments. Evidence suggests differences in learners' prior skills and abilities, in conjunction with excessive didactic support, can cause overreliance on these external aids, which in turn prevents deeper learning, and pedagogical agents can provide tailored scaffolding to accommodate learners' individual needs. However, there is less evidence about the impact of abstract scaffolding, such as the sharing of non-verbal metacognitive information via a pedagogical agent's facial expressions, on self-regulated learning. To assess factors in the passing of non-verbal metacognitive information via pedagogical agents in a multimedia learning environment, we used growth modeling with self-reports, eye-tracking, and log-file data collected from fifty ($n = 50$) undergraduates at a large North American university as they learned about human body systems while using MetaTutor-IVH, a multimedia learning environment with a pedagogical agent. We controlled for participant characteristics (perceived utility of emotions for self- and other-centered positive and negative emotions) and characteristics of the metacognitive monitoring information provided by a pedagogical agent (expression type and expression congruency) to assess factors in non-verbally communicating metacognitive information. Results suggest that learners attend to pedagogical agents less over time, but this rate of change is weaker when an agent is providing an expression that is congruent with the ground truth of the environment. Further, only the perceived information utility of other-centered negative emotions has a significant effect on this duration, suggesting learners are driven to consult pedagogical agents to avoid embarrassment or shame. We discuss design implications of these findings for technology-based learning environments.

Keywords Pedagogical agents · Metacognition · Affect detection and recognition · Individual differences · Multilevel methods · Science learning

Self-regulation refers to an individual's ability to dynamically monitor and regulate performance through the modification and regulation of subgoals, plans, and strategies to achieve

✉ Megan Wiedbusch
megan.wiedbusch@ucf.edu

1 School of Modeling, Simulation & Training, University of Central Florida, Orlando, FL, USA

2 Department of Computer Science, North Carolina State University, Raleigh, NC, USA

an overall goal (Winne, 2018). Using effective self-regulatory skills is critical when acquiring content knowledge and skills that are transferable to new contexts, including technology-based learning environments (Schunk et al., 2018). During self-regulated learning (SRL), learners must actively monitor and regulate their cognitive and metacognitive processes while selecting, organizing, and integrating different multimedia instructional materials presented in the learning environment (Azevedo & Dever, 2022; Greene & Azevedo, 2007, 2009). SRL has been found to be challenging for most learners and therefore has been supported in technology-based learning environments with the assistance of external aids such as cognitive and metacognitive scaffolds using a variety of methods including pedagogical agents (Azevedo et al., 2007; Greene & Azevedo, 2007; Johnson & Lester, 2016, 2018).

In this study, we address the issue of poor self-regulatory learner behaviors by examining how metacognitive scaffolding that is abstract and non-verbal (i.e., using a pedagogical agent's facial expressions of emotions) may impact the behaviors of learners interacting with a multimedia learning environment. This work highlights novel and dynamic approaches to the way information can be non-didactically communicated to learners and how this information impacts learner metacognitive monitoring over time. Specifically, we examined the way in which various emotional expressions (e.g., confusion) could relay information about the multimedia content a learner has examined to support the evaluation of that content and whether learners attended to non-didactic metacognitive information.

We begin by reviewing previous metacognitive monitoring and emotion-based research grounded in Efklides' (2019) multifaceted and multilevel model of SRL (MASRL). We primarily review studies that utilize eye-tracking and non-verbal facial expression scaffolding. We then introduce the MetaTutor-IVH learning environment and experimental paradigm used during our study, with a short discussion on the assumptions about metacognition and affect during multimedia learning that are made within this environment. Next, we describe our study and modeling approach to capturing multilevel individual differences and trial effects on the duration in which learners attend to the pedagogical agent during the non-verbal communication of metacognitive information. We conclude by reviewing our findings anchored around the Efklides' (2019) MASRL model and previous literature in addition to a discussion and recommendation for the future direction of non-didactic scaffolding approaches.

Self-regulated learning in multimedia learning environments with pedagogical agents and eye-tracking

Including multimedia content (i.e., pictures, graphs, diagrams, text, audio, etc.) has been shown to increase learning outcomes during SRL compared to text alone (Johnson & Lester, 2016, 2018; Mayer, 2022; Mayer & Fiorella, 2022; Schweppe & Rummer, 2016). Learning with multimedia requires the construction and integration of mental models to develop and modify plans, strategies, and learning goals (Azevedo & Dever, 2022; Mayer, 2022). With the increasing availability of eye-tracking technology, eye-tracking has been used to examine and measure underlying cognitive and metacognitive processes from multimedia learning (Alemdag & Cagiltay, 2018; Azevedo & Dever, 2022; Dever et al., 2021; Fiorella & Pilegard, 2021; Mudrick et al., 2019; Stull et al., 2018; Wiedbusch & Azevedo, 2020). Gaze metrics have been used to measure attention allocation on various multimedia elements (e.g., Dever et al., 2021), and can also measure the (re-)evaluation of content (e.g., Wiedbusch & Azevedo, 2020), which ultimately point to the construction and integration of the content into a coherent mental model.

However, complex learning with multimedia is difficult for learners to self-regulate, and empirical evidence shows that individuals with these skills and abilities, when supported via scaffolding techniques by pedagogical agents, outperform their peers that receive no assistance (Azevedo & Cromley, 2004; Azevedo & Dever, 2022; Azevedo et al., 2019; Bannert et al., 2009; De Boer et al., 2012; Dignath & Buttner, 2008; Jansen et al., 2019; Kramarski, 2018). Pedagogical agents act as simulated scaffolding support systems taking on various roles as teachers, tutors, peers, etc., during the learning process and can introduce social, motivational, and affective benefits to otherwise sterile learning environments (Johnson & Lester, 2016, 2018; Sinatra et al., 2021). Specifically, they can monitor and measure learners as they progress through their learning to offer tailored scaffolding through various methods such as hints, feedback, interactive demonstrations, and even customizing new problems based on the learner's current misconceptions or learning incompetencies (Johnson & Lester, 2018).

Despite the benefits of this support, there is a balance that needs to be established to foster and promote developing SRL skills without being used as the sole method of SRL skill use. That is, learners who are given too much external support rely heavily on pedagogical agents and thus develop minimal SRL skills (Boekaerts, 1999). The amount of scaffolding needed and levels of pre-existing SRL abilities differs between learners (Aleven et al., 2016). Therefore, pedagogical agents need to be adaptive based on the different aptitudes and characteristics of individual learners (Johnson & Lester, 2018).

Learning environments and technologies should be individualized based on these differences, but there are many underlying theoretical considerations that still need further exploration. For example, can essential hints and crucial information be provided to learners in non-traditional manners, such as through non-verbal clues, that help avoid over-directing the learner into overreliance on scaffolding? Previous research has shown that social reactions can be triggered by body movements that encourage focus and engagement with learning content (Krämer & Bente, 2010), but it was unclear if this had any direct impact on learning. A recent meta-analysis on the benefits of pedagogical agents (Castro-Alonso et al., 2021) showed an effect size of $g+=0.20$ on learning with various factors influencing learning-benefits such as agent gender and appearance. When agents used non-verbal communications, there were similar effect sizes when they used eye gaze ($g+=0.26$) and gesturing ($g+=0.23$), and a moderately significant effect of either static facial expressions or dynamic expressions ($g+=0.42$). However, these studies focused on learning outcomes and the role of social factors, and the agents did not provide metacognitive information about the relevancy of the multimedia content in relation to the learning goal. More research is needed to understand the role of these design features on metacognition and SRL, especially as it relates to the use of facial expressions of emotions exhibited by pedagogical agents in supporting learners' metacognitive monitoring and judgements.

Metacognitive monitoring

Metacognition is commonly described as thinking about one's thinking (Flavell, 1979). This monitoring and reasoning about an individual's cognition, however, encompasses multiple facets, including memory, emotion, motivation, and knowledge to help inform and regulate this phenomenon. Multiple empirical studies about metacognition suggest that there are multiple types of cues that can help learners monitor and regulate their cognition such as theory-based cues (e.g., beliefs about their own expertise), or experience-based cues (e.g., how easy a comparable problem was for them to solve; Jaeger & Wiley, 2014; Koriat, 1997; Serra & Dunlosky, 2005). Additionally, there is growing evidence that metacognition, and specifically learning, cannot be viewed purely as an individual process but should also account for various social and cultural aspects (Efklides, 2008; Lehtinen, 2003; Vauras et al., 2008). Efklides' (2019) multifaceted and multilevel model of metacognition suggest that there is a meta-meta-level of learning that contains judgements about one's own and other's metacognitive experiences, knowledge, and skills, representing a social level of metacognition. This can include multiple components that help interpret and encode these values such as judgments of performance, semantic understanding of conversation, and emotions (Efklides, 2019). Efklides (2019) proposed a model in which self-regulated learning is a dynamic process that involves both metacognitive, affective, and motivational components. Specifically, the Metacognitive and Affective model of SRL (MASRL) places affect within both the person level (tying together one's self-concept and motivation) as well as in the 'Task x Person' level (tying together cognition and affect/effort self-regulation; see Fig. 1).

In this way, this model theoretically explains how 'Task x Person' level processing is guided by 'Person' level driven decisions, but can also be revaluated and overridden through task and cognitive monitoring. For example, metacognition and affect, i.e., a person's subjective experiences, can influence how much effort or attention to invest. However, this effort or attention allocation can be overridden if as the learner monitors their cognition finds that too much (or too little) attention and effort are being used given their current perception of performance. We thus chose to anchor our study in Efklides' (2019) MASRL model because it captures the bidirectional influence of individual data-driven differences that shape (and conversely are shaped by) self-regulatory behavior unlike other models of SRL. It is also one of the few models that provides more than just a cursory nod to the role of affect and motivation within SRL.

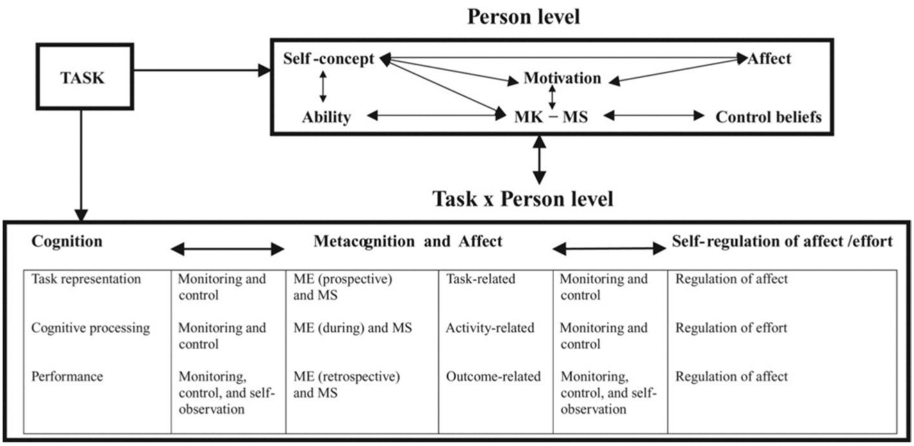

Fig. 1 Efklides (2011) MARSL Model. Permissions have been requested from Taylor & Francis (ref:_00D0 Y35Iji._5007THoJ3Q:ref)

Emotions

Academic-related emotions (Pekrun, 2016) are pervasive in school settings and can be achievement-related (e.g., enjoyment of learning, pride, boredom) as well as social in nature (e.g., admiration, envy, contempt). These emotions function as factors for student motivation and often play an important role in student's ability to self-regulate (Efklides, 2011; Pekrun, 2016; Pekrun et al., 2002). Specifically, emotions can serve as both antecedents and reactions to selected information processing strategies that influence how effectively a learner will cognitively and metacognitively comprehend presented information (D'Mello & Graesser, 2012; Pekrun & Linnenbrink-Garcia, 2012a, b; Taubet al., 2018a, b). That is, as learners engage with and metacognitively monitor their learning with instructional materials, they may currently be experiencing confusion after identifying inconsistencies between prior knowledge and new information that then leads to behaviors to alleviate that confusion.

Emotions also serve as reactions to cognitive processes, such as feeling frustrated if that new strategy proved unfruitful in their learning goals as detected by their metacognitive monitoring. This cyclic and recursive nature has made academic emotions an increasingly important focus for the study of self-regulated learning and academic achievement (Ahmed et al., 2012; Efklides, 2005; Linnenbrink, 2006; Linnenbrink-Garcia & Pekrun, 2011; Schutz & Lanehart, 2002; Schutz & Pekrun, 2007; Schutz & Zembylas, 2009).

Emotions have been conceptualized and classified along many dimensions when discussing them within academic contexts. One such distinction is made about valence, which contrasts positive emotions (i.e., pleasant emotions such as joy, pride, cheerful, etc.) and negative emotions (i.e., unpleasant emotions such as jealousy, embarrassment, anger, etc.; Pekrun, 2006). Studies have found that positive and negative emotions have both beneficial and detrimental effects on learning (Cloude et al., 2020; D'Mello & Graesser, 2012; Forgas, 2002; Pekrun & Linnenbrink-Garcia, 2012a, b).

For example, learners will choose to participate in more experiences they believe will induce negative emotions (e.g., fear) if they find present or future utility in the experience (Tamir & Ford, 2012). This perceived utility of emotion has been defined as an emotion's worth for accomplishing one's goals, which has been theorized to play a precursory role in emotion regulation and affect and considered a moderately fixed individual difference variable while controlling for the actual emotional experience (Chow, 2018; Chow, Berenbaum, & Flores Jr., 2013; Chow et al., 2015). Research has shown that a learner who finds the emotion of appreciation useful also is likely to perceive others as benevolent, feel appreciation more often and strongly, and demonstrate prosocial behaviors (Chow, Berenbaum, & Flores Jr., 2013).

Perceived emotion utility can be further conceptualized as either self-serving emotions (i.e., serving the needs of the individual who is experiencing the emotion) or other-serving emotions (i.e., serving the needs of others). For example, pride (positive) and jealousy (negative) are both self-serving emotions while appreciation (positive) and embarrassment (negative) are other-serving emotions. While the perception of emotion utility has not been examined within the context of its influence on SRL, some evidence has shown self-reflection of one's affective experiences during cognitive processing can lead to causal attributions (Metallidou & Efklides, 2001). It is therefore not difficult to imagine that reflection about how much utility various emotions carry in context may also lead to causal attributions (e.g., I think feeling pride when I am praised for doing good work provides me information that my tutor sees me working hard. Therefore, I take time and effort to look for indications of praise from them and increase or decrease my learning strategies to elicit that response).

Purpose of the study

In this study we focused on examining the individual differences of perceived utility of emotions on how learners behave (i.e., allocate attention) towards pedagogical agents providing metacognitive information nonverbally while learning with MetaTutor-IVH. For our study, participants learned about various human body systems (e.g., respiratory system, circulatory system, etc.) with MetaTutor-IVH. MetaTutor-Intelligent Virtual Human (IVH) is a self-paced and linearly structured multimedia environment, which was specifically designed to study how learning and metacognitive were influenced by a pedagogical agent (Azevedo et al., 2019). In this current construction of MetaTutor-IVH, the pedagogical agent is not intelligent, but future development will use incoming learner physiological and behavioral responses (e.g., user facial expression, self-report measures, eye-tracking, etc.) to intelligently adapt its responses (e.g., providing individualized adaptive verbal metacognitive scaffolding while coordinating congruent facial expressions). The current iteration seeks to study how users may be influenced or interact with the agent without knowing that it is not intelligent. Within this environment, learners are presented multimedia content about the human digestive, nervous, endocrine, circulatory, lymphatic, muscle, integumentary, urinary, and respiratory systems, and asked to answer a multiple-choice question, while being prompted to provide multiple metacognitive judgments throughout their learning (see MetaTutor-IVH subsection in Methods for explanation about the experimental paradigm and environment).

While much research has investigated the role of pedagogical agents in learning technologies, we are unaware of any work that has explicitly examined the role of non-verbal information being provided through facial expressions about metacognitive monitoring judgements, or the potential factors that could influence the use, disuse, or misuse of this information. To address this gap, we investigate the role of metacognition, affect, and SRL using the MASRL model. This leads us to the following research question:

> *How do learner SRL behaviors change when controlling for 'Person-level' characteristics and 'Task x Person level' characteristics?*

The SRL behaviors or interest include how long a learner examines a pedagogical agent providing metacognitive monitoring information. We classify an individual's perceived utility of emotions for self and other centered positive emotions as Person-level characteristics. Additionally, we classify metacognitive monitoring information provided by the agent via expression type and congruency as Task x Person level characteristics. In other words, our main research question driving the model building of this work asks,

We hypothesized that there will be a change in duration looking at the agent providing metacognitive information over time spent in MetaTutor-IVH. Additionally, we predict there will be an effect of duration accounted for by multiple perceived utility of both self- and other-centered emotions (both positive and negative) as well as effects from both the expression type and congruency of the trials. Our hypothesis is motivated by the MASRL model's theoretical implications for the dynamical unfolding of SRL as informed by metacognition and affect. Specifically, because this model highlights both the influence of 'Person' level features 'Person x Task' level features on SRL, we expect both the participant (i.e., 'Person') level characteristics and trial-specific (i.e., 'Person x Task') characteristics to influence participant behaviors. In our study, these behaviors refer to the duration in which a participant examines the agent during their judgment (via facial expression). If a participant fails to spend an adequate amount of time looking at the agent while they provide their judgment, the learner will not be able to interpret and incorporate this new information into their metacognitive knowledge or current working mental model, suggesting they do not wish to allocate effort to doing so. However, we would also expect the overall time needed to interpret the non-verbal information to decrease as learners become more efficient at gleaming the information the agent's scaffolding provides. Our work will provide empirical evidence for one of the major assumptions of the MASRL model that states 'Person' level features can influence 'Task-Person' level features and vice versa.

Methods

Participants

Sixty undergraduate students from a large North American university were recruited to learn about human body systems using MetaTutor-IVH. Ten participants were removed for incomplete eye-tracking data. Of the remaining 50 participants, 59% were female, and ages ranged from 18 to 29 ($M = 20.31$, $SD = 2.60$). All participants were monetarily compensated \$10 an hour up to \$30 and the study was approved by an IRB prior to all data collection.

After providing consent, participants sat in front of a computer and were calibrated to a SMI RED 250 eye-tracker and Shimmer 3 + wireless bracelet (used to capture skin conductance activity but not reported in this manuscript, see Fig. 2). The eye-tracker used a sampling rate of 60 Hz (set lower for integration with iMotions Attention Tool). The 9-point calibration resulted in a gaze position accuracy error of about 0.4 visual degrees and a spatial resolution of 0.03°.

All participants were required to first complete a demographic questionnaire, an 18-item multiple choice biology content pretest, and multiple questionnaires assessing emotions and motivation (i.e., Achievement Emotions Questionnaire [AEQ (Pekrun et al., 2011)], Emotion Regulation Questionnaire [ERQ (Gross & John, 2003)], Perceived Affect Utility Scale [PAUSe (Chow & Berenbaum, 2012)] prior to interacting with MetaTutor-IVH. For this study's analysis we only use the self-reports from the PAUSe questionnaire and thus do not further explain the other questionnaires to maintain brevity. During learning with MetaTutor-IVH, participants' actions, eye-tracking, skin conductance, and self-reported metacognitive judgments were all collected, however only eye-tracking and self-reported metacognitive judgments were used in this analysis.

Participants were instructed that they were about to learn about different body systems by reading text, inspecting diagrams, and answering questions about the content. A short video introduced the environment to the participants, explaining the general experimental paradigm and what they could expect on each trial. Participants were informed the pedagogical agent would provide a metacognitive judgment of its own about how relevant it felt the text and diagram were but were not told the agent was not always accurate. We purposefully did not tell the learners about the agent accuracy to force learners to have to evaluate the agent's reaction in comparison to their own judgements. If the agent was always accurate, and the participants informed of such, they would no longer have to evaluate

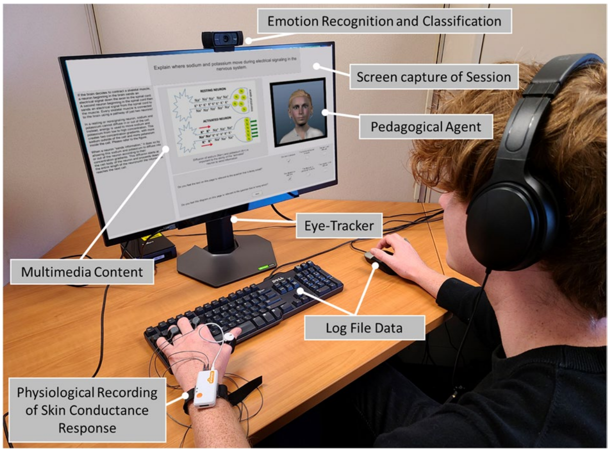

Fig. 2 Experimental setup depicting an instrumented participant

the presented information until they saw the agent's judgement. Instead, we designed the agent to provide additional information to the student to also consider. This serves a similar purpose that non-verbal feedback from a tutor or peer. Once the learning session was completed, participants completed the questionnaires once more (not used in this study), were debriefed about the agents' behavior and study purpose, compensated, and thanked for their participation.

Perceived affect utility scale (PAUSe)

This self-report questionnaire measures three facets of the degree of emotional utility of self-serving and other-serving emotions for (1) providing information, (2) motivating, and (3) fostering goal-directed behaviors (Chow & Berenbaum, 2012). Participants are first asked to think about "*things [they] generally seek to accomplish in everyday life, or the things [they] typically try to do. Some examples of goals are 'getting along with others', 'trying to be the center of attention', 'trying to help others', and 'trying to do what is best for myself'*" (Chow & Berenbaum, 2012). They are then asked to rate the extent to which they agree with three sets of questions each populated by a list of specific emotions on a 1 (strongly disagree) to 5 (strongly agree) scale. The first set of questions stated, *'Feeling [emotion] lets me know how well or poorly I'm doing in achieving my goals'*, to measure informational utility. The second set of questions, focused on motivational utility, stated, *'Feeling [emotion] motivates me to achieve my goals'*. Finally, the third set of questions, measuring goal-directed behavior facilitation utility, stated, *'Feeling [emotion] makes it easier for me to do things that will help me to achieve my goals'*. Participants then received an average across emotions aggregated by valence (i.e., positive or negative) and self/other-serving for each of these sets of questions (See Fig. 3). Simply, for each

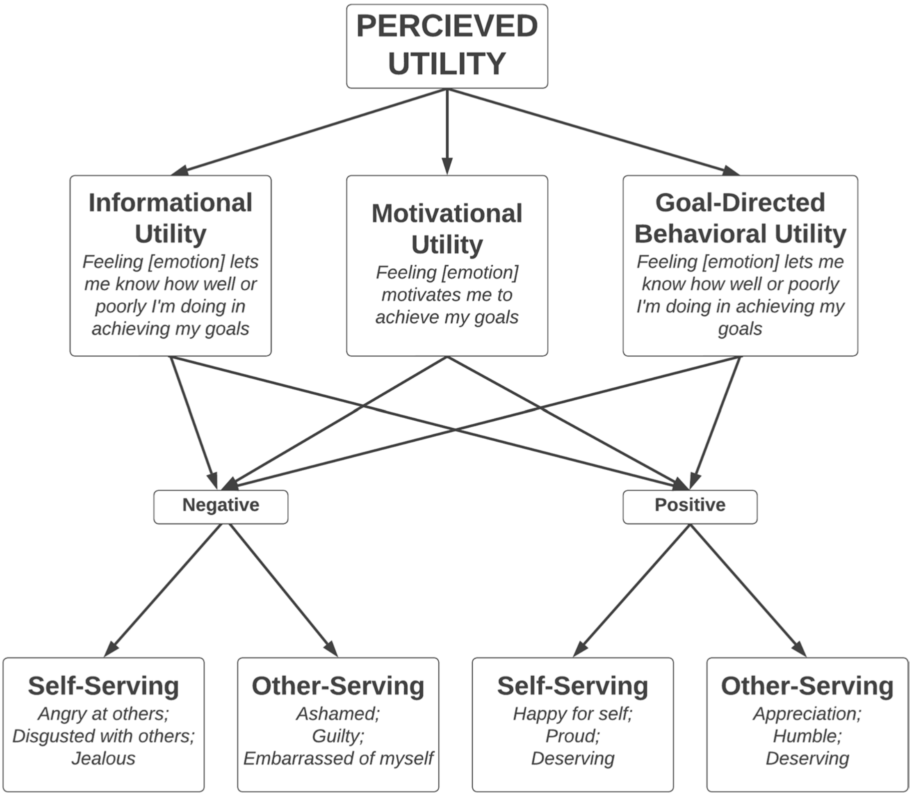

Fig. 3 A graphical diagram showing the twelve intersections of PAUSe measured attributes (Chow & Berenbaum, 2012). Perceived utility has three major branches, each with Self-and Other-serving emotions that are further categorized as either negative or positive

of the three facets (informational utility; motivational utility; and goal-directed behavior facilitation), participants had an average rating for (1) Self-serving negative emotions (angry at others, disgusted with others, jealous), (2) Self-serving positive emotions (happy for self, proud, deserving), (3) Other-serving negative emotions (ashamed of myself, guilty, embarrassed of myself), and (4) Other-serving positive emotions (appreciation, humble, respectful).

MetaTutor-IVH

Participants learn about the various human biological systems by completing 18 self-paced linear-sequenced experimental trials using MetaTutor-IVH (see Fig. 4). MetaTutor-IVH has 18 trials in a within-subjects design in which trials are defined in a 3 (*content relevancy*: fully relevant, text somewhat relevant, diagram some-what relevant) × 3 (*agent facial expression*: congruent, incongruent, neutral) × 2 (*question type*: function, malfunction) manner.

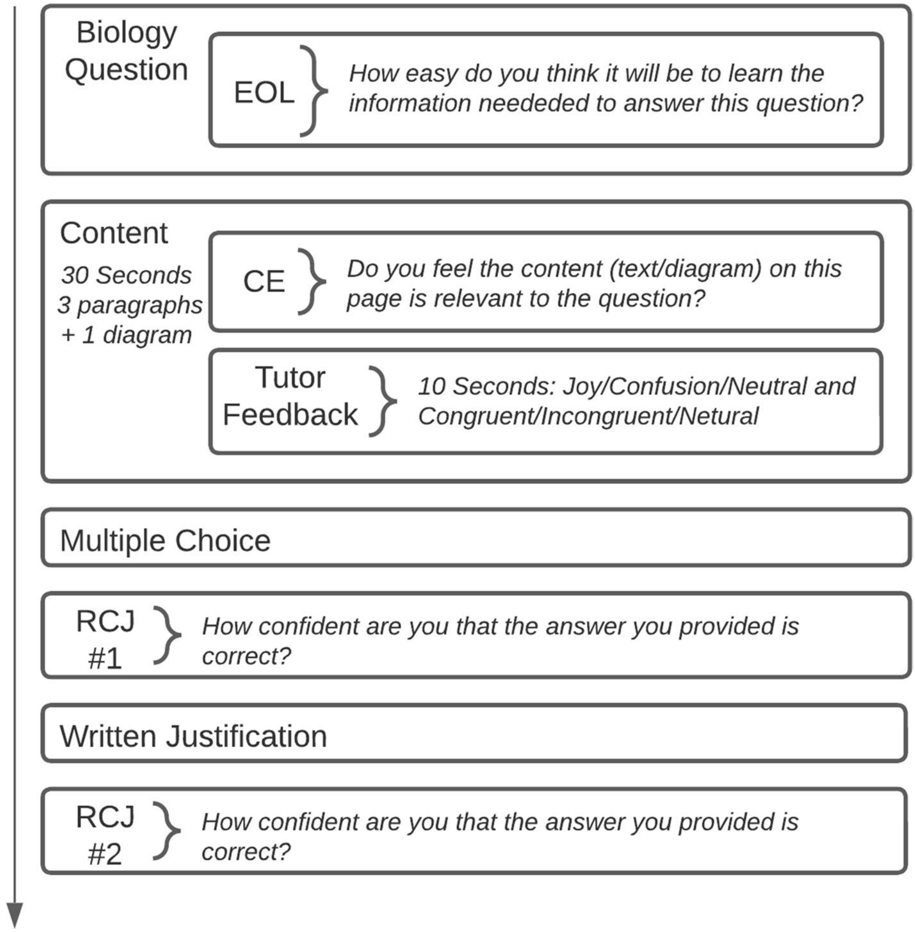

Fig. 4 A Typical MetaTutor-IVH Experimental Trial

Content relevancy describes the source of the pertinent information for the posed question. All slides had information that answered the multiple-choice question, but the location of this information varied. Fully relevant content meant that relevant information could be found in both the text and diagram, while text-somewhat relevant meant the pertinent information was located in the diagram and diagram-somewhat relevant meant the pertinent information was located in the text. Previous research has shown that the metacognitive selection of more relevant content is associated with higher performance (Dunlosky & Lipko, 2007).

Trials also varied by the facial expression of the agent during their given judgment. This expression could be congruent (i.e., showing confusion when either text/diagram were only somewhat relevant or joy when text/diagram were fully relevant), incongruent (i.e., showing joy when either text/diagram were only somewhat relevant or confusion when text/diagram were fully relevant), or neutral (showing no expression; see Fig. 5). This feature

Fig. 5 The agent's non-verbal emotion expressions are provided as a meta-cognitive monitoring judgment about content relevance displayed to participants every trial. From left to right: Joy, confusion, and neutral

was chosen explicitly to examine if non-verbal metacognitive information would impact a learner's metacognitive strategy use or behaviors while learning with an artificial peer. By varying the type of expression, we would be able to further examine if positive or negative emotions impacted this human-agent interaction.

Finally, trials varied by the question type (function versus malfunction), where each body system was represented in 2 trials and asked about the body functions when they were working correctly (function) and incorrectly (malfunction). This factor was included in the design of the environment to examine the impact of question wording and content might impact the interaction between the human and the agent. However, as we found no significant difference in the time spent examining the agent between function and malfunction trials ($t(824) = 0.522$, $p = 0.6$) we chose not to include this factor in our analysis.

The MetaTutor-IVH environment first provides a science question that learners will have to answer at the end of the trial. They are then asked to provide an ease-of-learning metacognitive judgment by answering the question '*How easy do you think it will be to learn the information needed to answer this question?*' on a 1-unit incremental scale from 0 to 100.

Ease of learning judgments must be made prior to any content, and therefore made through inferences guided by prior knowledge, contextual information, and general domain knowledge (Nelson & Narens, 1990; Jemstedt, Kubik, & Jonsson, 2017). They guide a learners' attention and effort allocation but have been found to be poor or sometimes only moderate predictors of learning (Leonesio & Nelson, 1990; McCarley & Gosney, 2005; Son & Metcalfe, 2000; Jonsson & Kerimi, 2011). We argue these judgements, contextualized within the MASRL model, occur in the "Person" level as it relies heavily on metacognitive knowledge and skills, and Efklides argues this is the level in which a learner makes decisions about their learning (Efklides, 2019).

Next, MetaTutor-IVH provides a multimedia content slide (see Fig. 6) which comprises of the posed question along the top edge of the slide, text (3 paragraphs, Flesch-Kincaid readability score range: 9.1–12.5; M = 10.5) along the left-hand side, a diagram in the middle of the slide, an agent in the top right-hand corner of the slide, and a judgment response panel along the bottom. Learners have 30 s on

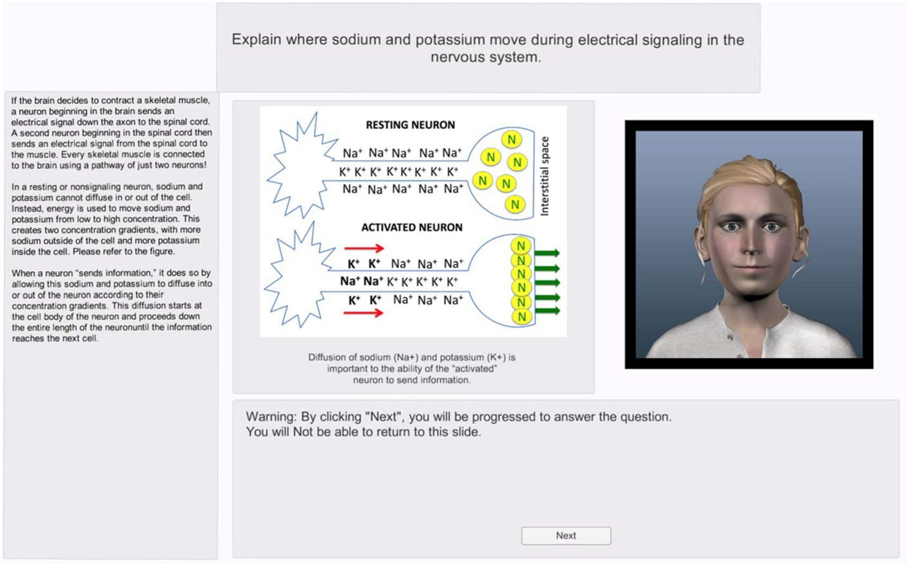

Fig. 6 A prototypical content slide used in MetaTutor-IVH. Note: This one specifically covers the Human Nervous System

this slide before being prompted to provide a content evaluation about how relevant they think the presented information is by answering *'Do you feel the text/diagram on this page is relevant to the question being asked?'* on a 3-point scale (1: text/diagram is relevant; 2: text/diagram is somewhat relevant; 3: text/diagram is not relevant). Content evaluations refer to the monitoring of informational content relative to a learner's goals (Greene & Azevedo, 2009). According to the MASRL model, this evaluative process occurs within the 'Task x Person' level as it is specific to the type and quality of information provided within the task in conjunction with one's evaluation of that relevancy and goals. That is, specific task processing (i.e., the selection, organization, and integration of multimedia content) must occur which is internally monitored and then externalized when responding to the CE question. For example, if a learner recognizes that the diagram is not relevant to the question, they may regulate how they will allocate their additional efforts after providing this CE. Additionally, they must regulate any emotions, such as frustration or confusion, caused by the irrelevant information as they continue their studying.

Once learners submit their judgment, the agent provides a 10 s judgment of how relevant it feels the text and diagram are with a silent emoted expression (i.e., joy, confusion, or neutral; see Fig. 5). This judgment provides non-verbal metacognitive information for the learner to interpret within the context of the task and, according to the MASRL model, we argue occurs within both the 'Person' level and the 'Person-task' level.

We are unaware of any previous work that has used the MASRL model in conjunction with non-verbal pedagogical agents providing metacognitive judgements of their own, and this study aims to directly address this component of the model. We will show how stable 'Person' level elements (perceived utility of emotions for self- and other-centered positive and negative emotions) and varying 'Person-task' level elements (expression type and expression congruency) both contribute to changes in behavior and interactions with pedagogical agents during complex learning. For example, if a learner reported that both the diagram and text were fully relevant, but the agent showed confusion, the learner must (1) evaluate and interpret the agent's reaction and then (2) determine how this reaction will affect their studying and next behavioral steps. This type of SRL requires a learner to integrate not only incoming agent-reaction based information, but also their own metacognitive monitoring of the content, and stable perceptions of emotion of the utility of the agent's emotion during the trial.

Participants could continue to study until ready to answer the question, after which they make a retrospective confidence judgment by answering '*How confident are you that the answer you provided is correct?*' on a 1-unit incremental scale from 50 to 100 (a score of 50 indicated they had simply guessed). Following their confidence judgment, learners justify their response in a text-based free response section followed by another retrospective confidence judgment based on their justification. These types of reflective monitoring evaluations occur within the 'Task x Person' level of the MASRL model and can later help inform a learner's metamemory as well as their regulation of effort and affect in relation to how they perceive their performance on a particular trial. After this final judgment, the next trial begins.

Data

Missing data Ten participants were removed due to incomplete eye-tracking data during the entire experiment. Out of the remaining 50 participants, 73 individual trials (out of 900) were removed as we were unable to parse out if a participant's duration reporting 0 s could be contributed to the fact the participant did not examine the agent on a given trial or if the measurement tool failed to correctly collect data. This means not all participants had 18 trials worth of data; however, our modeling approach is able to handle such cases and thus we chose not to impute missing data.

Data structure A total of 827 trials at the micro level (Level 1) are nested within 50 participants at the macro level (Level 2). Participants' self-reported perceived utility of emotions for self and other-centered positive and negative emotions are located at Level 2, while characteristics of the information provided by the agent (expression type and congruency) are located at Level 1. Variables basic descriptive statistics are provided in Table 1. Please note that we only report on the Level 2 explanatory variables that remained after our dimension reduction methods were applied.

Participant level variables Our Level 2 covariates include participants' self-reported perceived utility of emotions for self and other-centered positive and negative emotions. These include twelve average scores that range from 1 (strongly disagree) to 5 (strongly agree). These twelve variables are distributed among three perceived utilities, positive–negative valence, and self-other centered emotions (See Fig. 3). Simply, for each of the three facets (informational utility; motivational utility; and goal-directed behavior facilitation), participants had an average rating for (1) Self-serving

Table 1 Descriptive Statistics for Model Variables

Variable	M	SD	Min	Max
Dependent variable				
Duration (Seconds; Not transformed)	4.122	2.934	0.251	14.595
Duration (Box-Cox ($\lambda = 0.328$) transformed)	0.000	1.000	-2.200	2.346
Explanatory variables at the participant level				
Perceived Informational Utility of Self-Serving Positive Emotions	3.865	0.659	2.000	5.000
Perceived Informational Utility of Other-Serving Negative Emotions	3.593	1.165	1.000	5.000
Perceived Informational Utility of Other-Serving Positive Emotions	3.952	0.815	2.000	5.000
Perceived Motivational Utility of Other-Serving Negative Emotions	3.590	1.100	1.000	5.000
Perceived Behavior Facilitation Utility of Other-Serving Negative Emotions	3.516	1.024	1.333	5.000
Explanatory variables at the Environment/Trial level				
Expression Type	1.948	0.806	1.000	3.000
Expression Congruency	1.952	0.804	1.000	3.000

Note: Duration is the only variable that we normalized. No other variables have been centered or scaled. All explanatory variables at the participant level are averages of three values measured on a 5-point scale (1–5); Expression type is measured on a 3-point scale (1–3), and Expression Congruency is measured on a 3-point scale (1–3)

negative emotions (angry at others, disgusted with others, jealous), (2) Self-serving positive emotions (happy for self, proud, deserving), (3) Other-serving negative emotions (ashamed of myself, guilty, embarrassed of myself), and (4) Other-serving positive emotions (appreciation, humble, respectful).

Environment/trial level variables Our Level 1 covariates include two environmental descriptions for the emotion that the agent expressed when sharing its metacognitive judgment about the content relevancy. First, expression type could be (1) Confusion, (2) Joy, or (3) Neutral (see Fig. 5). Second, expression congruency could be either (1) Congruent (i.e., showing confusion when either text/diagram were only somewhat relevant or joy when text/diagram were fully relevant), (2) incongruent (i.e., showing joy when either text/diagram were only somewhat relevant or confusion when text/diagram were fully relevant), or (3) neutral (showing no expression).

Dependent variable For our study, dwell-time was the only outcome. Specifically, we were only interested in the total duration (seconds) participants fixated on the agent (area of interest [AOI]) while the agent provided non-verbal, emotion feedback. We classified fixations according to the Dispersion-Threshold Identification algorithm (Salvucci & Goldberg, 2000), which classifies fixations as a minimum of 80 ms on an area of interest less than 100 pixels. All sequential fixations within one AOI were summed to measure the dwell time on that AOI. Due to the non-normal distribution of dwell times, for our study we normalized durations using a Box-Cox transformation ($\lambda = 0.328$) before fitting our model. Interpretations of our intercepts include both the transformed value (as indicated from our model) and an interpretable inverse-transformed value (back into seconds).

Results

Dimension reduction

Due to the large number of potential explanatory variables and our given sample size, we used factor analysis (FA) to reduce the number of dimensions within our data. While there are several dimension reduction methods that can be used (e.g., principal component analysis, linear discriminant analysis, random forest), we chose to use FA as it would also allow us to simultaneously look at the reliability measures for the PAUSe questionnaire.

Before performing FA, we first examined the correlation matrix among our variables and found a Kaiser–Meyer–Olkin (KMO) adequacy measure of 0.64 which both suggest our data has acceptable factorability (Kaiser, 1974). Additionally, we conducted Bartlett's Test of Sphericity ($\chi 2 = 7264.66$, $df = 66$, $p < 0.005$) which indicated factor analysis may be useful for our data. To determine the number of factors to extract we examine the Scree plot (see Fig. 7) which suggested 4 factors.

As we cannot reasonably assume the factors will be independent, and therefore uncorrelated, we used an oblique rotation type and conducted FA using the 'Psych' package in R version 4.1.3 (Revelle, 2022). Our Root Mean Square Error of Approximation (RMSEA) fit index was 0.06 suggesting a close fit (Hu & Bentler, 1999). Figure 8 reports the factor loading of our variables. These factors show high reliability with the labels from the PAUSe questionnaire such that factor 4 (explaining 32% of the variance within our data) relates to the perceived motivational and behavioral utility of self- and other-serving negative emotions. Factor 3 (explaining 26% variance) can be defined as the perceived (motivational, behavioral, and informational) utility of other-serving positive emotions. Factor 1 (explaining 26% variance) can be defined as the perceived (motivational, behavioral, and informational) utility of self-serving positive emotions. Finally, factor 4 (explaining 17% of variance) can be defined as the perceived informational utility of self- and other-serving

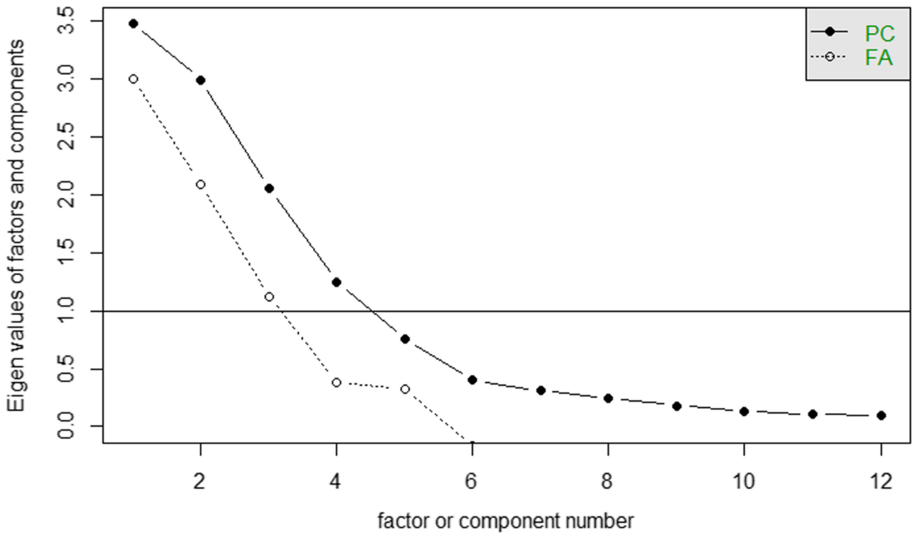

Fig. 7 Scree plot- used to determine number of factors based off where the "elbow" of the eigenvalue line plot levels off

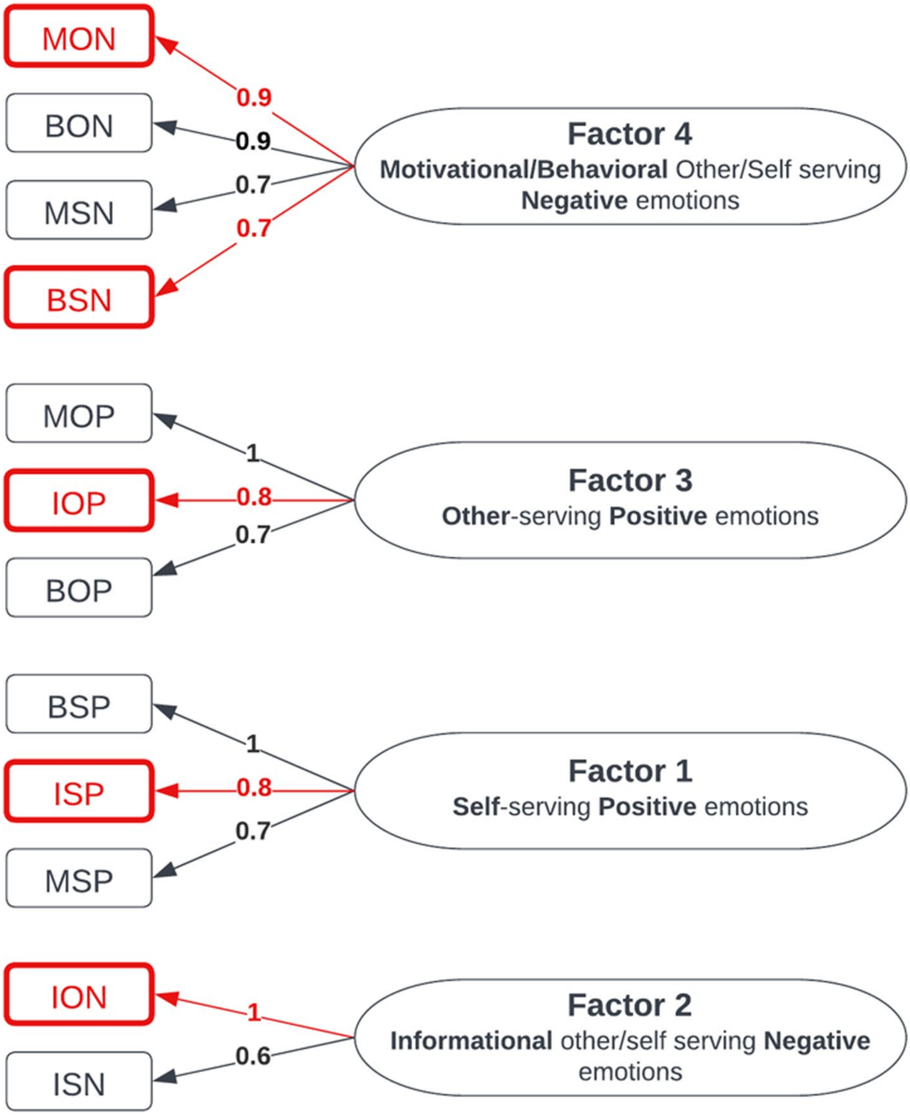

Fig. 8 Factor Loading Diagram. Red variables indicate which variables from each factor based on correlational analysis were chosen for model building. ISN: Perceived Informational Utility of Self-Serving Negative Emotions; ISP: Perceived Informational Utility of Self-Serving Positive Emotions; ION: Perceived Informational Utility of Other-Serving Negative Emotions; IOP: Perceived Informational Utility of Other-Serving Positive Emotions; MSN: Perceived Motivational Utility of Self-Serving Negative Emotions; MSP: Perceived Motivational Utility of Self-Serving Positive Emotions; MON: Perceived Motivational Utility of Other-Serving Negative Emotions; MOP: Perceived Motivational Utility of Other-Serving Positive Emotions; BSN: Perceived Behavior Facilitation Utility of Self-Serving Negative Emotions; BSP: Perceived Behavior Facilitation Utility of Self-Serving Positive Emotions; BON: Perceived Behavior Facilitation Utility of Other-Serving Negative Emotions; BOP: Perceived Behavior Facilitation Utility of Other-Serving Positive Emotions

negative emotions. These factors show a clear delineation between self- and other-serving positive emotions across all utility types. However, for negative emotions, these factors suggest that the split is more informative between utility types (with a grouping of motivational and behavioral versus information utility split) and not person-focused (other versus self).

To reduce dimensions of this instrument for modeling building, using the 'caret' package (Kuhn, 2022), we then identified highly correlated variables within each factor (greater than 0.7). One variable of any identified pair was then selected based on how well it loaded into their respective factors. This resulted in a reduction to 5 variables from our original 12 (see Table 1; see red colored factors in Fig. 8). We further reduce dimension in our model building process by using a backward-elimination approach with our participant-level variables (Level 2 covariates).

Model building process and assumptions

We developed several linear growth models in R version 4.1.3 (using the lmer package; Bates, Machler, Bolker, & Walker, 2014) to find if there is a relationship between how long a learner examines a pedagogical agent peer providing metacognitive monitoring information when controlling for participant-level characteristics (perceived utility of emotions for self and other centered positive and negative emotions) and characteristics of the metacognitive monitoring information provided by the agent (expression type and expression congruency). Our two-level hierarchical linear growth models used maximum likelihood (ML) to estimate our fixed and random effects, and model fits were assessed using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC; Akaike; Akaike, 1974' Raftery et al., 1995). Using a within-subjects experimental design, we examine dwell time during experimental trials (i.e., Level 1, $N = 827$) nested within participants (i.e., Level 2, $N = 50$), which is more than the recommended minimum Level 2 sample size of 30 (Maas & Hox, 2005; Paccagnella, 2011).

We first began by building (Model 1) an unconditional means model with only an intercept and its random effects. Building upon Model 1, we estimated an unconditional growth model (Model 2) by examining how much within-participant variability could be attributed to systematic changes over trials. In this model we assume slopes are fixed across participants (i.e., a random-intercept, fixed-slope model). We then constructed an unconditional growth model with a random-intercept and random-slope (Model 3) and compared both unconditional growth models (Models 2 and 3) against our unconditional means model (Model 1) and each other. Results suggest that there is not a significant difference between Model 2 and Model 3 (fixed versus random slope), but that both outperform our unconditional means model (Model 1). Looking at the ICC for both models led us to adopting Model 3 going forward for a random-intercept, random-slope model. Model 4 built upon Model 3 by adding in our Level 1 environment/trial-level predictors (i.e., expression type and expression congruency). Model 5 built upon Model 4 further incorporating participant-level variables (Level 2 covariates) which will then undergo a backward-elimination approach. Our final model, Model 6, was a random-intercepts random-slopes model estimated using only significant explanatory variables from previous models. Table 2 presents all growth models' coefficients and variance components while examining the participant- and trial-level influences on the duration a participant examines a pedagogical agent as it provides metacognitive monitoring information. All models have both fixed estimations (representing the estimated coefficients) and a random component (representing the

Table 2 Fixed and Random Effect Growth Models of Duration on an Agent Providing Metacognitive Information

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed Effects						
Intercept	1.792* (0.165)	2.326* (0.233)	2.325* (0.249)	3.141* (0.524)	3.212* (0.120)	3.314* (0.591)
Time-level factors						
Trial		-0.166* (0.018)	-0.166* (0.019)	-0.166* (0.016)	-1.349* (0.016)	-0.168* (0.018)
Agent Expression				-0.535 (0.282)		
Agent Expression Congruency				-1.01* (0.295)	-0.402* (0.130)	-1.454* (0.130)
Individual-level factors						
ISP					-2.353 (0.198)	
ION					-0.812* (0.131)	-0.293* (0.132)
IOP					-1.393 (0.158)	
MON					-1.302 (0.144)	
BSN					-1.057 (0.152)	
Random Effects (standard deviations of those random intercepts across participants)						
Intercept	0.890*	0.976*	1.363*	2.639*	2.146*	5.344*
Trial			0.002*	0.002*	0.002*	0.003*
Agent Expression				2.640		
Agent Expression Congruency				3.009*	0.367*	0.365*
Residual (Within-participant variance)	7.701*	6.894*	6.826*	4.637*	5.058*	5.018*
Model fit statistics						
AIC	4094.5	4011.7	4015.0	3822.6	3801.7	3797.1
BIC	4108.6	4030.6	4043.3	3893.4	3872.4	3849.0

*Standard errors are in parentheses. ISP: Perceived Informational Utility of Self-Serving Positive Emotions; ION: Perceived Informational Utility of Other-Serving Negative Emotions; IOP: Perceived Informational Utility of Other-Serving Positive Emotions; MON: Perceived Motivational Utility of Other-Serving Negative Emotions; BSN: Perceived Behavior Facilitation Utility of Self-Serving Negative Emotions variability of the intercept and slope across participants and trials respectively). We report model fit statistics at the bottom.

Data used in our final model's Level 1 and Level 2 residuals were screened for linearity and normally distributed and homogeneous variance. We reviewed linearity and homogeneity of our Level 1 residuals by reviewing Q-Q plots, which plot residuals against predicted values. No discernable patterns in the plot suggest this assumption for our model was met. We reviewed linearity and homogeneity of our Level 2 predictors by examining Empirical Bayes residuals in Q-Q plots. No discernable patterns in the plots suggest this assumption for our model was also met.

Model 1: Unconditional means model

First, we estimated an unconditional means model. Based on this model, we can examine the participant effect on duration of looking at the agent providing metacognitive monitoring information. Our Model 1 can be written as:

Level One:

$$\boldsymbol{Duration}_{ti} = \pi_{0i}\boldsymbol{trial}_{ti} + e_{ti}$$

Level Two:

$$\pi_{0i} = \beta_{00} + r_{0i}$$

where $\boldsymbol{Duration}_{ti}$ is the dwell time of trial t for participant i; π_{0i} is the average duration for the ith participant; β_{00} is the average duration across all participants at time 0 (trial 1); r_{0i} is the deviation of each participant's mean duration from the grand mean (i.e., the random effect of participant i); and e_{ti} is a trial-level residual. The overall mean dwell time (across participants) is estimated as 1.79 (4.09 s)[1] which is significantly different from zero (t (50.37) = 24.85, $p < 0.005$). The between-participant (Level 2) variance in duration is estimated as $\widehat{\sigma}^2_{r0} = 0.89$, 89% and the within-participant between-trial (Level 1) variance residual is estimated as $\widehat{\sigma}^2_{e} = 7.70$, giving a total variance of 8.59. The intraclass correlation (ICC), or proportion of total variance that is explained by the variance between participants, is 0.104. Thus, approximately 10% of dwell time on a pedagogical agent providing metacognitive information is between participants, and 90% of the variance is between trials within a given participant.

Models 2 & 3: Unconditional growth models

Next, we built upon Model 1 to estimate how much within-participant variability can be attributed to systematic changes over time (trials). We estimated an unconditional growth model introducing time (trial). Model 2, a random-intercept fixed-slope model, can be written as:

Level One:

$$\boldsymbol{Duration}_{ti} = \pi_{0i} + \pi_{1i}\boldsymbol{trial}_{ti} + e_{ti}$$

[1] *Note: All models used the transformed version of duration and therefore we provide the value of the transformed duration in addition to the interpretable inverse-transformed value in seconds.*

Level Two:

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10}$$

where β_{10} is the growth slope, or expected yearly rate of change in duration for participant i over the 18 trial period. Before interpreting the coefficients, we also built Model 3, a random-intercept fixed-slope unconditional growth model, which can be written as:

Level One:

$$\boldsymbol{Duration}_{ti} = \pi_{0i} + \pi_{1i}\boldsymbol{trial}_{ti} + e_{ti}$$

Level Two:

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

where r_{1i} is the variance of the random slope between participant variance in growth slopes. We compare both Model 2 and Model 3 against our unconditional means model (Model 1) and find that they are both statistically significantly better fits ($\chi^2=84.71$, $df=1$, $p<0.005$, $\chi^2=85.42$, $df=3$, $p<0.005$, respectively). When we compare Model 3 against Model 2 (adding in a random slope) using maximum likelihood estimates, we find there is not a significant difference between the models ($\chi^2=0.714$, $df=2$, $p=0.67$). However, the ICC for Model 3 (0.166) provides more evidence of clustering than Model 2 (0.124) that we might be able to explain with our environment/trial level covariates and thus providing an overall better fitting model. If we find that neither of these covariates are significant, however, we plan on using Model 2 for the remaining model building process for parsimonious reasons.

According to Model 3, the mean duration for all participants in trial 1 (β_{00}) is 2.32 (5.63 s). The mean trial change in duration for participants (β_{10}) is -0.17, which is statistically significant ($t(50.88)=-8.85$, $p<0.005$) suggesting that the longer a participant is learning with MetaTutor-IVH, the less they look at the agent while it provides metacognitive information. The variance in within-participant deviations ($\hat{\sigma}_e^2$) is 6.83 and the variance between participants on trial 1 ($\hat{\sigma}_{r0}^2$) is 1.36. The variance between participants in rates of change in durations during the experiment ($\hat{\sigma}_{1i}^2$) for participant i is 0.002. The estimated within-participant variance,$\hat{\sigma}_e^2$, decreased by about 11% from our unconditional meals model, suggesting that 11% of within-participant variability in duration looking at the agent providing metacognitive information can be explained by a linear decrease over time (trials).

Model 4: Random-intercept random-slope growth model with level 1 covariates

Next, building upon Model 3, we built a model with our Level 1 covariates, expression type and expression congruency. Model 4 can be written as:

Level One:

$$\boldsymbol{Duration}_{ti} = \pi_{0i} + \pi_{1i}\boldsymbol{trial}_{ti} + \pi_{2i}\boldsymbol{expression}_{ti} + \pi_{1i}\boldsymbol{congruency}_{ti} + e_{ti}$$

Level Two:

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

$$\pi_{2i} = \beta_{20} + r_{2i}$$

$$\pi_{3i} = \beta_{30} + r_{3i}$$

where β_{20} is the expected effect of agent expression over trials with a variance of r_{2i} and β_{30} is the expected effect of agent congruency over trials with a variance ofr_{3i}. Agent expression for participant i is estimated as $-0.54 + r_{2i}$, however this was not a significant predictor (t (10.22) = -1.90, p = 0.09). Agent congruency for participant i, however was a significant predictor (t (13.93) = -3.44, $p < 0.005$), and is estimated to be$-1.01 + r_{3i}$. This suggests that accounting for trial, participant i on a trial with an agent providing a congruent emotion (congruency = 1) are expected to have a duration that is $1.01 + r_{3i}$ units (transformed seconds) lower than on a trial with either an incongruent or neutral emotion. That is, participants show evidence that they are metacognitively detecting when the agent is not accurate and examining that agent for longer, perhaps to decipher why. Comparing Model 4 and Model 3, we find that adding in Level 1 covariates make for a statistically significant better fitting model ($\chi^2 = 8.71$, $df = 4$, $p = 0.04$).

Model 5: Random-intercept, random-slope growth model with level 1 and Level 2 covariates

Model 5 built upon Model 4 and introduced all 4 potential explanatory variables to then undergo backwards elimination to further reduce the number of explanatory variables within the model. Model 5 is defined as:

Level One:

$$\boldsymbol{Duration}_{ti} = \pi_{0i} + \pi_{1i}\boldsymbol{trial}_{ti} + \pi_{1i}\boldsymbol{congruency}_{ti} + e_{ti}$$

Level Two:

$$\pi_{0i} = \beta_{00} + \beta_{01} * \left(informational_{self_{positive_i}}\right) +$$

$$\beta_{02} * \left(informational_{other_{negative_i}}\right) + \beta_{03} * \left(informational_{other_{positive_i}}\right) +$$

$$\beta_{04} * (motivational\_other\_negative_i) + \beta_{05} * (behavioral\_self\_negative_i) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} * \left(informational_{self_{positive_i}}\right) +$$

$$\beta_{12} * \left(informational_{other_{negative_i}}\right) + \beta_{13} * \left(informational_{other_{positive_i}}\right) +$$

$$\beta_{14} * (motivational\_other\_negative_i) + \beta_{15} * (behavioral\_self\_negative_i) + r_{1i}$$

$$\pi_{2i} = \beta_{20} + \beta_{21} * \left(informational_{self_{positive_i}}\right) +$$

$$\beta_{22} * \left(informational_{other_{negative_i}}\right) + \beta_{23} * \left(informational_{other_{positive_i}}\right) +$$

$$\beta_{24} * (motivational\_other\_negative_i) + \beta_{25} * (behavioral\_self\_negative_i) + r_{2i}$$

As Table 2 shows, for perceived informational utility, only the perceived utility of other-centered negative emotions was a significant explanatory variable ($t = -2.96$, $p = 0.004$). Compared to Model 4 (unconditional random-intercept random-slope model), Model 5 is a significantly better fitting model ($\chi 2 = 20.97$, df $= 1$, $p < 0.005$). Model 5 suggests that for every 1 point more average utility in -centered negative emotions a participant perceives, holding all else constant, the rate of change in the duration spent looking at the agent decreases by 0.293 (0.389 s). We failed to find any significant effects on duration from the perceived motivational utility measures or perceived goal-directed behavioral facilitation measures.

Model 6: Final model

To build our final model, we only included the significant predictors from all previous models in a two-level conditional random intercept random slope growth model. Model 6 can be written as follows:

Level One:

$$\boldsymbol{Duration}_{ti} = \pi_{0i} + \pi_{1i}\boldsymbol{trial}_{ti} + \pi_{1i}\boldsymbol{congruency}_{ti} + e_{ti}$$

Level Two:

$$\pi_{0i} = \beta_{00} + \beta_{01} * (informational\_other\_negative_i) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} * (informational\_other\_negative_i) + r_{1i}$$

$$\pi_{2i} = \beta_{20} + \beta_{21} * (informational\_other\_negative_i) + r_{2i}$$

The perceived informational utility of other-centered negative emotions was entered in our model as a Level 2 predictor for the slope (rate of change in duration) while agent expression congruency was entered as a Level 1 predictor. Both predictors were uncentered.

According to our final model, the mean duration for all participants in trial 1 (β_{00}) with an average perceived informational utility of other-centered negative emotions (β_{01}) is 3.31 (9.42 s), which is statistically significantly different from zero ($SE = 0.59$, t $(111.07) = 15.94$, $p < 0.005$). On average and across participants agent expression congruency was negative and statistically significantly related to duration on the agent providing metacognitive information. The average effect (i.e., slope) across participants for agent expression congruency can be represented as a decrease of 0.40 (1.45 s) for every additional trial ($SE = 0.02$, t $(51.10) = -9.83$, $p < 0.005$). Simply, the duration a participant examines an agent providing non-verbal metacognitive monitoring information decreases over time (trials), and the rate of change in this duration is lower for participants who perceive less informational utility in other-centered negative emotions compared to participants who perceive higher utility. Additionally, the rate of change in this duration is also lower on trials in which the agent is providing a congruent emotion expression compared to non-congruent or neutral.

Relative to our unconditional means model, our final model is a significantly better fit model ($\chi^2 = 313.31$, $df = 8$, $p < 0.005$). This is further confirmed by lower AIC and BIC values in our final model compared to our unconditional means model. A likelihood ratio test also suggests that significant variance still exists for participants ($p < 0.005$), or simply, the differences between participants' duration are still unaccounted for with our final model. The proportion reduction in within-participant variance (Level 1;R^2_{level1}) was 0.26, suggesting that adding agent expression congruency as a predictor of duration reduced the within-participant variability by 26% (i.e., accounting for 26% of the person-level variance in duration). The between-participant variance explained at Level 2 (i.e., variation between participants) increased after including the perceived informational utility of other-centered negative emotions.

Discussion

Based on the final model and model building process, we have estimated that there are both individual-level and environment/trial-level effects on the duration a participant spent looking at a pedagogical agent providing non-verbal metacognitive information while learning with MetaTutor-IVH. Our models fully support our hypothesis that we would find a change in duration over time (i.e., trials) spent in MetaTutor-IVH. We argue this is potentially attributable to two major factors. First, as participants become oriented to the learning environment, less time is needed to navigate attention towards the general location of page elements such as where to click next, or where to find the text and diagram. In other words, there is evidence learners were habituating to the environment over the experiment. Second, the novelty of the agent lowers over time in addition to learners potentially becoming faster at interpreting and analyzing the non-verbal cues represented in the agent's facial expressions. That is, the learners become habituated to the agent and its behavior as well as becoming more efficient in interpreting abstract non-verbal metacognitive information.

Across all models, and within our final model, we saw that there was a fixed and random effect of trial on duration. Our results also somewhat support our hypothesis that we would find an effect of duration accounted for by multiple perceived utility of both self- and other-centered emotions (both positive and negative). Only the perception of informational utility of other-centered negative emotions had a significant effect on participants' duration on the agent. We interpret this result to suggest that emotions such as shame and embarrassment could drive participants to look at the agent more. This in conjunction with the fact that no perceived utility goal-directed behavioral facilitation was found significant in our modeling could suggest that learners are more worried about being seen as incorrect or making wrong judgements than actually performing better. Additionally, our FA results, while also used for dimension reduction, provide some evidence for the reliability of the PAUSe metric showing factor loadings of variables that match the theoretical assumptions of the measure.

This modeling provides empirical support for Efklides' (2019) MASRL model of SRL which characterizes SRL into two levels. The informational utility of other-centered negative emotions are relatively static ‘Person’ level features that we showed influenced the way in which learners interacted with the environment. Specifically, it affected the way in which they chose to collect additional metacognitive information via the non-verbal pedagogical agent. This highlights the importance of not only considering one’s current emotional state, but also general disposition towards the role of emotions during learning on an individualized basis.

Finally, we found that agent expression was surprisingly not an influencing factor in the duration learners spent looking at the agent. This suggests that positive and negative emotions do not attract more attention or require additional interpretation time but rather the underlying congruency matters. That is, learners spent less time looking at the agent when its expression represented the ground truth. This could suggest longer attention was required to reevaluate one’s own judgment or compare the agent’s judgment to their own. Similar to the previous result, our model found ‘Person x Task’ level features also influenced the attention allocation on metacognitive monitoring information. This suggests learners were not merely using the agent as a verifier of their own judgment but wanted to incorporate new (and accurate) information into their current metamemory and metacognitive knowledge (Nelson & Narens, 1990). Surprisingly, it was the metacognitive information that we found impacted behavior and regulation of effort and attention, not affect (i.e., the agent’s expression). This could be attributed to the fact that it was not an emotion that the learner was experiencing themselves, but rather observing.

Overall, our results align with the MASRL model, and show the role that metacognition and affect have within SRL (Efklides, 2019). That is, we show that both ‘Task x Person’ and ‘Person’ level processing influence behaviors during SRL. Additionally, it is aligned with previous work that suggests metacognitive judgments can be used to predict effort and attention allocations (Wiedbusch & Azevedo, 2020), and support overall learning outcomes (e.g., Azevedo & Dever, 2022; Jansen et al., 2019).

This study extends previous work on the role of affect in SRL in two major ways. First, it extends previous work by not looking at the role of current learner emotions, but rather a learner’s perception of the role of emotions as a stable trait. If a learner is highly concerned with appearing ashamed in front of others, they may consult nonverbal feedback more often than a learner who is not highly concerned. This has important design implications in that by considering trait-level perceptions of emotion utility, individualized scaffolding can be personalized based on the type of information that would be most impactful for a learner (e.g., motivational versus informational versus behavioral). Non-verbal information as a scaffolding technique is not a fix-all solution for all student needs, however, this work suggests it could have a significant behavioral impact on some learners and their SRL behaviors. Second, our work provides empirical evidence for the MASRL model assumption about ‘Task x Person’ level and ‘Person’ level processing using dynamical analysis approaches via growth modeling. Unlike previous work that uses self-report measures of current affective states (e.g., Cloude et al., 2021; Wortha et al., 2019), or facial expressions (e.gHarley et al., 2015; Lajoie et al., 2021; Taub et al., 2018a, 2018b), our model utilized a stable trait on the perception of emotions.

Limitations

Due to sample size and convergence limitations, we did not include any potential interaction effects in our modeling. With a larger sample size, this might be possible and reveal more significant effects of our explanatory variables that were hidden in our current models. For example, as participants move through the experiment (and therefore across trials), it is possible that their perception of utility for certain emotions could alleviate some of the natural decline in duration that we see in our current models. Additionally, we found evidence that the congruency of the agent's expression had a significant effect on duration spent looking at the agent. We interpreted this to suggest that participants were metacognitively aware to the discrepancy of the agent's accuracy. To further corroborate such a claim, we should only look at trials in which the participant is accurate in their metacognitive judgement. This would significantly reduce our trial (Level 1) sample size, however, and therefore was not included. Future work should consider one's metacognitive accuracy and bias about content when examining the social aspects of tutoring interactions.

In addition to our results, we need further analysis to further substantiate our interpretations of performance-avoidance behaviors we saw with our modeling through other measures (such as the AEQ questionnaire). Examining potential relationships with this other measure could reveal interesting behaviors of how and when learners look for complex metacognitive interpretations outside of their own judgments. Finally, our modeling did not account for other stable demographic characteristics that could be highly correlated with perceived emotion utility including age, gender, ethnicity, and race.

Future directions and implications for intelligent learning systems

Technology-based learning environments should be designed to be individualized and adaptive based on stable and fluctuating differences, but there are many underlying theoretical considerations that still need further exploration (Azevedo & Wiedbusch, in press; Loderer et al., 2020). Our work attempted to address some of these differences by examining the relationship between how long a learner examines a pedagogical agent peer providing metacognitive monitoring information when controlling for participant-level characteristics (perceived utility of emotions for self and other centered positive and negative emotions) and characteristics of the agent (expression type and expression congruency).

As we discussed above, emotions have been shown to impact learning outcomes (Dever et al., 2021; D'Mello & Graesser, 2012; Forgas, 2002; Pekrun & Linnenbrink-Garcia, 2012a, 2012b; Sinclair et al., 2018), but our study highlights the degree to which an individual's perception of emotions can also impact one's approach to learning. This is important to consider when designing advanced learning technologies. Specifically, these learning environments should not be designed to merely elicit, or help regulate, affect as it is occurring, but also to factor in more stable features around affect such as the perception of utility of those emotions. For example, if a learner does not feel other-centered emotions are very important, but places great value in self-centered emotions, a system could prompt more reflection on one's current state and rely less on embedded pedagogical agents' facial expressions to elicit various affective responses (like prompting difficult questions to induce confusion).

While we have found many interesting findings, there is still much work that needs to be done to help make more adaptive and intelligent learning systems. Indicatively, our study focused only on the content-evaluative metacognitive judgements, or judgements that occur during content presentation and relate to the relevancy of the content to the current learning goal (Azevedo & Cromley, 2004; Greene & Azevedo, 2009). Future work should consider a more expansive set of metacognitive judgments such as ease of learning judgements or retrospective confidence judgments. For example, the MASRL model suggests if a learner is feeling anxious prior to learning what they think will be a difficult topic, they might choose to allocate more time and effort to that subject. If, however, they also place high value on other-serving negative emotions (e.g., embarrassment), a system should aim to help alleviate some of that anxiety or provide reassurance if the learner performs poorly. Or conversely, if during reflection the learner reveals they think they're doing particularly well but their performance suggests otherwise, this might signal to the underlying system that the learner's metacognitive judgements are not accurate and need further calibration or prompting to better reflect on their current learning. That is, other behaviors besides attention allocation, such as reflection accuracy or response quality, could also be used to study the effects of metacognition and affect on SRL. Furthermore, this study highlights the need to consider non-task related metacognitive evaluations, such as evaluations of social dynamics. How (in)accurate am I in not only interpreting the facial expressions of others, but what do I believe that information a social interaction is relaying and how does that impact my future learning decisions? The current MASRL model does not account for these type of evaluations or social monitoring that may have an impact on motivation and affect, but could be expanded to do so.

When designing future learning environments, instructional designers should consider how different attention drawing mechanisms might affect the duration of learners looking at pedagogical agents (Dai, Jung, Postman, & Louwerse, 2022). For example, consider a hint popping up as the system detects a decrease in allocated attention, suggesting that the agent could be supplying important information which could cause the learner to pause, reflect, and ultimately revisit their metacognitive monitoring to reevaluate if they are truly ready to keep learning. Other work should focus on other potential manipulations that may have influenced how long a participant looked at the agent. Interpreting a non-verbal facial expression can be effortful and abstract. If learners find diminishing value in these interactions, this could have accounted for the decrease we saw over time. Providing more explicit feedback from the agent on the same monitoring knowledge could prompt learners to pay more attention and use the shared metacognitive knowledge. Finally, we suggest that future research also focus on other aspects of emotions to assess utility. For example, examining the facial expressions of the learner themselves might reveal instances of mimicry that suggest the learner is trying to decipher what exactly the agent is relaying to them (see EASI model; van Kleef, 2009). This could further substantiate claims about the effortfulness of non-verbal clues or provide insight about what learners are cognitively and metacognitively doing when they are examining the agent during this period. As we move forward making these systems increasingly adaptive and individualized, it will be important for future research to continue to address these issues.

Acknowledgements The authors would like to thank the members from the SMART Lab at UCF and the IntelliMedia Group at NCSU for their assistance.

Funding The research presented in this paper is supported by funding from the National Science Foundation, DRL 1431532, Roger Azevedo.

Data availability The data that support the findings of this study are available from the corresponding author, MW, upon reasonable request.

Declarations

Disclosure of potential conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

Research involving human participants and/or animals University IRB approval was received prior to recruitment and data collection.

Informed consent Informed consent was obtained from all participants involved in this research prior to any data collection.

References

Ahmed, W., Minnaert, A., Kuyper, H., & van der Werf, G. (2012). Reciprocal relationships between math self-concept and math anxiety. *Learning and Individual Differences, 22*(3), 385–389.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Pedagogical Intelligence in Education, 26*(1), 205–223.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*(3), 523.

Azevedo, R., & Dever, D. (2022). Metacognition in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *Cambridge handbook of multimedia* (3rd ed., pp. 132–141). Cambridge University Press.

Azevedo, R., & Wiedbusch, M. (in press). Theories of metacognition and pedagogy applied in AIED systems. In du Boulay (Ed.), *Handbook of Artificial Intelligence in Education*. The Netherlands: Springer.

Azevedo, R., Greene, J. A., & Moos, D. C. (2007). The effect of a human agent's external regulation upon college students' hypermedia learning. *Metacognition and Learning, 2*(2–3), 67–87.

Azevedo, R., Mudrick, N. V., Taub, M., & Bradbury, A. E. (2019). Self-regulation in computer-assisted learning systems. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 587–618). Cambridge University Press.

Bannert, M., Hildebrand, M., & Mengelkamp, C. (2009). Effects of a metacognitive support device in learning environments. *Computers in Human Behavior, 25*(4), 829–835.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*(6), 445–457.

Castro-Alonso, J. C., Wong, R. M., Adesope, O. O., & Paas, F. (2021). Effectiveness of multimedia pedagogical agents predicted by diverse theories: A meta-analysis. *Educational Psychology Review, 33*, 989–1015.

Chow, P. I. (2018). Psychology therapies and interventions that raise subjective well-being. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of well-being*. DEF Publishers.

Chow, P. I., & Berenbaum, H. (2012). Perceived utility of emotion: The structure and construct validity of the Perceived Affect Utility Scale in a cross-ethnic sample. *Cultural Diversity and Ethnic Minority Psychology, 18*(1), 55.

Chow, P. I., Berenbaum, H., & Flores, L. E., Jr. (2013). The role of perceived utility of emotion in interpersonal sensitivity and depression. *Journal of Social and Clinical Psychology, 32*(8), 859–877.

Chow, P. I., Berenbaum, H., & Wang, C. (2015). Examining the validity and reliability of an abridged version of the Perceived Affect Utility Scale (PAUSe). *European Journal of Psychological Assessment, 33*(5), 365–376.

Cloude, E. B., Dever, D. A., Wiedbusch, M. D., & Azevedo, R. (2020). Quantifying scientific thinking using multichannel data with Crystal Island: Implications for individualized game-learning analytics. *Frontiers in Education, 5*, 572546. https://doi.org/10.3389/feduc

Cloude, E. B., Wortha, F., Dever, D. A., & Azevedo, R. (2021, September). Negative emotional dynamics shape cognition and performance with MetaTutor: Toward building affect-aware systems. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1–8). IEEE.

Dai, L., Jung, M., Postma, M., & Louwerse, M. (2022). A systematic review of pedagogical agent research: Similarities, differences and explored aspects. *Computers & Educational 190*.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157.

de Boer, H., Donker-Bergstra, A. S., Kostons, D. D. N. M., Korpershoek, H., & van der Werf, M. P. (2012). *Effective Strategies for Self-regulated Learning: A Meta-analysis*. Groningen: GION onderzoek/onderwijs.

Dever, D. A., Wiedbusch, M. D., Cloude, E. B., Lester, J., & Azevedo, R. (2021). Emotions and the Comprehension of Single versus Multiple Texts during Game-based Learning. *Discourse Processes*, 1–22, https://psycnet.apa.org/doi/https://doi.org/10.1080/0163853X.2021.1950450

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231–264.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232.

Efklides, A. (2005). Introduction to the special section: Motivation and affect in the self-regulation of behavior. *European Psychologist, 10*(3), 173–174.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6–25.

Efklides, A. (2019). Gifted students and self-regulated learning: The MASRL model and its implications for SRL. *High Ability Studies, 30*(1–2), 79–102.

Fiorella, L., & Pilegard, C. (2021). Learner-generated explanations: Effects on restudying and learning from a multimedia lesson. *Educational Psychology, 41*(1), 45–62.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906.

Forgas, J. P. (2002). Feeling and doing: Affective influences on interpersonal behavior. *Psychological Inquiry, 13*(1), 1–28.

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*(3), 334–372.

Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology, 34*(1), 18–29.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*(2), 348.

Harley, J. M., Bouchet, F., Hussain, M. 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.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*(1), 1–55.

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34*, 58–73.

Jansen, R. S., Van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review, 28*, 100–292.

Jemstedt, A., Kubik, V., & Jönsson, F. U. (2017). What moderates the accuracy of ease of learning judgments? *Metacognition and Learning, 12*(3), 337–355.

Johnson, W. L., & Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial Intelligence in Education, 26*(1), 25–36.

Johnson, W. L., & Lester, J. C. (2018). Pedagogical agents: Back to the future. *AI Magazine, 39*(2), 33–44.

Jönsson, F. U., & Kerimi, N. (2011). An investigation of students' knowledge of the delayed judgements of learning effect. *Journal of Cognitive Psychology, 23*(3), 358–373.

Kaiser, H. F. (1974). *An Index of Factorial Simplicity. Psychometrika, 39*, 31–36. https://doi.org/10.1007/BF02291575

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349.

Kramarski, B. (2018). Teachers as agents in promoting students' SRL and performance: Applications for teachers' dual-role training program. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 223–239). Routledge/Taylor & Francis Group.

Krämer, N. C., & Bente, G. (2010). Personalizing e-learning. The social effects of pedagogical agents. *Educational Psychology Review*, *22*, 71–87.

Kuhn, M. (2022). caret: Classification and regression training. R package version 6.0-93. https://CRAN.R-project.org/package=caret

Lajoie, S. P., Zheng, J., Li, S., Jarrell, A., & Gube, M. (2021). Examining the interplay of affect and self-regulation in the context of clinical reasoning. *Learning and Instruction, 72*, 101219.

Lehtinen, E. (2003). Computer-supported collaborative learning: An approach to powerful learning environments. In E. de Corte, L. Derschaffel, N. Entwistle, & J. Van Merrienboer (Eds.), *Powerful*

learning environments: Unraveling basic components and dimensions (pp. 35–54). Pergamon/Elsevier Science Ltd.
Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 464.
Linnenbrink, E. A. (2006). Emotion research in education: Theoretical and methodological perspectives on the integration of affect, motivation, and cognition. *Educational Psychology Review, 18*(4), 307–314.
Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement: Introduction to the special issue. *Contemporary Educational Psychology, 36*(1), 1–3.
Loderer, K., Pekrun, R., & Lester, J. (2020). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction, 70,* https://doi.org/10.1016/j.learninstruc.2018.08.002.
Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92.
Mayer, R. E. (2022). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 57–72). Cambridge University Press.
Mayer, R. E., & Fiorella, L. (Eds.). (2022). *The Cambridge handbook of multimedia* (3rd ed.). Cambridge University Press.
McCarley, J. S., & Gosney, J. (2005, September). Metacognitive judgments in a simulated luggage screening task. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 49, No. 17, pp. 1620–1624). Sage CA: Los Angeles, CA: SAGE Publications.
Metallidou, P., & Efklides, A. (2001). The effects of general success-related beliefs and specific metacognitive experiences on causal attributions. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 325–347). Kluwer Academic Publishers.
Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Computers in Human Behavior, 96*, 223–234.
Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation, 26*, 125–173. Academic Press.
Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models. *Methodology*.
Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341.
Pekrun, R. (2016). Academic emotions. In *Handbook of motivation at school* (pp. 120–144). Routledge.
Pekrun, R., & Linnenbrink-Garcia, L. (2012a). Academic emotions and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 259–282). Springer.
Pekrun, R., & Linnenbrink-Garcia, L. (2012b). *Handbook of emotions in education*. Taylor & Francis.
Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*(2), 91–105.
Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48.
Raftery, A. E., Givens, G. H., & Zeh, J. E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association, 90*(430), 402–416.
Revelle, W. (2022) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.9.
Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78).
Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 1–15). Routledge/Taylor & Francis Group.
Schutz, P. A., & Lanehart, S. L. (2002). Emotions in education. *Educational Psychologist, 37*(2), 67–68.
Schutz, P. A., & Pekrun, R. (2007). Introduction to emotion in education. In *Emotion in education* (pp. 3–10). Academic Press.
Schutz, P. A., & Zembylas, M. (2009). Introduction to advances in teacher emotion research: The impact on teachers' lives. In P. A. Schutz & M. Zembylas (Eds.), *Advances in teacher emotion research* (pp. 3–11). Springer.
Schweppe, J., & Rummer, R. (2016). Integrating written text and graphics as a desirable difficulty in long-term multimedia learning. *Computers in Human Behavior, 60*, 131–137.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1258–1266.

Sinatra, A. M., Pollard, K. A., Files, B. T., Oiknine, A. H., Ericson, M., & Khooshabeh, P. (2021). Social fidelity in virtual agents: Impacts on presence and learning. *Computers in Human Behavior, 114*, 106562.

Sinclair, J., Jang, E. E., Azevedo, R., Lau, C., Taub, M., & Mudrick, N. V. (2018). Changes in emotion and their relationship with learning gains in the context of MetaTutor. In *Intelligent Tutoring Systems: 14th International Conference*, ITS 2018, Montreal, QC, Canada, June 11–15, 2018, Proceedings 14 (pp. 202–211). Springer International Publishing.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 204.

Stull, A. T., Fiorella, L., & Mayer, R. E. (2018). An eye-tracking analysis of instructor presence in video lectures. *Computers in Human Behavior, 88*, 263–272.

Tamir, M., & Ford, B. Q. (2012). When feeling bad is expected to be good: Emotion regulation and outcome expectancies in social conflicts. *Emotion, 12*(4), 807.

Taub, M., Azevedo, R., & Mudrick, N. V. (2018a). How do different levels of AU4 impact metacognitive monitoring during learning with intelligent tutoring systems? In *International Conference on Intelligent Tutoring Systems* (pp. 223–232). Springer, Cham.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2018b). 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.

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science, 18*(3), 184–188.

Vauras, M., Salonen, P., & Kinnunen, R. (2008). Influences of group processes and interpersonal regulation on motivation, affect and achievement. *Advances in motivation and achievement*. *Social Psychological Perspectives, 15*, 275–314.

Wiedbusch, M. D., & Azevedo, R. (2020). Modeling metacomprehension monitoring accuracy with eye gaze on informational content in a multimedia learning environment. In *ACM symposium on eye tracking research and applications* (pp. 1–9).

Winne, P. H. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology, 88*(1), 9–20.

Wortha, F., Azevedo, R., Taub, M., & Narciss, S. (2019). Multiple negative emotions during learning with digital learning environments–Evidence on their detrimental effect on learning from two methodological approaches. *Frontiers in Psychology, 10*, 2678. https://doi.org/10.3389/fpsyg.2019.02678

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.