Eye tracking and early detection of confusion in digital learning environments: Proof of concept

Mariya Pachman, Amaël Arguel
Macquarie University, Australia

Lori Lockyer
Macquarie University, Australia; University of Technology Sydney, Australia

Gregor Kennedy, Jason M. Lodge
University of Melbourne, Australia

Research on incidence of and changes in confusion during complex learning and problem-solving calls for advanced methods of confusion detection in digital learning environments (DLEs). In this study we attempt to address this issue by investigating the use of multiple measures, including psychophysiological indicators and self-ratings, to detect confusion in DLEs. Participants were subjected to two intrinsically confusing insight problems in the form of visual digital puzzles. They were asked to solve problems while their eye trajectories were recorded and these data were triangulated with self-ratings of confusion and cued retrospective verbal reports. All participants had a significant increase in fixations on relevant (i.e., related to the solution) and not-relevant areas at an early stage of the problem-solving process. However, only fixations on not-relevant areas were positively correlated with confusion ratings. Moreover, participants who significantly solved the problem differed in their fixations duration on relevant and not-relevant areas from non-solvers. The importance of early detection of confusion and the affordances of emerging technologies for this purpose are discussed.

Introduction

Increasingly higher education learning activities are delivered online or in technology-enabled formats. As students in this context are often considered to be self-directed learners, independently timing their learning activities, such activities tend to include minimal guidance and support (VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003; Yamagata-Lynch, Do, Skutnik, Thompson, Stephens & Tays, 2015). Students often encounter difficulties in this context: in particular, in interpreting tasks set by their teachers and maintaining their engagement with online tasks (Waycott, Dalgarno, Kennedy, & Bishop, 2012). In fact, confusion is a quite common state when learning about complex topics in digital learning environments (Baker, D'Mello, Rodrigo, Graesser, 2010). There are a variety of reasons why online or technology-enabled tasks could be confusing: no timely intervention from a teacher may be available; learners may have difficulties understanding the content or solving a problem; or they may have difficulties in following the optimal learning trajectory. Along with confusion, learners often experience frustration or total disengagement if their confusion lasts for too long (D'Mello & Graesser, 2014). But confusion is also deemed to offer expanded learning opportunities in some cases (Lehman, D'Mello, & Graesser, 2012). That is why the moment when confusion is first experienced is often considered a turning point in the learning process: confusion from this point can develop in a detrimental fashion because of the reason mentioned above if not resolved (VanLehn et al., 2003), or be beneficial for learning because of learner's deliberations on the content (D'Mello, Lehman, Pekrun, & Graesser, 2014). Potentially, one of the keys to keep confusion from contributing negatively to learning would be an early detection of confusion before it becomes non-constructive.

Much of the research on confusion detection is focused on improving intelligent tutoring systems (ITS) by embedding confusion, frustration, and boredom detection features. Less is known about confusion in non-ITS digital learning environments. Further, existing detection systems are only partially automated and often require costly human intervention for classification purposes (see D'Mello & Graesser, 2010). Avenues for creation of fully automated confusion detection systems, which are suitable for more traditional, non-ITS digital learning environments need to be explored.

In this paper we aim to extend the understanding of changes in levels of confusion in digital learning environments by deriving a set of parameters for an early detection of confusion, using materials that are known to be confusing with the addition of eye-tracking. Further, we create the case for a development of an early confusion detection system based on capacities of modern educational technologies underpinned by learning analytics and user gaze recognition features.

We start with a review of confusion research, research on eye movements in situations of cognitive disequilibrium, and confusion detection models in ITS. Then, we investigate learners' behaviour in a confusing problem-solving situation employing eye tracking as a way to extend our understanding of the changes in levels of confusion. The obtained results shed light on which parameters should be considered for an early detection of confusion in digital learning environments and lead to the recommendations for particular types of technologies to be employed for this purpose.

Background

Cognitive disequilibrium and the types of confusion

Confusion is triggered by cognitive disequilibrium (D'Mello & Graesser, 2014). Cognitive disequilibrium can be defined as a state experienced by an individual during learning when obstacles to the normal flow of the learning process are encountered. These obstacles might include uncertainties, errors, anomalous information, or simply new information that contradicts an individual's prior knowledge (D'Mello & Graesser, 2012). In this state an individual is often unsure what to do next. At the same time, the arising contradictions are found to make individuals elaborate on, and engage in, a deeper cognitive processing of the learning material (D'Mello et al., 2014). The resulting outcomes help promote conceptual change and transfer of learning (D'Mello & Graesser, 2012, 2014; Limón, 2001). An exploration of cognitive-affective states - emotions related to learning - generated by cognitive disequilibrium served as a starting point for research on confusion (see D'Mello & Graesser, 2012; Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Grasser & Olde, 2003; Lehman, et al., 2012). Simply put, confusion is seen as an affective expression of cognitive disequilibrium (Lehman et al., 2012). When encountering contradictory information leads to uncertainties and results in cognitive disequilibrium learners feel confused.

Learners, therefore, need to manage their confusion and to engage in confusion resolution activities, such as problem-solving, before they can move on with their learning (D'Mello & Graesser, 2014). The question of *confusion management* however, is not a straightforward one. The difficulty of managing confusion is that there is no single pathway or a single method identified in the literature, but rather a collection of methods that could help learners return to the stage of equilibrium or a smooth processing of new information (D'Mello & Graesser, 2014). One of the explanations as to why there is no unified perspective on how learners manage their confusion is that individual differences, such as age, motivation, personality, and prior knowledge all have an impact on confusion regulation (Lehman et al., 2012). Thus, without addressing individual differences, it would be hard to talk about an instructional formula for the regulation of confusion. At the same time, several common strategies, such as greater use of scaffolding, prompting self-regulation, and providing feedback are mentioned in the literature as prospective generic ways to manage confusion and cognitive disequilibrium (D'Mello et al., 2014; Lehman et al., 2012). Designing digital learning environments containing these features would allow learners to work on the needed combination of methods to return to equilibrium.

Properly managing and resolving one's confusion is crucial for further learning. Only in cases when confusion is properly managed and resolved can the benefits of confusion be harvested, leading to learning gains (D'Mello & Graesser, 2014; Lehman, et al., 2012). Confusion that is deemed beneficial for learning is also called constructive confusion.

Conversely, unresolved confusion may hinder learning (D' Mello & Graesser, 2014; VanLehn, et al., 2003). In particular, D' Mello and Graesser (2014) have found that learners who managed to at least partially resolve

their confusion had significant learning gains in comparison with a group who left their confusion unresolved. Confusion that lasts for too long is also found to have undesirable consequences for learning (Liu, Pataranutaporn, Ocumpaugh, & Baker, 2013). In case of long-lasting confusion, and in the absence of adequate scaffolds, learners were found to give up or get frustrated and disengage from the task (Baker et al., 2010; D'Mello & Graesser, 2014). This type of detrimental confusion is often referred to as non-constructive confusion.

An early detection of confusion in digital learning environments could thus be used as a basis for offering a combination of strategies to help learners manage confusion when they are in danger of following a non-constructive pathway. Such possibilities to empower learners with confusion managing strategies have been discussed in the literature. For example, Lehman et al. (2012) suggested using direct hints and explanations, or even interventions to help learners understand benefits of confusion. D'Mello and Graesser (2014) noted that learners' scholastic aptitude was an important factor influencing confusion resolution. Thus, individualised learning paths through material based on learners' scholastic aptitude could be a good way to help learners manage their confusion. However, few researchers who have considered confusion management strategies have paid attention to the pre-requisite question of how confusion can be detected for the provision of hints and support.

The other known caveat with an early detection of confusion is the dynamic nature of confusion. Confusion is a process rather than an instant occurrence (D'Mello et al., 2014), and this process can produce different outcomes depending on the timing of the measurement. For example, partially resolved confusion results in a lower overall mean confusion rating in comparison with unresolved confusion (see D'Mello & Graesser, 2014; Fig. 2, p. 111). At the same time the authors found that partially resolved confusion peaks relatively early in the learning episode in comparison with unresolved confusion. If their learners were compared at the beginning of the learning process the difference would probably be opposite to the overall results: learners who partially-resolved their confusion would score higher on confusion in comparison with those whose confusion remained unresolved. Thus, choosing suitable measurement slots and understanding the difference of confusion ratings within these slots from overall averages is an important question to consider.

Early detection of confusion

Given that confusion can potentially become non-constructive if not properly regulated, the early detection of incidents of confusion could help prevent this transformation. The importance of the early detection of confusion and other cognitive-affective states, such as frustration and boredom, has been widely discussed in ITS research (Craig, D'Mello, Witherspoon, & Graesser, 2008; D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008; D'Mello & Graesser, 2010; Rodrigo & Baker, 2011; Zeng, Pantic, Roisman, & Huang, 2009).

Initially, researchers who considered detection linked it to a single source of data, from either facial expressions or dialogue utterances (Zeng, et al., 2009). For example, cognitive-affective states were classified by evaluating learners' postures (D'Mello & Graesser, 2009), monitoring facial expressions (McDaniel et al., 2007), and analysing vocabulary and semantics in tutorial dialogues (D'Mello, Dowell, & Graesser, 2009). Detection models that were based on data from a single source had high error rates (D'Mello & Graesser, 2010). In later studies, researchers pooled data from several sources, such as conversational cues, gross body language, and facial features which represented a promising new avenue for ITS research (see D'Mello & Graesser, 2010). However, detection processes described in these research studies were not completely automated. In fact, ITS research is typically based on costly methods of classifying cognitive-affective states, including confusion, by external judges (D'Mello & Graesser, 2010; Rodrigo & Baker, 2011). Even though D'Mello and Graesser's (2010) multimodal system was quite effective in detecting confusion based on facial expressions and discourse correlates (also see Craig et al., 2008; D'Mello, et al., 2008), human judges were still needed to classify these states. Using a totally automated system is suggested as a future research direction by the authors.

The question is how to extend the findings from this existing ITS research into more generic digital learning environments. Usually, digital learning environments, such as learning management systems, do not require learners to have verbal interactions with the system, thus, it is not possible to use verbal cues to detect confusion

which has such value in ITS-based studies. There are clear opportunities to develop both more cost-effective and simpler ways to detect learner confusion in traditional learning management systems and other modern educational technologies. The emergence and maturity of innovative research technologies, such as eye tracking and learning analytics are one such opportunity. Eye or gaze tracking refers to the recording learners' eye movements using a simple web camera connected to the computer screen and further filtering of these data using sophisticated data sorting algorithms. Learning analytics refers to the affordances of contemporary educational technologies to compile and analyse learners' mouse clicks, interactions, time on task, and other non-verbal interactions within digital learning environments. When combined, these two techniques could be developed and implemented in DLEs for early confusion detection.

Using eye tracking for detecting of confusion

Eye tracking has been used as a successful technique for in-depth investigations of general problem-solving as well as situations involving cognitive disequilibrium (Graesser et al., 2005; Knoblich, Ohlsson, & Raney, 2001). Below, we present the findings from several studies that have used eye-tracking for various investigations of this sort, focusing on eye movement parameters used. Specifically, we focus on total fixations duration.

In problem-solving research the increased gaze fixation time on relevant to the solution parts of the problem has been found to occur at the moment directly preceding the successful problem solution (Ellis, Glaholt, & Reingold, 2011; Knoblich, Ohlsson, & Raney, 2001). In studies of cognitive disequilibrium the most successful learners were found to have longer overall fixation duration on device components relevant to the solution (Graesser et al., 2005). It should be noted however, that first of all, in the Graesser et al. (2005) study learners' confusion was not measured directly. It could only be inferred from their success in resolving cognitive disequilibrium introduced through a device breakdown scenario. Second, cognitive disequilibrium resolution (performance) was not measured directly but rather inferred from the quality of the questions asked by participants. Since neither confusion nor performance were measured directly it is difficult to judge whether the most successful learners would have experienced overall higher or lower levels of confusion in comparison to unsuccessful learners. It might have been that the most successful learners had higher confusion at the early stages of the problem-solving and lower overall confusion in comparison with unsuccessful learners, as in D'Mello and Graesser (2014).

Finally, DeLucia, Preddy, Derby, Tharanathan, and Putrevu, (2014) investigated participants' eye movements when participants remotely operated two different devices. They measured confusion using a subjective Likert-type measure (e.g., *I was confused*) and were not able to find consistent common correlation patterns between the variables for both devices, but only for several tasks performed with a second device. In particular, they found that higher confusion ratings were positively correlated with the total fixation time on the whole screen, mean fixation duration (long fixations) and task completion time (longer task completion). Unfortunately, these researchers did not report on correlations between confusion ratings and total fixation duration for relevant areas and other areas of the screen. They simply did not have this differentiation. Thus, it is not clear whether any significant correlations would remain if confusion ratings were linked with fixations on relevant parts of the screen.

Although eye-tracking emerges as an effective technique in gathering detailed data related to cognitive disequilibrium and incidents of confusion, the existing findings are limited. We believe that exploring correlations of self-rated confusion and fixations on relevant parts of the screen could be beneficial not only for research on confusion but could also help with identifying parameters for automated methods of early detection of confusion.

The present study

This study investigated a set of parameters for an early detection of confusion in non-ITS digital learning environments. It is expected that using eye-tracking will broaden our understanding of confusion and potential parameters associated with its early detection. Thus, exploring eye-tracking behaviour should help us (1) find

correlates of confusion and, based on these data, (2) derive measures of an early detection of confusion in digital learning environments.

Following Graesser et al. (2005) and DeLucia et al. (2014) we adopted a working assumption that longer overall fixation durations denotes a greater amount of cognitive processing. The longer people fixate on relevant to solution areas of the problem in total, the higher their chance to resolve cognitive disequilibrium. We should be cautious, however, about hypothesising a relationship between longer total fixation durations and confusion ratings: although Graesser et al. (2005) have used eye-tracking in situations involving cognitive disequilibrium, they did not include measurements of confusion in their design. DeLucia et al. (2014), on the other hand, did not differentiate between relevant and not-relevant areas of the screen when reporting a positive correlation between total fixation durations and confusion ratings. Finally, D'Mello and Graesser (2014) used the same experimental materials as Graesser et al. (2005) and found that partially resolved confusion led to the higher problem-solving/troubleshooting performance than unresolved confusion. The partially resolved confusion group also rated their confusion higher than the unresolved confusion group for the first half of the problem-solving process. However, D'Mello and Graesser (2014) did not use eye tracking in their study. Thus, our hypothesis is partially based on the confusion research and partially on problem-solving research using eye-tracking.

Based on these previous studies it is expected that the level of learners' self-reported confusion will be positively correlated with fixations on relevant areas of the problem. In other words, high total fixation durations on relevant areas will be correlated with high confusion ratings.

Method

Participants

Fourteen young adults (university students from a large metropolitan Australian university) volunteered to participate in the study. The recruitment was conducted via an internal university employment website. All of the participants were novices in regard to the insight problems used in the experiment. This was confirmed by their statements during cued-retrospective reporting. Participants were compensated at the rate $15 per hour.

Materials

We used multimedia based insight problems to generate confusion in this study. An insight problem is a problem that requires the learner to shift his or her perspective and view the problem in a novel way in order to achieve the solution (Dow & Mayer 2004). In lay terms, learners need to have an "Aha!" moment to solve such a problem. Insight problems are considered inherently challenging (Knoblich et al., 2001) and serve as a good instructional material for the exploration of confusion (see Andres, Andres, Rodrigo, Baker, & Beck, (2015).

Insight problems presented to our participants were transformation puzzles, in which the pieces could form two different layouts showing different pictures (the missing square puzzle, Figure 1, and the 13 crystal skulls puzzle, Figure 2). Both problems were developed using Mathematica 10 (Wolfram Research, 2014) and presented as on-screen simulations with learner control: participants were encouraged to manipulate a scrollbar to transform the problem from its initial state to the final state and to draw comparisons when needed.

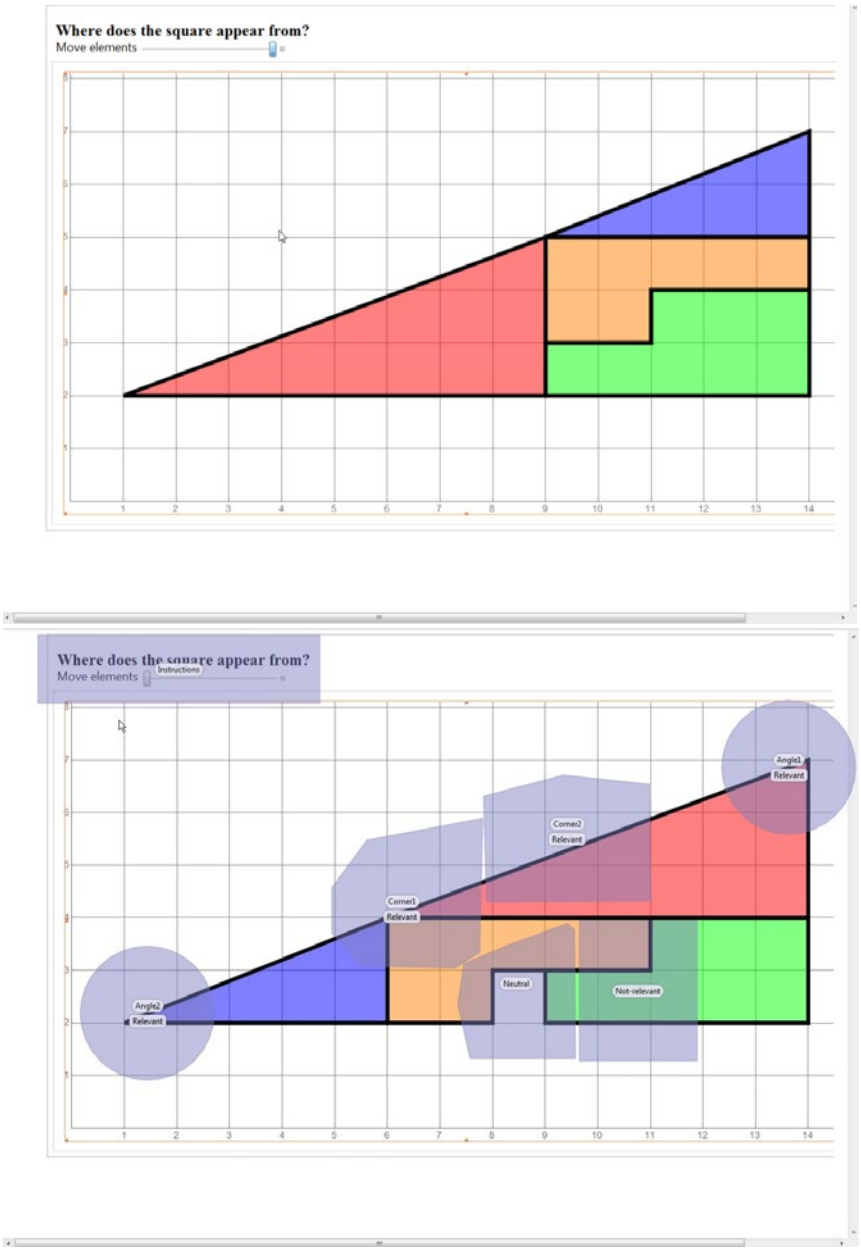

Figure 1. The initial and the final positions of the missing square puzzle with areas of interests (AOIs) marked in grey. The initial question was: "Where does the white square appear from?" (Picture by Krauss / CC BY-SA 4.0)

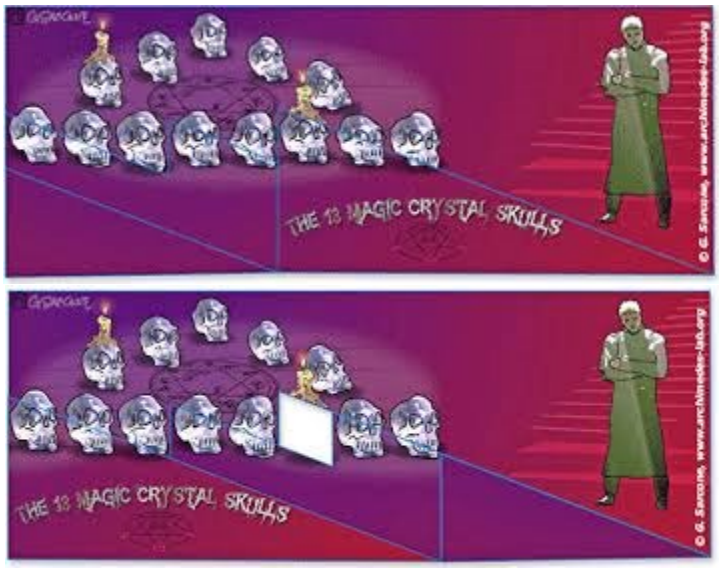

Figure 2. The initial and the final positions of the 13 crystal skulls puzzle. The initial question was: "Where does the 13th skull disappear to?" (Picture by Gianni Sarcone / CC BY-NC-ND 3.0 US)

Paper-based materials consisted of a laminated A4 format sheet of instructions and six 4.5"x3" laminated feedback cards as well as four laminated A4 format sheets representing the initial and final positions of each problem. Feedback card 1 (hint 1) used a questioning technique probing participants to make a comparison between specific parts of the puzzle. Feedback card 2 used a rhetoric question and gave an immediate answer in regard to the other specific parts of the puzzle. Computer-based materials consisted of a visual-spatial abilities test and a self-rated confusion scale. The materials were administered via secure sequence of Qualtrics™. Tobii T120 Eye Tracker integrated into a 17" TFT monitor with an angular resolution of less than 0.5° was used to record participants' eye movements and to replay their gaze trajectories back to them at the retrospective cued reporting stage. Sampling frequency of 60 Hz was used for the current study. A laptop with installed Tobii Studio 2.3 software operated the calibration of the eye tracking system and acquisition of data. Computer-based materials were all presented on the eye tracker monitor, and the mouse connected to the laptop was used to manipulate the interactive materials. Sony ICD series audio recorder was used to gather participants' verbalisations at the cued retrospective reporting stage.

Data sources and measures

A visual-spatial abilities test, The Card Rotation test with a 3 minute time limit (Ekstrom, French, & Harman, 1976) was used to measure learners' visual-spatial abilities. Visual-spatial abilities tests such as The Card Rotation test and The Paper Folding test (Ekstrom, et al., 1976) are routinely used in research on multimedia learning, since visual-spatial abilities mediate the effects of learning with instructional multimedia (Mayer & Sims, 1994; but also see Paik & Schraw, 2013). Specifically, learners with low visual-spatial abilities experience greater difficulty in completing a task containing visual and verbal information (Mayer & Sims, 1994). Since our tasks contained both visual and verbal information, we wanted to be able to estimate the influence of this factor (i.e., visual-spatial abilities) on the final performance.

Participants' behaviours (e.g., wrong answers, early feedback requests) and the timing of the responses during the problem-solving task were recorded by the experimenters using an observation sheet. Participants' gaze trajectories and fixations were recorded with the eye tracker. Audio records of participants' verbal reporting were collected. A self-rated confusion measure ("Please, select the number below that best represents your level of confusion as experienced in this time point", Likert scale 1 to 10) similar to the measure used by D'Mello and Graesser (2014) was administered to the participants, and they rated their confusion for each 1-minute

interval of the problem-solving phase. The measure was considered to be a viable method to document changes in confusion levels because the previous research established that the retrospective confusion ratings of this sort correlate with online recordings of facial expressions and body language (D'Mello & Graesser, 2010; McDaniel et al., 2007). Finally, participants problem-solving success (solved/did not solve) served as a performance measure.

The study used cued-retrospective reporting (Van Gog, Paas, Van Merriënboer, & Witte, 2005) to collect self-rated confusion measures and verbal protocols. Participants were presented with their own gaze trajectories after the problem-solving task as a cue to retrospectively describe the thoughts they had during a problem-solving process; they were also self-rating their confusion level for each 1-minute time interval of the task (see D'Mello et al., 2014). For their study, Van Gog et al. (2005) gathered participants' verbalisations in concurrent, retrospective and cued-retrospective reporting conditions. Participants' eye-movements were recorded while they were problem-solving. Both concurrent and cued-retrospective reporting resulted in more detailed learning process descriptions than simple retrospective reporting (Van Gog et al.'s, 2005). Since Van Gog et al. (2005) were only collecting participants' verbalisations, and we were collecting confusion measures in addition to verbal protocols, concurrent reporting had a danger of introducing substantial interruptions into the learning process. Thus, we employed cued-retrospective reporting to gather a detailed verbal data at the same time minimising the risk of learning process interruptions while frequent confusion measures were taken.

Eye-tracking data and confusion self-ratings were analysed for the first 2 minutes of the problem-solving period. To remind, the independent problem-solving took place within the first 2 minutes, after that learners' gaze and confusion ratings could have been influenced by feedback. Our goals in this study were educational, and we wanted to assure learners did not feel helpless facing unfamiliar complex problems. Thus, all the learners were provided with feedback (hint 1 after 2 minutes and hint 2 after 4 minutes of the overall problem-solving period).

Procedure

Experimental manipulations took place in a laboratory setting. All participants were tested in individual sessions of approximately 45-55 minutes. First, all the participants were asked to read the paper-based instructions explaining the context of the study, given the opportunity to ask study-related questions if they had them, then asked to read and sign consent forms. Participants were then pre-tested on their visual-spatial abilities, seated in front of the eye-tracker and probed to solve their first problem after the calibration. If they did not produce either an incorrect or correct solution or ask for a hint within the first 2 minutes, hint 1 was suggested to them. Alternatively, if they asked for a hint earlier than in 2 minutes time, they were probed to further elaborate on the task and were provided with the hint after 2 minutes if they still needed it. This interval (2 minutes) was considered a minimum sufficient time for successful problem-solving to take place, based on pilots. Hint 2 was suggested in the similar circumstances after the first 4 minutes. During the response participants were instructed to turn away from the eye tracker and explain their prospective solution using paper sheets representing the initial and final positions of each problem. If a participant failed to produce a correct solution within 10 minutes the solution was given to the participant. During the problem-solving process, an experimenter recorded participant questions, their timing, and the timing of the solution by filling in an observation sheet. After problem 1 was resolved, the same procedure was repeated with the problem 2. Participants were then shown the recordings of their eye movements during both problem-solving activities and were asked to think aloud and explain their thought processes. They also rated their past confusion levels on self-rated measure in 1-minute intervals. Finally, participants were debriefed, compensated, and discharged.

Visual-spatial abilities and problem 2

Visual-spatial abilities scores (M = 4.88; SD = 2.04) were not significantly correlated with confusion ratings for either problem 1 (r = .05) or problem 2 (r = .05). Nor did they influence problem-solving outcomes: Wald (1, 13) = 2.64; p = .10, Exp(b) = 1.88 for problem 1, n_s = 7, n_{ns} = 7 (problem 2 had only three non-solvers and logistic regression could not be calculated). Since the results were not significant we did not include visual-spatial abilities in any further calculations.

Problem 2 was included for a possible replication of the trends found with problem 1. However, the data show that participants spent significantly less time: M_2 = 5 min 25 sec; M_1 = 7 min 40 sec; Wilks' lambda = 0.69, $F(1, 13) = 5.90$, $p = .030$, and were significantly less confused: M_2= 5.17; M_1= 7.23; Wilks' lambda = 0.41, $F(1, 13) = 18.43$, $p = .001$ with problem 2 in comparison to problem 1. Participants' verbalisations seem to point at the fact that problem 1 served as kind of pre-training for solving problem 2 (i.e., "The first one gives you a hint for the second one"). Thus, only problem 1 was considered for data analysis.

Results

A range of pre-processing of the data was required in order to arrive at the variables used in this investigation. To calculate total fixation durations a Tobii default fixation filter was applied to the raw data, such that if a participant engaged in glances within a radius of 35 pixels for more than 75 milliseconds they were determined to be fixations. An initial look at problem-solving times of participants revealed that problem-solving times were uneven and ranged from 1 minute 40 seconds to more than 10 minutes for some of the unsuccessful problem-solvers (M = 7 min 40 sec; SD = 2 min 57 sec).

The main set of analyses responded to the hypothesis that the level of learners' self-reported confusion will be positively correlated with fixations on relevant areas of the problem. First, we describe the changes in self-reported confusion ratings. There was a significant decrease in confusion ratings for all participants between minute 1 (M = 8.69; SD = 1.93) and minute 2 (M = 6.85; SD = 3.05), Wilks' lambda = 0.39, $F(1,13) = 19.20$, $p < .01$. All participants seemed to believe they understood the problem well enough and were on the right track for solution.

Second, we report on eye fixations particularly as they related to instructions, non-relevant, neutral, and relevant areas of interest (AOIs). Data from all four relevant AOIs (see Figure 1) were combined for this analysis, since the original multiple relevant AOIs were drawn to avoid any misinterpretations due to the dynamic nature of the environment. Thus, only the areas that could be unequivocally related to the problem solution in various positions of the scrollbar were considered relevant AOIs. The results demonstrated that the whole group had longer total fixation times focusing at relevant AOIs (Wilks' lambda = 0.57, $F(1,13) = 9.95$, $p = .01$) at minute 2. While participants had spent on average about 8 seconds fixating at relevant areas at minute 1, this number had doubled at minute 2 (M_{r1} = 7.71; SD_{r1} = 6.19; M_{r2} = 15.41; SD_{r2} = 10.67). Simultaneously, the whole group also had longer total fixation times focusing at not-relevant AOIs (Wilks' lambda = 0.65, $F(1,13) = 6.89$, $p = .02$) between minutes 1 and 2. At the beginning participants spent on average about 8 seconds fixating at not-relevant areas, this number had increased at minute 2 (M_{n1} = 7.65; SD_{n1} = 4.17; M_{n2} = 11.49; SD_{n2} = 6.32).

Finally, in regard to correlations, a comparison of average participants' fixations for the first 2 minutes and average confusion ratings for the first 2 minutes demonstrated a medium sized positive correlation between fixating on not-relevant AOI and increase in confusion ratings (Pearson's r =.59, $p = .03$).

To more closely investigate the finding that participants were fixating significantly more on not-relevant and relevant AOIs between minutes 1 and 2, we introduced problem-solvers and non-solvers groups as a between-subjects factor in the analysis. The results demonstrate that solvers were fixating more on relevant information in comparison with non-solvers: $F(1,12) = 5.38$, $MSE = 82.62$, $p = .04$, although both groups had a total overall increase in fixations on relevant information between minutes 1 and 2: $F(1,12) = 5.38$, $MSE = 82.62$, $p = .04$ (Figure 3). It was also clear that non-solvers were fixating somewhat more on not-relevant information in comparison with solvers: $F(1,12) = 4.75$, $MSE = 32.94$, $p = .05$, although both groups had a total overall increase in fixations on not-relevant information between minutes 1 and 2: Wilks' lambda = 0.65, $F(1,12) = 6.47$, $p = .03$ (Figure 3).

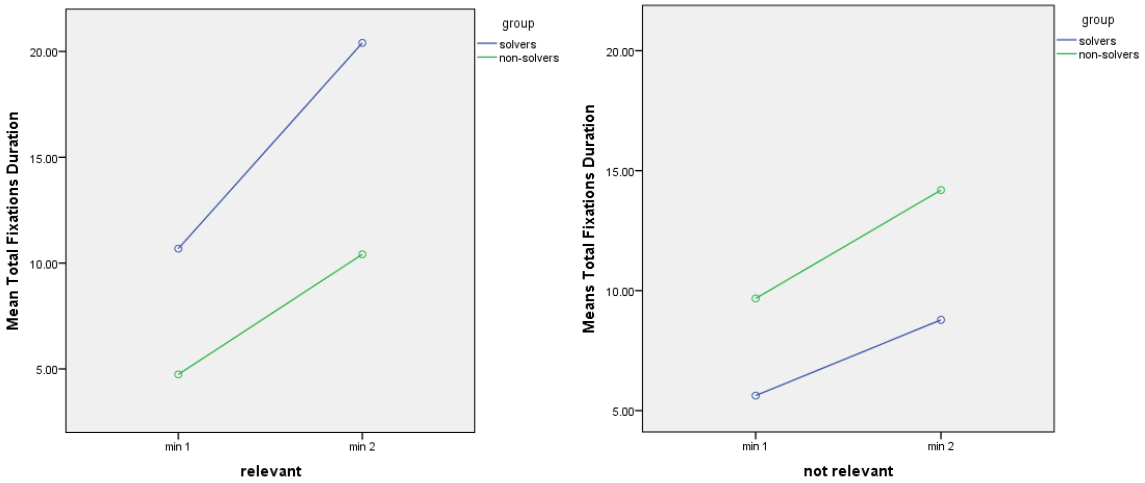

Figure 3. Comparison of solvers and non-solvers total fixations on not-relevant and relevant AOIs during the first 2 minutes of the problem-solving.

Expected changes in confusion self-ratings were not consistent with the findings above. Despite the original finding that all participants rated their confusion lower between minutes 1 and 2, the group differences or interaction were not significant for changes in confusion self-ratings: $F(1,12) = 1.58$, $MSE = 11.33$, $p = .23$ for a group and Wilks' lambda = 0.89, $F(1,11) = 1.3$, $p = .28$ for the interaction term.

Discussion

This study used potentially confusing digital material (insight problems) with eye-tracking and self-report measures to extend the understanding of changes in the levels of confusion. The ultimate aim was to derive a set of parameters for an early detection of confusion in digital learning environments.

We hypothesised that learners' self-reported confusion would positively correlate with fixations on relevant areas of the problem. The results did not support this hypothesis. Rather, they have supported an alternative hypothesis that self-reported confusion is positively correlated with fixations on not-relevant elements of the problem (the finding that was not particularly discussed in the literature). One of the possible explanations of this finding is that we have focused on the first 2 minutes of the problem-solving instead of the whole solution period because of the uneven solution times and feedback provided after the first 2 minutes. In Graesser et al.'s (2005) study, troubleshooting period was limited to 90 seconds and participants were well aware of it and able to adjust their strategies accordingly. The other point mentioned before is that Graesser et al. (2005) did not use measures of confusion. Their reasoning about being or not being able to resolve an arising cognitive disequilibrium was based on the quality of participants' questions (serving as indicators of performance). Thus, we do not have confusion measurements from their study. We can only assume that learners' confusion was growing at the beginning of the learning process until some of the learners were able to identify the potential problem area on a screen; they reached an "aha" moment. At the same time our finding that all the participants fixated more on relevant areas from minute 1 to minute 2 potentially adds to Graesser et al.'s (2005) results with effective learners during the 90 sec troubleshooting sessions. As a reminder, Graesser et al. (2005) have found that only effective learners fixated more on relevant areas. Ineffective learners fixated on relevant areas at the level of chance (randomly). It could well be that our current findings are aligned with DeLucia et al.'s (2014) results (i.e., confusion ratings are positively correlated with fixations), but since the authors did not report correlations between confusion and fixations on specific areas of the screen, this conclusion cannot be made.

In a broader sense our finding that self-reported confusion has a positive correlation with fixations on not-relevant area has links to some of the findings from problem-solving research. In particular, Hodgson, Bajwa, Owen, & Kennard, (2000) found that poor problem-solvers attempting a logical puzzle fixate more on irrelevant units of the puzzle than good problem-solvers. In our case all participants fixated significantly more on not-relevant AOIs between minute 1 and 2, but non-solvers fixated more than solvers (similar to Hodgson et al., 2000). These long fixations were however not accompanied by a significant increase in confusion for either of the two groups (7 participants in each). The found correlation was only true for the whole sample. It is quite possible, however, that there might have been insufficient power to detect an effect because of small sample sizes of the solver and non-solver groups.

The other possibility is that non-solvers were still at the stage of cognitive disequilibrium without an idea of how to resolve it while solvers could foresee a potential for solution (they had longer total fixations on relevant areas) and rated their confusion somewhat lower than non-solvers during minute 2 (M_s = 5.86; SD_s = 3.67; M_{ns} = 8.00; SD_{ns} = 1.79). As mentioned in the results, the group difference in confusion self-ratings was not statistically significant. Besides, the solvers group was not very homogenous in their confusion ratings for minute 2 (large standard deviation).

Our findings have potential theoretical implications for confusion research and specifically, for early detection of confusion. There is a possibility that instead of tracking whether learners fixate on areas relevant for task completion or a problem-solution area, confusion researchers should first evaluate learners' fixations on not-relevant parts of the screen coupled with relatively high confusion ratings. While fixating on relevant areas is conducive for a successful problem solving and a higher performance, fixating on not-relevant parts and relatively high confusion ratings help detecting potential non-constructive confusion cases.

In practice, a combination or a choice of strategies to manage confusion could be provided to the learners fixating on not-relevant elements of the material and rating confusion relatively high at the beginning of the learning process. As we have discussed, feedback, advanced scaffolding, and other methods could in time potentially help such learners resolve their cognitive disequilibrium and avoid non-constructive confusion.

Besides, practical implications of our findings are relatively easy to implement using the existing technologies. While early fixations on not-relevant areas of the screen could be assessed using one of the low cost web cam based gaze recognition applications, such as xLabs (https://xlabsgaze.com/), PyGaze (http://www.pygaze.org/) or GazeHawk (http://gazehawk.com/), regular confusion ratings could be embedded within a learning management system. Combined in the learning analytics engine, these data could serve as a basis for creation of a fully automated early warning system used in the existing not-ITS digital learning environments. Specifically, learners could be probed to rate their confusion at 30 second to 1 minute intervals after they have accessed a particular resource, or to simply click an emoticon-based button when they feel uncertain about the presented information. At the same time the system will detect if they fixate on not-relevant features starting from the moment they have accessed a particular resource.

While our finding suggests a relatively simple clear-cut confusion detection method, further testing and fine tuning of this method in the real technology-enabled classrooms is a must. The implementation of confusion management strategies mentioned above also requires future research. Overall, however, the multimodal method of an early confusion detection presented in this paper could serve as an example of using simple recognition parameters to stimulate fully automated confusion detection.

Limitations

One of the limitations of the current study is the post-factum division on successful and unsuccessful problem-solvers. While the consequences of insight problem-solving process cannot be predicted (Knobich et al., 2001), it is safe to assume that some people will solve such problems fast and some will solve the problems slowly, while others will not solve a problem at all within an allotted time. While post-factum division on groups complicate the making of inferences from the obtained results, such divisions are quite common in problem-

solving and confusion research alike (see D'Mello & Graesser, 2014; Graesser et al., 2005; Knobich et al., 2001). A second, limitation is that since problem 1 seemed to serve as pre-training for problem 2 it affected our within-subjects analysis. A design including independent problems could allow for detailed repeated measures comparisons and a discussion on the role of the task features in confusion detection. Third, individual differences in problem-solving could have influenced participants' problem-solving trajectories but we failed to collect the demographic data to be able to assess this influence. Fourth, although think aloud data was collected it has not been systematically analysed, but rather used for triangulation of the eye-tracking data, that is participant gazing at not-relevant areas and talking about an incorrect solution. Fifth, we have included a pre-test to assess visual-spatial abilities of our participants, but the results of this test did not have a significant influence on further problem-solving success. Possibly, a more extensive assessment of visual-spatial abilities in future studies could help uncover the influence of this factor on the final performance. Finally, the dynamic nature of an on-screen stimuli and the limits of the technologies did not allow for a more detailed analysis of learning trajectories in relation to the moving puzzle pieces.

Future directions

While this study provides a proof-of-concept for early confusion detection, it does not test the validity of generic confusion management methods (i.e., advanced scaffolding, introduction of self-regulatory techniques) for a confusion resolution. Further research could shed the light on effectiveness of these interventions after confusion is detected. First, learners could receive some information on confusion and how to manage confusion in digital learning environments. Second, they could be pre-trained on self-regulatory techniques. Third, they could be shown a video or a simulated example of their peer managing confusion in a similar situation. Fourth, further research should seek to replicate the results of this study and possibly investigate additional indicators for confusion detection, such as a change of posture, pulse, and in facial muscles activity. A word of caution, however, should be added in terms of implementation of these potentially invasive methods in educational settings. Finally, the results of our study could be evaluated in a realistic higher education context, once an early warning system based on the discussed confusion detection parameters is implemented within learning management system. Overall, an implementation of such system could help promote learning and avoid detrimental outcomes of non-constructive confusion.

Acknowledgements

This research is funded by the Science of Learning Research Centre - A Special Research Initiative of the Australian Research Council (SR120300015).

References

Andres, J. M. A. L., Andres, J. M. L., Rodrigo, M. M. T., Baker, R. S., & Beck, J. B. (2015). An investigation of eureka and the affective states surrounding eureka moments. In H. Ogata et al. (Eds.), *Proceedings of the 23rd International Conference on Computers in Education.* China: Asia-Pacific Society for Computers in Education.

Baker, R. S. J. D, D'Mello, S., Rodrigo, M., & Graesser, A. (2010). Better to be frustrated than bored: The incidence and persistence of affect during interactions with three different computer-based learning environments. *International Journal of Human-computer Studies, 68*(4), 223–241. http://dx.doi.org/10.1016/j.ijhcs.2009.12.003

Craig, S., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote-aloud during learning with AutoTutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition and Emotion, 22*(5), 777-788. http://dx.doi.org/10.1080/02699930701516759

DeLucia, P., Preddy, D., Derby, P., Tharanathan, A., & Putrevu S. (2014). Eye movement behavior during confusion: Toward a method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, September 2014, 58*(1), 1300-1304. http://dx.doi.org/10.1177/1541931214581271

D'Mello, S., Craig, S. D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction,*

18(1-2), 45-80. http://dx.doi.org/10.1007/s11257-007-9037-6

D'Mello, S., Dowell, N., & Graesser, A. C. (2009). Cohesion relationships in tutorial dialogue as predictors of affective states. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of 14th International Conference on Artificial Intelligence In Education*, (pp. 9-16). Amsterdam: IOS Press.

D'Mello, S., & Graesser, A. (2009). Automatic detection of learners' emotions from gross body language. *Applied Artificial Intelligence*, *23*(2), 123-150. http://dx.doi.org/10.1080/08839510802631745

D'Mello, S., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction*, *20*(2), 147-187. http://dx.doi.org/10.1007/s11257-010-9074-4

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145-157. http://dx.doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S., & Graesser, A. (2014). Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica*, *151*, 106-116. http://dx.doi.org/10.1016/j.actpsy.2014.06.005

D'Mello, S., Lehman, B. Pekrun, R., & Graesser, A. C. (2014). Confusion can be beneficial for learning. *Learning & Instruction*, *29*(1), 153-170. http://dx.doi.org/10.1016/j.learninstruc.2012.05.003

Dow, G. T., & Mayer, R. E. (2004). Teaching students to solve insight problems. Evidence for domain specificity in training. *Creativity Research Journal*, *16*(4), 389-402. http://dx.doi.org/10.1080/10400410409534550

Ekstrom, R., French, J., & Harman, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: ETS.

Ellis, J., Glaholt, M., & Reingold, E. (2011). Eye movements reveal solution knowledge prior to insight. *Consciousness and Cognition*, *20*(3), 768-76. http://dx.doi.org/10.1016/j.concog.2010.12.007

Graesser, A., Lu, S., Olde, B., Cooper-Pye, E., & Whitten S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory & Cognition*, *33*(7), 1235-1247. http://dx.doi.org/10.3758/BF03193225

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, *95*(3), 524–536. http://dx.doi.org/10.1037/0022-0663.95.3.524

Hodgson, T., Bajwa, A., Owen, A., & Kennard, C. (2000). The strategic control of gaze direction in the Tower-of-London task. *Journal of Cognitive Neuroscience*, *12*(5), 894-907. http://dx.doi.org/10.1162/089892900562499

Knobich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29*(7), 1000-1009. http://dx.doi.org/10.3758/BF03195762

Lehman, B., D'Mello, S. K., & Graesser, A. C. (2012). Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, *15*(3), 184-194. http://dx.doi.org/10.1016/j.iheduc.2012.01.002

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction*, *11*(4-5), 357-380. http://dx.doi.org/10.1016/S0959-4752(00)00037-2

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. (2013). Sequences of frustration and confusion, and learning. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.) *Proceedings of the 6th International Conference on Educational Data Mining Conference, Memphis, TN*, 114-120.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, *86*(3), 389-401. http://dx.doi.org/10.1037/0022-0663.86.3.389

McDaniel, B. T., D'Mello, S. K., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial features for affective state detection in learning environments. In D. S. McNamara, & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 467-472). Austin, TX: Cognitive Science Society.

Paik, E. S., & Schraw, G. (2013). Learning with animation and illusions of understanding. *Journal of Educational Psychology*, *105*(2), 278-289. http://dx.doi.org/10.1037/a0030281

Rodrigo, M. M. T., & Baker, R. S. J. D. (2011). Comparing learners' affect while using an intelligent tutor and an educational game. *Research and Practice in Technology Enhanced Learning*, *6*(1), 43–66. http://dx.doi.org/10.1007/978-3-540-69132-7_9

Van Gog, T., Paas, F., Van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology, Applied, 11*(4), 237-244. http://dx.doi.org/10.1037/1076-898X.11.4.237

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209 –249. http://www.jstor.org/stable/3233810

Waycott, J., Dalgarno, B., Kennedy, G., & Bishop, A. (2012). Making science real: Photo-sharing in biology and chemistry. *Research in Learning Technology, 20*. http://dx.doi.org/10.3402/rlt.v20i0.16151

Wolfram Research (2014). *Mathematica 10* [computer program]. Champaign, IL: Author.

Yamagata-Lynch, L. C., Do, J., Skutnik, A. L., Thompson, D. J., Stephens, A. F., & Tays, C. A. (2015). Design lessons about participatory self-directed online learning in a graduate-level instructional technology course. *Open Learning, 30*(2), 178-189. http://dx.doi.org/10.1080/02680513.2015.1071244

Zeng, Z., Pantic, M., Roisman, G. & Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39-58. http://dx.doi.org/10.1109/TPAMI.2008.52

Corresponding author: Mariya Pachman, korotenkom@yahoo.com

Please cite as: Pachman, M., Arguel, A., Lockyer, L., Kennedy, G., & Lodge, J. M. (2016). Eye tracking and early detection of confusion in digital learning environments: Proof of concept. *Australasian Journal of Educational Technology, 32*(6), 58-71. http://dx.doi.org/10.14742/ajet.3060