JOURNAL OF LEARNING ANALYTICS

Volume 8(3), 28–44. https://doi.org/10.18608/jla.2021.7161

Can Strategic Behaviour Facilitate Confusion Resolution? The Interplay Between Confusion and Metacognitive Strategies in Betty's Brain

Yingbin Zhang[1], Luc Paquette[2], Ryan S. Baker[3], Jaclyn Ocumpaugh[4], Nigel Bosch[5], Gautam Biswas[6], Anabil Munshi[7]

Abstract

Confusion may benefit learning when it is resolved or partially resolved. Metacognitive strategies (MS) may help learners to resolve confusion when it occurs during learning and problem solving. This study examined the relationship between confusion and MS that students evoked in Betty's Brain, a computer-based learning-by-modelling environment where elementary and middle school students learn science by building causal maps. Participants were sixth graders. Emotion data were collected from real-time observations by trained researchers. MS and task performance information were determined by analyzing the action logs. Pre- and post-tests were used to assess learning gains. The results revealed that the use of MS was a function of the state of student confusion. However, confusion resolution was not related to MS behaviour, and MS did not moderate the effect of confusion on student task performance in Betty's Brain or on learning gains.

Notes for Practice

- Does confusion facilitate or hamper learning? That may depend on whether it is resolved or not, but it is not clear what factors influence confusion resolution.
- This study examined the interplay between metacognitive strategies (MS) and confusion. It found that confusion was accompanied by self-regulated behaviour, suggesting that creating situations that may evoke confusion for students may promote learning regulation. However, we could not establish that such situations would necessarily result in enhanced learning.
- In order to use confusion to promote learning, instructors need to provide students with support to help them resolve confusion. This study found that the use of MS did not improve confusion resolution. Thus, providing students with only MS support may be insufficient to take the advantage of confusion. Student cognitive skills and motivation must also be considered.
- Confusion rarely occurred naturally in this study. Even within the high-confusion group, on average, only 6% of the emotion observed was confusion. Thus, instructors may need to create activities that spark student confusion if they want to use confusion to enhance learning, while also ensuring sufficient support for confusion resolution.

Keywords

Confusion, confusion resolution, metacognitive strategy, cognitive disequilibrium

Submitted: 29/04/20 — **Accepted:** 09/03/21 — **Published:** 01/10/21

Corresponding author [1]Email: yingbin2@illinois.edu Address: Department of Curriculum & Instruction, University of Illinois at Urbana-Champaign, 1310 S. Sixth Street, Champaign, IL 61820, USA. ORCID ID: https://orcid.org/0000-0002-2664-3093

[2]Email: lpaq@illinois.edu Address: Department of Curriculum & Instruction, University of Illinois at Urbana-Champaign, 1310 S. Sixth Street, Champaign, IL 61820, USA. ORCID ID: https://orcid.org/0000-0002-2738-3190

[3]Email: ryanshaunbaker@gmail.com Address: Graduate School of Education, University of Pennsylvania, 3700 Walnut St., Philadelphia, PA 19104, USA. ORCID ID: https://orcid.org/0000-0002-3051-3232

[4]Email: ojaclyn@upenn.edu Address: Penn Center for Learning Analytics, University of Pennsylvania, 3700 Walnut St., Philadelphia, PA 19104, USA.

[5]Email: pnb@illinois.edu Address: School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820, USA. ORCID ID: http://orcid.org/0000-0003-2736-2899

[6]Email: gautam.biswas@vanderbilt.edu Address: Department of EECS/ISIS, Vanderbilt University, 1025 16th Ave. South, Nashville, TN 37212, USA. ORCID ID: https://orcid.org/0000-0002-2752-3878

[7]Email: anabil.munshi@vanderbilt.edu Address: Institute for Software Integrated Systems, Vanderbilt University, 1025 16th Ave S., Nashville, TN 37212, USA. ORCID ID: https://orcid.org/0000-0002-3366-9048

1. Introduction

Previous research has suggested that confusion may promote robust learning (Craig, Graesser, Sullins, & Gholson, 2004; D'Mello, Lehman, Pekrun, & Graesser, 2014; D'Mello & Graesser, 2014a). Whether confusion facilitates or hampers learning depends on whether it is resolved or not (D'Mello & Graesser, 2014b). If a computer-based learning environment (CBLE) can intentionally cause confusion among learners, and then provide scaffolding to promote confusion resolution, it may successfully promote deeper understanding of the material being learned (Lehman & Graesser, 2014). The scaffolding should have the potential to help learners understand and invoke strategies to overcome the confusion. However, it is unclear what strategies support confusion resolution.

Metacognitive strategies (MS) — strategies for monitoring and controlling cognitive processes (Pintrich, Smith, Garcia, & McKeachie, 1993) — are an essential component of self-regulated learning (e.g., Efklides, 2011) and a powerful predictor of academic performance (e.g., Ohtani & Hisasaka, 2018). Recent research has raised the possibility that MS may play an important role in confusion resolution (Di Leo, Muis, Singh, & Psaradellis, 2019; Muis, Psaradellis, Lajoie, Di Leo, & Chevrier, 2015), and researchers have also suggested promoting self-regulation strategies to help learners take advantage of their confused states in CBLEs to become better learners (Arguel, Lockyer, Kennedy, Lodge, & Pachman, 2019; Arguel, Lockyer, Lipp, Lodge, & Kennedy, 2017). However, to date, there is little direct evidence to show that self-regulation and metacognitive strategies can improve confusion resolution.

The research presented in this paper aims to address the gap in the links between confusion and the use of MS by examining whether confusion invokes changes in MS behaviour, and how students go about performing confusion resolution when working with Betty's Brain, an open-ended CBLE. This study indicates that confusion may induce strategic regulation of behaviour, but the MS that students use may not be sufficient for confusion resolution. The rest of this section reviews studies about confusion and MS and discusses their potential relationship.

1.1. Confusion During Learning

1.1.1. What Causes Confusion?

Confusion is an emotion that arises when learners have difficulties during knowledge acquisition and putting the knowledge to use (Pekrun & Stephens, 2012; Silvia, 2010). It has been widely observed across multiple learning environments (e.g., D'Mello, 2013; Rodrigo et al., 2008; Yang, Kraut, & Rose, 2016). According to Mandler's (1990) interruption (discrepancy) theory, confusion follows cognitive disequilibrium, which arises when there is a conflict between new information and an individual's prior knowledge, or when the individual's current knowledge structure cannot assimilate the new information. This notion has been supported by a set of systematic studies (D'Mello et al., 2014; D'Mello & Graesser, 2014b; Lehman, D'Mello, & Graesser, 2012). These studies found that confusion occurred more often in conditions that could cause cognitive disequilibrium. Based on these findings, D'Mello and Graesser (2014a) hypothesized that confusion is the affective signature of cognitive disequilibrium.

Pekrun and Stephens (2012) indicated that cognitive disequilibrium does not initially or inevitably trigger confusion. Learners initially experience surprise and curiosity when new information is incongruent with their knowledge. If this incongruity cannot be resolved, then confusion arises. If the incongruity seems to be impossible to resolve, frustration replaces confusion. Silvia (2010) surmised that events stemming from appraisals of high novelty and low comprehensibility lead to confusion, while events with appraisals of high novelty but high comprehensibility induce interest. Following Silvia's claim, Muis, Chevrier, and Singh (2018) argued that confusion more likely follows surprise when the task is quite complicated. This notion was supported by recent studies (Chevrier, Muis, Trevors, Pekrun, & Sinatra, 2019; Munzar, Muis, Denton, & Losenno, 2020). These studies found that the relative frequency of confusion was higher than curiosity when learners felt that information was novel and complex (Chevrier et al., 2019), and the transition from surprise to confusion was a function of complexity (Chevrier et al., 2019; Munzar et al., 2020).

Learner characteristics have an impact on the extent of confusion they experience during learning. For example, students with higher learning motivation may experience more confusion (Lehman, D'Mello, & Graesser, 2013). Muis, Pekrun et al. (2015) found that beliefs about the complexity of knowledge and the source of knowing predicted student confusion negatively during the process of learning about climate change. Another study they conducted showed that students' task value and academic control (values and perception of control for learning general mathematics and for solving the specific mathematical problem) also negatively predicted student confusion in the context of complex problem solving (Muis, Psaradellis et al., 2015). The association between task value and confusion was mediated by the extent of cognitive disequilibrium (Munzar et al., 2020).

To summarize, discrepant information or novel and complex information causes cognitive disequilibrium. Cognitive disequilibriu`m leads to confusion if it cannot be resolved right away. The effects of the discrepant events on confusion may depend on learner characteristics.

1.1.2. Effects of Confusion on Learning

The effects of confusion on learning are mixed. On the negative side, confusion has been associated with less frequent use of deep processing strategies and planning strategies (Di Leo et al., 2019; Muis, Psaradellis et al., 2015). Uninterrupted confusion might lead to frustration (D'Mello & Graesser, 2012), and this may cause students to feel lower levels of self-efficacy (Caprara et al., 2008). This, in turn, can lead to lower achievement or even worse, disengagement, which may result in dropping out of courses (Yang et al., 2016).

On the positive side, confusion may promote the use of MS (Di Leo et al., 2019; Muis, Pekrun et al., 2015) and benefit learning (D'Mello et al., 2014; D'Mello & Graesser, 2014b). D'Mello and Graesser (2014a) claimed that confusion "plays a prominent role in learning activities that are pitched at deeper levels of comprehension" (p. 290). The extent to which learners can benefit from confusion may depend on both their skills and their motivation. For instance, one study found that students with higher cognitive ability and drive had higher learning gains resulting from confusion than their peers (Lehman & Graesser, 2015).

1.1.3. Confusion Resolution

The mixed effects of confusion on learning may be related to its resolution. Poor learning outcomes may be partially due to persistent unresolved confusion (Bosch, D'Mello, & Mills, 2013; D'Mello & Graesser, 2014b). However, confusion need not be entirely resolved in order to benfit learning; completely and partially resolved confusion have both been observed to predict learning gains (D'Mello & Graesser, 2014b). These results could help explain the conflicting impact of confusion on learning. When confusion is entirely or partially resolved, learning outcomes improve, but when confusion is entirely unresolved, learners are unable to modify their existing knowledge structure and assimilate the new information. Instead, they learn nothing, and may become frustrated (D'Mello & Graesser, 2012) and doubt their abilities (Caprara et al., 2008). This may be why prolonged confusion was negatively related to learning, while short-term confusion was positively related to learning (Liu, Pataranutaporn, Ocumpaugh, & Baker, 2013). Therefore, it is critical to understand how confusion can be resolved so that intervention and scaffolding can be designed and given to students who are likely to experience sustained confusion.

While this paper discusses confusion resolution, it is important to clarify that confusion resolution does not mean that learners directly resolve their confusion. Instead, what learners tackle is the state of cognitive disequilibrium, which is the cause of the confusion (D'Mello et al., 2014; D'Mello & Graesser, 2014b). In other words, confusion resolution excludes situations where learners simply ignore their confusion and engage in the next task. D'Mello et al. (2014) indicate that confusion can be resolved if a) learners have the knowledge and skills to resolve it, or b) the learning environment provides scaffolding to help them resolve it. The first prerequisite is supported by findings in D'Mello and Graesser's (2014b) study, where learners with higher ACT (a standardized college entrance test) scores were more likely to partially resolve their confusion in learning the functioning of everyday devices, such as an electric bell and a toaster. These prerequisites suggest that confusion resolution requires special skills, and simply attempting to resolve confusion without the necessary skills is insufficient to lead to its resolution (Lehman & Graesser, 2015).

1.2. Metacognitive Strategies

1.2.1. Metacognitive Strategies, Knowledge, and Experiences

Metacognitive strategies (MS) refer to the deliberate use of cognitive skills to regulate cognition (Efklides, 2008). Generally, the components of MS include goal setting, planning, self-monitoring, self-control, and self-evaluation (Dent & Koenka, 2016). MS, metacognitive knowledge, and experiences are intercorrelated but distinct (Efklides, 2008). It is worth distinguishing the three facets of metacognition to avoid ambiguity. Metacognitive knowledge is declarative knowledge and belief about the world, such as tasks, goals, self, others, and cognitive processing and skills (Flavell, 1979). Metacognitive experiences are what a person feels about their thoughts or is aware of during task processing, such as the feeling of confidence and judgment of knowing (Efklides, 2006).

1.2.2. Impact of Metacognitive Strategies on Learning

Many studies have found the benefits of MS for learning. In an early literature review, MS was one of the most powerful predictors of academic performance (Wang, Haertel, & Walberg, 1990). Some subsequent studies did not find a strong association between MS and learning (e.g., Hargrove & Nietfeld, 2015; Muis, Pekrun et al., 2015), but a recent meta-analysis, which focused on studies in elementary and secondary school, indicated that the correlation between learning and MS differed significantly (though not by a large magnitude) across specific components of MS, disciplines, grades and how learning and

MS are measured (Dent & Koenka, 2016). For example, the average Pearson correlations between MS and learning were r = 0.21 in math, 0.23 in English/language arts, 0.26 in science, and 0.34 in social studies. The results were replicated in another meta-analysis (Ohtani & Hisasaka, 2018), which also found that the relationship between MS and learning was moderated by disciplines, grades, and the measurement approach of MS.

1.3. Confusion and Metacognitive Strategies

Nelson and Narens' (1990) two-level metacognitive system provides a helpful framework for understanding the relationship between cognitive disequilibrium and MS. The system is about overall cognitive processing. The two levels are the *object-level* and the *meta-level*. Cognition about the external world is in the object-level, and monitoring and evaluating cognition is at the meta-level. The meta-level receives information from the object-level to monitor cognitive activities about the external world and send instructions to the object-level to control these cognitive activities. The controlling function of the meta-level manifests as the use of MS (Efklides, 2006). Under the two-level metacognitive system, learners need to activate MS to resolve confusion. When learners encounter discrepant information, cognitive disequilibrium may appear at the object-level. Information about the object-level flows to the meta-level, and the meta-level detects the inconsonant state at the object-level. To restore cognitive equilibrium, the meta-level must modify cognitive activities at the object-level through the controlling function. In other words, learners should behave according to MS so that they can integrate the discrepant information and their existing knowledge model.

The metacognitive and affective model of self-regulated learning (MASRL; Efklides, 2011) is also useful for understanding the association between confusion and the use of MS. The MASRL model emphasizes the interactions of affect, motivation, and metacognition in self-regulated learning (SRL). Under this model, cognitive disequilibrium during task processing leads to affective reactions, such as surprise and confusion. At the same time, cognitive disequilibrium also contributes to metacognitive experiences, such as the feeling of difficulty and the estimate of effort and time. The metacognitive experiences and affective states, in turn, may trigger decisions about conducting MS behaviour. In other words, cognitive disequilibrium may indirectly influence the use of MS. Therefore, it is reasonable to expect that learners change MS behaviour when they are confused.

Taking these theories together, changes in MS behaviour may co-occur with confusion because both follow cognitive disequilibrium. In order to make cognition return to the state of equilibrium, learners need to control cognitive activities by invoking MS. However, the interplay between confusion and MS needs to be further researched.

1.4. Current Study

The current study combined the observations of emotion and the action logs of student activities in a CBLE, Betty's Brain, to answer four research questions examining the relationship between confusion and MS. As described above, confusion may be accompanied by changes in MS behaviour. Additionally, confusion is experienced as an unpleasant emotion (Baker, D'Mello, Rodrigo, & Graesser, 2010) and may motivate learners to regulate their cognition to resolve it. This motivated the first research question (RQ1): *Is the frequency of MS behaviours different when students are confused compared with when they are not confused*?

The use of MS may be necessary for learners to restore cognition to an equilibrium state and resolve confusion. To study this hypothesis, the current study investigated the second question (RQ2): *Is confusion resolution related to increases in the use of MS*?

Confusion can benefit learning when it is resolved or partially resolved (D'Mello & Graesser, 2014b; Lehman & Graesser, 2014). If MS contributes to confusion resolution, the effect of confusion on learning may depend on student abilities to apply MS. For students experiencing the same level of confusion, those who use more MS may have better task performance within the Betty's Brain environment and pre–post learning gains than those who use MS less often. Thus, the third and fourth research questions were as follows: (RQ3) *Do MS moderate the relationship between confusion and task performance*? (RQ4) *Do MS moderate the relationship between confusion and learning gain*?

2. Betty's Brain

Betty's Brain is an open-ended CBLE (Biswas, Segedy, & Bunchongchit, 2016; Leelawong & Biswas, 2008). Students learn about scientific phenomena, such as climate change, thermoregulation, or river ecosystems, by teaching a virtual pedagogical agent, named Betty. Students teach Betty by building a causal map of the scientific phenomenon, in which a causal (*cause-and-effect*) relationship is represented by a pair of concepts connected by a directed causal link (see Figure 1). To build this map, students have access to hypermedia resource pages (*Science Book* in Figure 1) on relevant scientific concepts. Students can evaluate their causal modelling progress by asking Betty to take quizzes graded by a mentor agent, Mr. Davis, or by querying Betty with *cause-and-effect* questions related to what she has been taught so far. Betty's quiz grades (correct and

incorrect answers), along with her explanations of her answers to specific questions, can help the student track her progress, and, in turn, their own progress. In more detail, looking at Betty's correct and incorrect answers, students can identify problems (e.g., incorrect links or missing links) in their causal map. They can then improve their understanding of the topic by re-reading the science book and tracking the explanations Betty provides to correct their perceived problems with their causal map.

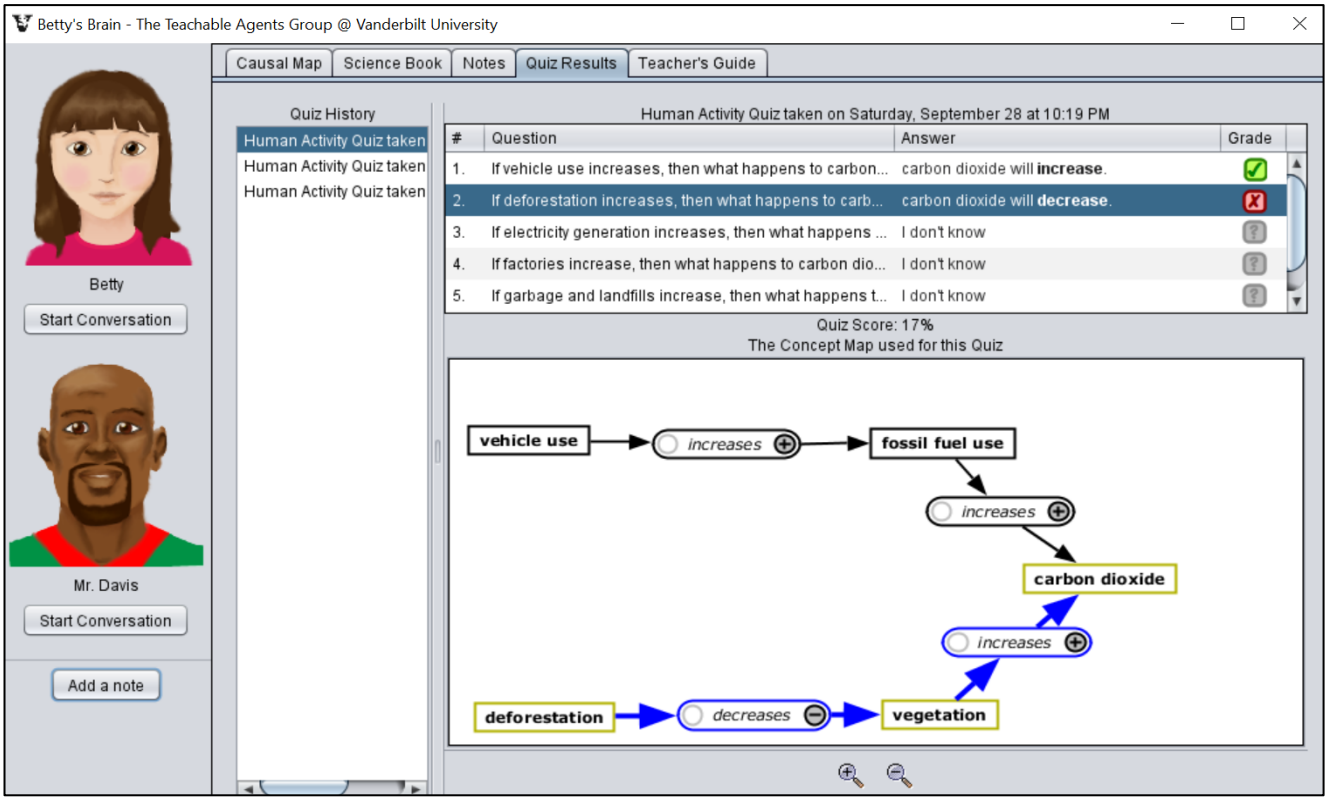

Figure 1. Screenshot of quiz results and the link chain that Betty used to answer the quiz question.

While using Betty's Brain, student activities can be grouped into six primary categories: 1) *read* (read a page in the resources), 2) *take notes* (create, view or edit a note), 3) *edit* (edit the causal map, including adding or deleting a concept, adding, deleting or modifying a causal link, marking a causal link as correct or incorrect), 4) *query* (ask Betty a cause-and-effect question based on the concepts and links on her map so far), 5) *quiz* (assess the state of the map by having Betty take a quiz), and 6) *explain* (ask Betty to explain her answer to a cause-and-effect query or check the chain of links she used to answer a quiz question, in order to probe Betty's reasoning).

In the current study, students learned about climate change by building a causal map to teach Betty. The expert map (a map with all correct causal links) on this topic contains 22 scientific concepts and 27 causal links. For details about this topic, see (Kinnebrew, Segedy, & Biswas, 2017, p. 147).

3. Methods

3.1. Participants and Procedures

This study, run in the 2016–2017 school year, included 93 sixth-grade students from four classrooms in an urban public school in Tennessee. There were four classrooms with 23 to 24 students in each classroom. Four or five students were seated at a table, but they worked individually on separate laptops. Overall, the school had 700 students in grades 5–8. Within this school, around 40% of the students were from underrepresented minorities, and around 8% of the students enrolled in the free and reduced lunch program. The study lasted seven school days. On day 1, students spent 30–45 minutes on completing a paper-based pre-test that assessed their knowledge of climate change and causal relationships. On day 2, they received a 30-minute training about how to use Betty's Brain. In the next four days, they spent 45 to 50 minutes per day constructing the causal map to teach Betty. On the final day, students completed a post-test identical to the pre-test.

3.2. Observations of Emotion

While students were working on Betty's Brain, two trained observers recorded their affective states via the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; Ocumpaugh, Baker, & Rodrigo, 2015), which uses holistic observations that consider the context of students actions, behaviours, speech, and facial/bodily expressions, in line with how Planalp, DeFrancisco, and Rutherford (1996) describes the process humans normally use to identify affective states. That is, rather than relying exclusively on whether or not a student's eyebrows undergo specific actions when they are confused (since scowls can also indicate engaged concentration), coders pay attention to both facial expressions and the ways in which the student is interacting with their environment. This technique, which contextualizes a constellation of features rather than relying on a fixed set of facial expressions to determine student affect, is also in line with research by Cowen et al. (2019). The protocol was enforced by the Android App HART (Ocumpaugh et al., 2015), which adds a timestamp to each observation recorded. With the timestamps, the observations could be aligned with student action logs in the Betty's Brain environment.

The observers used a momentary time sampling method (Meany Daboul, Roscoe, Bourret, & Ahearn, 2007), where they coded students individually in a pre-determined order. This sampling method is intended to result in a representative sample of emotions and behaviour. The observers recorded both student emotions (i.e., boredom, confusion, delight, engaged concentration, and frustration) and behaviour (i.e., on-task, on-task conversation, off-task), though the current study only considers the emotion observations. Observers recorded the first emotion and behaviour they saw and had up to 20 seconds to decide the student affective state. At the beginning of the study, the first observer trained the second observer in using BROMP. Once the second observer had been trained, both observers coded the same students at the same time to calculate inter-rater reliability and validate their coding. The inter-rater reliability was acceptable (Cohen's Kappa > 0.60). Following the training period, both observers coded separately to maximize the number of observations that could be collected during the study. There were 326 recorded observations of confusion as well as observations of boredom (233), delight (157), engaged concentration (4182), and frustration (251). On average, 53 observations were made for each student.

3.3. Metacognitive Strategies

We analyzed the action logs of students working on Betty's Brain to measure MS behaviour based on coherence analysis (CA; Segedy, Kinnebrew, & Biswas, 2015). CA interprets learner behaviour according to the relationship between actions (e.g., read and edit). Two ordered actions, x and y, are interpreted as being coherent if the second action, y, utilizes information generated by the first action, x, i.e., x supports y. For instance, in Figure 1, the quiz results inform the student that the quiz question about the relation between deforestation and carbon dioxide was answered incorrectly. After viewing these quiz results, if students read the resource pages that contain information about the relationship between the two concepts, the action of viewing quiz results supports the action of reading that resource page, and, therefore, these two actions are labelled as coherent. X and y need not be consecutive actions but they must occur within a pre-determined time interval. The time interval is chosen with the assumption that students are going to remember the information generated by action x when they execute action y. The choice of the interval is somewhat arbitrary, but in line with prior work on Betty's Brain (e.g., Segedy et al., 2015), we set the interval length for the current study to 5 minutes.

Prior work hypothesizes that action coherence is an indicator of task understanding and the use of metacognitive strategies (Kinnebrew et al., 2017). Indeed, if a student performs two ordered coherent actions, it suggests that they monitored previous cognitive processes (*the first action*, e.g., viewing quiz results), showed awareness by paying attention to the information generated by previous cognitive processes (*the results of the first action*, e.g., the link chain between deforestation and carbon dioxide was wrong), and regulated the later cognitive processes (*the second action*, e.g., reading resource pages about deforestation and carbon dioxide) based on the previously generated information. Therefore, performing coherent actions may suggest the use of MS by the student.

Five behavioural metrics were identified as measures of MS. Each metric was calculated as the frequency of a particular kind of action per minute so that these metrics were comparable across students, given that the total time of each student's work on Betty's Brain varied (for example, some students finished their map building task earlier than the others). Metric 1, *Quizzing frequency* can be linked to metacognitive monitoring, i.e., how students are progressing in their causal modelling task. Metrics 2–5 are directly based on CA. Incoherent actions (INC) that do not satisfy the four CA metrics were also examined in the analysis for RQ1 and defined here. We do not assume that the INC metrics represent suboptimal or bad strategies. Rather, they were mainly used to enhance our understanding of the differences in the use of MS when students were confused and not confused. Thus, INC metrics were not used in the analyses for RQs 2 to 4. The five CA metrics are as follows:

1. *Quizzing frequency*: the number of times the student had Betty take a quiz, per minute. This variable reflected how often students evaluated the accuracy of the map they created to teach Betty, and was a reflection of their checking progress about their own understanding of climate change. This metric was not directly a CA measure, but, as discussed, taking quizzes relates to self-monitoring strategies, where students are evaluating their current progress (e.g.,

Cicchinelli et al., 2018). Since self-monitoring is a metacognitive process, we included this metric in our list of CA metrics.

2. *Frequency of coherent viewing*: the number of viewing quiz results actions, per minute, that were coherent with later actions. This variable measured how often students used the quiz results to influence further actions. For example, a quiz view would be coherent if the student marked all of the links associated with a correct answer to a quiz question as correct. Coherent viewing might indicate self-monitoring behaviour because the information generated was used by the student in subsequent actions.

   The *frequency of incoherent viewing* is the number of viewing quiz result actions per minute that do not support later actions within the specified time window. For instance, in the Figure 1 example, the viewing quiz results action would be coherent if a student did one or more of the following actions after viewing the quiz results: 1) read the resource pages that contained information about the relationship between deforestation and carbon dioxide, 2) edited the links between the two concepts, or 3) deleted the link because they assumed it was incorrect, or they marked it as *maybe wrong*. In contrast, if the student did none of these actions, the viewing action would be incoherent.
3. *Frequency of coherent editing*: the number of coherent map edits per minute. It measured how often students edited the concept map (adding, deleting, or revising causal links) based on acquired information. This acquired information could come from reading the resource pages, or from quiz viewing actions. In other words, coherent editing might indicate informed map generation and refinement behaviour.

   The *frequency of incoherent editing* is the number of map edits per minute that are not based on quiz results or reading text related to the edited links. For example, in the Figure 1 example, after viewing the quiz results, if a student edited links between deforestation and carbon dioxide, this edit would be coherent. In contrast, if the student edited links between vehicle use and carbon dioxide, this edit would be incoherent because the quiz results indicated that the link between vehicle use and carbon dioxide was correct. It should be noted that a coherent edit is not necessarily correct, and an incoherent edit is not necessarily incorrect.
4. *Frequency of coherent reading*: the number of coherent reading page actions per minute. This variable measured how often students intentionally sought relevant information to improve their understanding based on the quiz results. If the quiz results suggested some links were wrong, the coherent reading meant that pages being read contained information about the wrong links. If the quiz results showed that all links were correct, the page being read contained information about new links not in the current causal map. Coherent reading might represent self-control behaviour because the page being read was contingent on the quiz results.

   The *frequency of incoherent reading* is the number of reading actions, per minute, not based on quiz results. For example, in Figure 1, after viewing the quiz results, if a student read the resource pages that contained information about the relationship between deforestation and carbon dioxide, that reading action would be coherent. By contrast, if the student read resource pages that did not contain any information about concepts in the links indicated by the quiz results as incorrect or possibly incorrect, the reading action would be incoherent.
5. *Frequency of coherent marking*: the number of coherent actions of marking (i.e., labelling) a link on the concept map, per minute. In addition to marking these links as "correct" or "maybe wrong," students could also delete these marks. This variable reflects how often, based on the quiz results, students understood what links on their map were correct or possibly incorrect, and annotated them accordingly. Coherent marking again represents constructive monitoring behaviour because the marking action translates quiz results into systematic checking of the causal maps. The mark could remind students of which parts of their understanding were incorrect, correct, or uncertain.

   The *frequency of incoherent marking* is the number of marking actions per minute not based on quiz results. For instance, in the Figure 1 example, after viewing the quiz results, if a student marks links between deforestation and carbon dioxide as "may be wrong," the marking action would be coherent. In contrast, if the student marks any link between the two concepts as "correct," the marking action would be incoherent.

Internally, the Betty's Brain system keeps a record of the resources pages that contain information about each of the causal links required to build the correct map. This information is used to establish links between quiz questions, the causal links required to answer a particular quiz question, and the corresponding resource pages relevant for each link. This information helps us compute the coherence of reading resource pages, editing causal links, and marking links with respect to prior student actions. The IDs of prior actions (every action is assigned a unique ID in Betty's Brain) that are coherent with the current action are also recorded. Based on these IDs, we distinguished coherent and incoherent viewing of quiz results actions.

In line with prior research on Betty's Brain (e.g., Segedy et al., 2015), we excluded some actions that were too brief, including viewing (quiz results) actions that were less than 2 seconds long, and reading (resource pages) actions that were less than 10 seconds long. These actions were discarded because they were likely to indicate only shallow engagement with that

part of the system. For example, short reading actions might indicate that students were just flipping through the resource pages quickly without reading the content.

3.4. Task Performance

The indicator of task performance in Betty's Brain was the final causal map score. It was the difference between the number of correct causal links and incorrect causal links in the student's final map. The correct causal links were those that appeared in an expert map on the climate change topic (students could not see this map).

3.5. Knowledge Tests

Learning gains were measured by the pre-test and post-test. These tests were identical in both questions and forms. The test assessed student knowledge of the science topic, climate change, and their causal reasoning abilities. It contained multiple-choice questions and short-answer questions. The appendix displays an example of a multiple choice question and a short-answer question. The answer to each short-answer question consisted of a fixed number of successive causal links from an expert-defined map. Students could score a maximum of 7 points for the multiple-choice questions and 9 points for the short-answer questions. The total maximum score attainable was 16 points.

3.6. Analyses

3.6.1. Analyses for RQ1

After aliging observation data with the action logs, an emotion observation matched a subset of the action logs in Betty's Brain. CA metrics derived from the action logs within the duration of an emotion were regarded as measures of MS behaviour during this emotion. Although the field observations for emotion recording were conducted using a 20-second time window, we analyzed the data with a window that included 30 seconds before and 30 seconds after the observation of the emotion. This window was based on prior research that found the average duration of emotions during learning is approximately 30 seconds (D'Mello & Graesser, 2012), and because it is impossible to know whether an observation was collected closer to the beginning or the end of a learner's emotion. For this study, CA and INC metrics were calculated for 326 observations of confusion and 4,823 observations of other emotions.

The Shapiro-Wilk test indicated that the normality assumption for the CA and INC metrics was violated ($p < 0.001$ for all metrics). Therefore, the robust independent *t*-test with 10% trimmed means (a robust measure of means that ignores the top and bottom 10% of data; Wilcox, 2011) was performed to compare CA and INC metrics during confusion and other affective states. The robust test is satisfactory even when normality and homoscedasticity assumptions are violated (Wilcox, 2011). The Benjamini-Hochberg correction was applied to control for false discovery rate (FDR). This correction adjusts the α value rather than the p value, so we only marked a result significant if its p value was lower than its adjusted α value. The explanatory measure of effect size ξ (i.e., "xi"), a robust version of Cohen's d, was estimated (Wilcox & Tian, 2011). A rule of thumb for ξ is that values of 0.10, 0.30, and 0.50 correspond to small, medium, and large effect sizes R (Mair & Wilcox, 2020). The analysis was implemented within the WRS2 package in R (Mair & Wilcox, 2020).

3.6.2. Analyses for RQ2

If confusion (i.e., the cognitive disequilibrium) is resolved, the learner's affective state is expected to transition to engaged concentration (D'Mello & Graesser, 2012; Liu et al., 2013) or delight. Otherwise, confusion continues, and persistent confusion can lead to frustration or boredom (Botelho et al., 2018; D'Mello & Graesser, 2012). Prolonged confusion is negatively related to learning (Liu et al., 2013). Therefore, if an observation of confusion is followed by an observation of engaged concentration or delight ("confusion → engaged concentration/delight"), this may represent confusion that was resolved. In contrast, if an observation of confusion is followed by an observation of confusion, frustration, or boredom ("confusion → confusion/frustration/ boredom"), this may represent confusion that was unresolved.

The current study referred to a pair of successive affect observations for the same student as an affect sequence, which we categorized based on valence. Specifically, positive affect sequences referred to "confusion → engaged concentration/delight," while negative affect sequences referred to "confusion → confusion/frustration/boredom." The affect sequences whose intervals were longer than three minutes were discarded because students might experience another affective state between two successive observations of affective states if the interval between the two observations was too long (e.g., Botelho et al., 2018; D'Mello & Graesser, 2012). Three minutes was used as the cut-off value because it corresponded to the average gap

between one observation and the next.[1] In order to answer RQ2, the robust *t*-test with 10% trimmed means was applied to compare CA metrics during confusion in positive affect sequences and negative affect sequences. As in RQ1, we applied the Benjamini-Hochberg FDR correction.

3.6.3. Analyses for RQ3 and RQ4

The five CA metrics were calculated for each student across their entire action log. Principal component analysis (PCA) with varimax rotation was applied to extract a factor from the five CA metrics, and the factor score was used as an overall indicator of MS. Multicollinearity was tested to examine whether the five metrics were suitable for PCA. Table 1 displays the correlation matrix. Seven of the ten Pearson product-moment correlations were greater than 0.4. Bartlett's test of sphericity indicated that sufficient correlations existed among the CA metrics ($\chi^2 / df = 18.80$, $p < 0.001$).

Table 1. Correlations Among Coherence Analysis Metrics

	1	2	3	4
1. Quizzing	–			
2. Coherent viewing	0.53	–		
3. Coherent editing	0.39	0.57	–	
4. Coherent reading	0.66	0.69	0.60	–
5. Coherent marking	0.20	0.55	0.25	0.41

The Kaiser-Meyer-Olkin (KMO) overall statistic was 0.76, indicating that a high proportion of variance in the CA metrics might be caused by underlying factors, and all single KMO values were greater than 0.5 (Table 2). Cronbach's alpha for these metrics was 0.83, indicating that the internal consistency of the CA metrics was good. Overall, the results suggested that these metrics were suitable for PCA. Table 2 displays the results of PCA.

Table 2. Results of the Principal Component Analysis

M (SD)	MSA	Loading	R^2
Quizzing	0.75	0.73	54%
Coherent viewing	0.78	0.88	77%
Coherent editing	0.80	0.74	55%
Coherent reading	0.76	0.89	79%
Coherent marking	0.69	0.59	35%
Total explained variance			60%

Student confusion scores were calculated as the ratio of the number of confusion observations to the total number of affect observations. Students were divided into high and low confusion groups, based on whether their confusion scores were higher or lower than the median, as well as high and low MS groups, based on whether their MS scores were higher or lower than the median. Data from ten students were discarded because either their pre-test or post-test scores were missing. Finally, there were 23 in the high-MS and high-confusion group, 20 in the high-MS and low-confusion group, 18 in the low-MS and high-confusion group, and 22 in the low-MS and low-confusion group.

The normality assumptions for the final map scores and the test scores were violated (the p-values from the Shapiro-Wilk test were 0.001 for the final map scores and 0.006 for the test scores). Thus, a robust 2-way ANOVA with 10% trimmed means (using the WRS2 package in R; Mair & Wilcox, 2020) was conducted to examine the relationship among MS, confusion, and map scores. The robust ANOVA produces a χ^2-distributed test statistic, Q. The p-value is calculated by comparing Q with an adjusted critical value, so the degrees of freedom are not reported (Mair & Wilcox, 2020). A three-way non-parametric repeated

[1] The gap varies because the time that observers needed to decide an observation was not constant. For example, if one student showed apparent confusion (e.g., scrunching up the nose and forehead and pursing the lips), observers might only need five seconds to conclude that this student's emotion was confusion; in another observation where this student might just purse the lips slightly, observers might need 15 seconds to make a decision. Classroom events might also cause pauses in observation.

ANOVA (using the nparLD package in R; Noguchi, Gel, Brunner, & Konietschke, 2012) was conducted to test the relationships among MS, confusion, and test scores. For the repeated ANOVA, the test time was the within-subject factor, and confusion and MS were between-subject factors. The non-parametric repeated ANOVA produces an ANOVA-type statistic (*ATS*), which can be approximated by the *F* distribution (Brunner & Puri, 2001).

4. Results

RQ1: Is the Frequency of MS Behaviour Different During and Outside the Duration of Confusion?

Table 3 shows the results of the robust *t*-test. The frequencies of coherent viewing were higher during confusion than during other affective states. This indicates that the information that students collected while they were confused was more likely to be used to support later actions. Students did not decrease the frequency of incoherent viewing when they were confused.

Coherent editing was the same when students were confused and not confused. However, they did less incoherent editing during confusion. In other words, when students were confused, they were less likely to edit a causal link without collecting any (new) information about this link. Both the frequencies of coherent and incoherent reading were higher when students were confused than not confused. This indicates that students might simply read more text when they are confused.

There were no statistically significant differences for quiz-taking actions, coherent and incoherent link marking actions. Overall, these results suggest that learners changed their use of MS when confused, primarily doing more reading and using the information from quiz results more frequently.

Table 3. Coherence Analysis Metrics During and Outside Confusion

M (SD)	During	Outside	ξ
Quizzing	0.13 (0.37)	0.16 (0.41)	0.06
Coherent viewing	0.52 (1.11)	0.26 (0.71)	0.25*
Incoherent viewing	0.33 (0.94)	0.24 (0.70)	0.07
Coherent editing	0.26 (0.56)	0.32 (0.67)	0.06
Incoherent editing	0.05 (0.22)	0.12 (0.47)	0.12*
Coherent reading	0.38 (0.66)	0.26 (0.54)	0.14*
Incoherent reading	0.23 (0.94)	0.04 (0.92)	0.11*
Coherent marking	0.05 (0.28)	0.04 (0.36)	0.05
Incoherent marking	0.06 (0.31)	0.08 (0.42)	0.05

Note: *$p <$ adjusted α.

RQ2: Is Confusion Resolution Related to Increases in MS Behaviours?

Table 4 displays the results of the robust *t*-test. There was no difference in the frequency of CA metrics during confusion in negative and positive affect sequences ($p >$ adjusted α for all metrics). Recall that an observation of confusion might be resolved confusion if it was followed by observations of positive affect, and unresolved confusion if followed by observations of negative affect. Thus, the result indicates that confusion resolution may not be related to more MS behaviour.

Table 4. Coherence Analysis Metrics in Negative and Positive Affect Sequences

M (SD)	Negative (N = 26)	Positive (N = 63)	ξ
Quizzing	0.23 (0.43)	0.08 (0.27)	0.29
Coherent editing	0.35 (0.56)	0.25 (0.59)	0.18
Coherent reading	0.19 (0.40)	0.32 (0.69)	0.07
Coherent marking	0.04 (0.19)	0.05 (0.37)	0.19
Coherent viewing	0.65 (1.32)	0.59 (1.50)	0.15

RQ3: Do MS Moderate the Relationship between Confusion and Task Performance?

Table 5 displays the final map scores of the different groups. The high-MS group's average map scores were more than twice as high as the low-MS group's. Within the high-MS group, students with high confusion had higher map scores than those

with low confusion. In contrast, this difference was reversed within the low-MS group. Nevertheless, the robust ANOVA showed that there was no interaction effect between confusion and MS ($Q = 0.08$, $p = 0.766$); the main effect of confusion was also not significant ($Q = 0.08$, $p = 0.776$). The main effect of MS was significant ($Q = 37.23$, $p = 0.001$).

Table 5. Final Map Scores

M (SD)	Low Confusion	High Confusion
Low MS	7.40 (8.30)	6.20 (7.65)
High MS	16.05 (8.94)	17.75 (5.89)

RQ4: Do MS Moderate the Relationship between Confusion and Learning Gains?

Table 6 displays the pre-test and post-test scores. In all groups, students generally got better scores in the post-test than in the pre-test, and the effect size was large ($\xi >= 0.5$). ξ was greater in the high-confusion group than in the low-confusion group, for both the high-MS group and the low-MS group. However, the difference in pre–post gain between the low-confusion and high-confusion groups was larger for students with high-MS (difference in ξ was 0.21) than for low-MS students (difference in ξ was 0.09). This finding suggested that MS might moderate the effect of confusion on learning. Indeed, the three-way non-parametric repeated ANOVA revealed a marginally significant interaction among time, confusion, and MS ($ATS = 3.11$, $p = 0.078$). To determine whether the interaction between confusion and time was different within different MS groups, we conducted a two-way non-parametric repeated ANOVA (confusion × time) within each MS group. The results showed an interaction between time and confusion within the high-MS group ($ATS = 6.14$, $p = 0.013$) but not in the low-MS group ($ATS = 0.09$, $p = 0.762$). Nevertheless, it should be noted that the moderation effect of MS on the relationship between confusion and learning was small and only marginally significant.

The three-way repeated ANOVA showed that the main effects of time and MS were significant ($ATS = 98.18$ for time and $ATS = 16.47$ for MS; both $p < 0.001$). Students performed better on the post-test than the pre-test, and students with high MS had higher test scores than those with low-MS. The main effect of confusion was not significant ($ATS = 1.89$, $p = 0.169$). No significant interaction was found between time and MS ($ATS = 1.51$, $p = 0.218$), time and confusion ($ATS = 2.53$, $p = 0.112$), or MS and confusion ($ATS = 0.00$, $p = 0.950$).

Table 6. Pre-Test and Post-Test Scores

MS	Confusion	M (SD)		ξ
		Pre-test	Post-test	
Low	Low	5.84 (2.56)	8.23 (3.19)	0.50**
	High	4.97 (3.00)	7.31 (2.41)	0.59***
High	Low	8.18 (2.67)	10.28 (2.5)	0.50***
	High	6.46 (3.12)	10.41 (3.42)	0.71***

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

5. Discussion

Confusion can benefit learning when it is partially or entirely resolved (e.g., D'Mello & Graesser, 2014b). To resolve confusion, learners need necessary knowledge and learning strategies or support from the environment (D'Mello et al., 2014). What specific strategies may contribute to confusion resolution is unclear. To address this gap, the current study examined the relationship between a set of metacognitive strategies (MS) and confusion in the context of Betty's Brain.

5.1. MS and Confusion

Results for RQ1 showed that the use of MS differed depending on whether the student was experiencing confusion. Specifically, this study found that coherent quiz-results viewing, and coherent and incoherent reading, were more common among students when they were confused, and incoherent editing was less frequent when they were confused. It is understandable that coherent viewing and reading were more frequent during confusion. Both might serve to help learners to find the discrepant information, which is the cause of cognitive disequilibrium (Mandler, 1990). For example, in this study, the correct link between vehicle use and carbon dioxide is "vehicle use increases fossil fuel use, which in turn increases carbon dioxide." Students might think vehicle use could directly increase carbon dioxide and miss the intermediate concept, fossil

fuel use, and thus, they would add a link of "vehicle use increases carbon dioxide." When the quiz result indicated the link was wrong, they might feel confused. This unpleasant emotion might drive them to investigate why the link was wrong, and thus they would look for and read the resource pages to discover the relationship between vehicle use and carbon dioxide. A decrease in incoherent edits during confusion is also reasonable. Cognitive disequilibrium may induce more effortful and analytic cognitive processing (Efklides, 2011). Thus, when students were confused, they might be more cautious and stop adding, deleting, or revising causal links without attempting to gain more understanding of the links by reading the resources.

One surprising finding is that incoherent reading increased when students were confused. This result seems to contradict with the MASRL model, which asserts that cognitive disequilibrium indirectly facilitates the deliberate use of MS via triggering affective reactions and metacognitive experiences (Efklides, 2011). However, if we put together the changes in coherent reading and viewing as well as incoherent reading and edits, an interesting pattern emerges: when confused, students read more text and decrease edits in the causal map. As an unpleasant emotion (Baker et al., 2010), confusion might motivate students to find an explanation for the discrepant information that caused their cognitive disequilibrium (Vogl, Pekrun, Murayama, Loderer, & Schubert, 2019; Vogl, Pekrun, Murayama, & Loderer, 2020). If students were confused by the quiz results, they read resource pages related to the quiz results (increase in coherent reading and quiz-results viewing). If they were confused by the text about a causal link in a resource page, they might read other resource pages that did or did not contain information about the causal link but explain the scientific concepts in the link or related links (we treated the latter as an increase in incoherent reading). As they needed time to search for the explanation, they would not edit links on the causal map until they collected information that helped them with the editing tasks (decrease in incoherent edits). In summary, this behaviour pattern during confusion might not be an efficient strategy in terms of the time spent on reading additional pages, but it may be an indication of trying to get more holistic knowledge on the science topic. It certainly is more constructive behaviour than the trial-and-error edits that some students perform. Therefore, in line with the MASRL model (Efklides, 2011), these results suggest that interruption in cognitive processing may drive learners to regulate their behaviour, in this case to gain more knowledge.

5.2. MS and Confusion Resolution

RQ2 investigated the relationship between confusion resolution and MS at a fine grain-size, i.e., whether there is more MS behaviour during resolved confusion than during unresolved confusion. The confusion followed by engaged concentration or delight might represent resolved confusion, while the confusion followed by confusion, frustration, or boredom might be unresolved (D'Mello & Graesser, 2012). Results showed no significant difference between the frequencies of MS behaviour within the confusion followed by engaged concentration or delight and within the confusion followed by confusion, frustration, or boredom. This suggests that MS behaviour may not be sufficient for resolving the cognitive disequilibrium that underlies confusion. However, it may also be possible that, in some cases, students were cycling to engaged concentration because they were shifting their focus to other scientific concepts and causal relations between these concepts that were less confusing, rather than resolving the cognitive disequilibrium. We cannot verify this hypothesis in our current research, but prior research has found that students might move on without handling confusion (Munzar et al., 2020).

Results for RQ1 suggest that MS may help learners to find the discrepant information; for example, the resource page containing information about the wrong link. However, in order to successfully integrate the information into their mental model, students needed to interpret it correctly, and this task might rely on their reading skills, such as the ability to read and understand the text, to translate information into the causal format, and to focus on the specific content related to the causal link on the page (Segedy, Biswas, & Sulcer, 2014).

5.3. MS, Confusion Resolution, and Learning

RQ3 and RQ4 examined confusion resolution and MS at a coarser granularity. Both the main effects of MS on final map scores and test scores were significant. This is in line with prior research, where students with high map scores did more coherent actions than those with low map scores (Segedy et al., 2014). The significant main effect of test time on test scores indicates that student understanding of climate change increased after using Betty's Brain. The interaction between test time and MS was not significant, so the improvement on understanding was not associated with MS. The improvement on understanding was also not related to confusion because the interactions between test time and confusion were not significant.

However, we found a marginally significant interaction among test time, confusion, and MS. Further analyses showed that the difference in pre–post gain between the low- and high-confusion groups was significant within the high-MS group but not significant in the low-MS group. This different effects of confusion on learning between high- and low-MS groups suggest that confusion resolution may be associated with MS. Students with high MS might be likely to resolve confusion, while students with low MS might be unable to resolve confusion. Resolved confusion can benefit learning (D'Mello et al., 2014;

D'Mello & Graesser, 2014b). Thus, the high-MS group benefited more from confusion than the low-MS group because the former might resolve more confusion than the latter.

Nevertheless, the interaction effect on test scores among time, confusion, and MS was small and only marginally statistically significant. In addition, the interaction effect on the map scores between confusion and MS was not significant. Several reasons might explain these results. First, MS may not be sufficient for resolving the cognitive disequilibrium underlying confusion. Additional cognitive skills may be necessary (Segedy, 2014). Moreover, in terms of self-regulated learning (SRL), motivation may also be necessary for confusion resolution because it plays a critical role in the regulation of emotion and behaviour (Azevedo, Behnagh, Duffy, Harley, & Trevors, 2012; Efklides, 2011). This may also partially explain why prior studies did not find a treatment effect for MS scaffolding on learning in Betty's Brain (e.g., Kinnebrew, Segedy, & Biswas, 2014). Future work should examine comprehensive motivational, cognitive, and metacognitive data to deeply investigate how these factors influence the process of confusion resolution.

Moreover, confusion did not frequently occur in the current study. Even within the high-confusion group, the average proportion of confusion in the affect observations was only 6%. The low frequency of confusion might limit the accumulative effect of confusion resolution. Even if a single confusion resolution benefited learning, the overall effect of confusion on the learning gain might be small and difficult to be revealed. Indeed, the interaction between confusion and time on test scores was not significant, indicating that confusion did not facilitate learning in this study. Therefore, even though MS might contribute to confusion resolution, their interaction effect on learning was marginal.

It is worth noting that in most prior studies that found the positive effect of confusion, confusion was induced intentionally with discrepant information (e.g., D'Mello et al., 2014; D'Mello & Graesser, 2014b; Lehman et al., 2012). In contrast, confusion appeared naturally in this study. Therefore, the discrepancy of new information encountered by students might not be as strong as in that earlier work. In addition, it is reasonable to expect that studies that intentionally spark confusion will have substantially more confusion than studies where confusion only occurs naturally. Most confused states might not reach the zone of optimal confusion, where the positive effect of confusion occurs (Arguel et al., 2019; D'Mello et al., 2014). Thus, in further research about confusion resolution, it may be more efficient to trigger confusion purposely rather than letting it occur naturally, to increase both its frequency and intensity. Nevertheless, if a finding can only be obtained through intentionally triggered confusion, it raises questions about the generality and scope of the phenomenon.

Another reason for our relatively weak results may be that the MS proficiency of sixth graders in this study was not strong enough for resolving confusion. Muis, Psaradellis et al. (2015) hypothesized that younger students might not posses the necessary strategies for confusion resolution. The MS of older elementary school children and even high school children are still developing (van der Stel & Veenman, 2014). For example, without intervention, the judgement of sixth and seventh graders about their comprehension of texts did not relate to testing performance, and they had difficulty in selecting the appropriate material for restudy (de Bruin, Thiede, Camp, & Redford, 2011).

5.4. Implications

The association between confusion and the change of MS behaviour implies that cognitive disequilibrium may not only lead to metacognitive experience (Efklides, 2011) and confusion (Mandler, 1990) but it may also indirectly induce the regulation of behaviour. This provided evidence for some assumptions of the MASRL model (Efklides, 2011). According to this model, cognitive interruption can trigger metacognitive experience and affective reaction, which in turn motivate bottom-up self-regulation of behaviour and emotions. Bottom-up self-regulation contrasts with top-down self-regulation, which is driven by the goal and plan that learners set based on personal characteristics such as competency, self-concept, and beliefs concerning learning. Given that confusion is measured in this study rather than cognitive disequilibrium, we cannot make any strong statement about the role of cognitive disequilibrium in bottom-up self-regulation. However, whether cognitive disequilibrium drives bottom-up self-regulation is an interesting question and worthy of further study.

Confusion occurred naturally in this study, and its occurrence was infrequent. The lack of frequent occurrence might limit the benefits of confusion resolution and may explain why there was no association between confusion and learning gain in our findings. Some of the previous findings discussed above suggest that it is useful to induce confusion in learners if they are experiencing an extended period without confusion. Timely and effective confusion inducement has three requirements. The first requirement is accurate detection of confusion (Calvo & D'Mello, 2010). An accurate affect detector makes it possible to determine when the desired condition is met — the student has not been confused for a considerable amount of time. In this case, our current automated detection of confusion in Betty's Brain remains imperfect (Jiang et al., 2018). It is possible that our current finding may lead to better affect detection in this environment — we have learned that students did more reading actions when confused, suggesting that information-searching behaviour may be useful in improving a confusion detector.

A second requirement is a sense of how much time without confusion is too much. There has been some research into the dynamics or half-life of affect (e.g., Botelho et al., 2018) and some work into the different learning outcomes associated with

different durations of confusion (Liu et al., 2013). Replicating this work in the context of Betty's Brain may make it possible to know when the time has arrived to induce confusion.

A third requirement is effective scaffolding for confusion resolution. Without such support, some students may not be able to resolve the triggered confusion, and a persistent confused state may evolve into frustration and boredom, which may harm rather than benefit learning (Botelho et al., 2018; D'Mello & Graesser, 2012). Researchers have suggested taking advantage of confusion in CBLEs by providing self-regulation strategy support (Arguel et al., 2017, 2019). This study indicates that MS alone may not be sufficient for confusion resolution. Cognitive and motivational factors should also be considered. For example, suppose students are confused about what to do next in Betty's Brain. In this case, the system delivers an MS prompt such as "think of what part of the map you are trying to build, perform a directed search of resource pages that talk about the relevant concepts, and read the text." If a student remains confused when they read resource pages, the system could deliver cognitive scaffolding like "recall the concepts whose relationships you want to know, pay attention to the verbs and modifiers connecting them, and take notes when necessary." If students are confused and become disengaged, using motivational support to re-engage them may be useful before delivering the MS and cognitive prompts. It is worth noting, however, that such a highly scaffolded learning design may not be beneficial for high-ability students because these students typically perform better in less-structured instructional environments (Cooper, 1993).

6. Conclusion

This study examined the relationship between confusion and metacognitive strategies (MS). The results showed that changes in MS behaviour co-occur with confusion, but that confusion resolution was not related to MS behaviour. Furthermore, MS did not moderate the effect of confusion on task performance and learning. These results demonstrate that MS may be a prerequisite but is not sufficient for confusion resolution.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from National Science Foundation under Grant #1561676 and the China Scholarship Council under Grant #201806040180.

References

Arguel, A., Lockyer, L., Kennedy, G., Lodge, J. M., & Pachman, M. (2019). Seeking optimal confusion: A review on epistemic emotion management in interactive digital learning environments. *Interactive Learning Environments, 27*(2), 200–210. https://doi.org/10.1080/10494820.2018.1457544

Arguel, A., Lockyer, L., Lipp, O. V., Lodge, J. M., & Kennedy, G. (2017). Inside out: Detecting learners' confusion to improve interactive digital learning environments. *Journal of Educational Computing Research, 55*(4), 526–551. https://doi.org/10.1177/0735633116674732

Azevedo, R., Behnagh, R., Duffy, M., Harley, J., & Trevors, G. (2012). Metacognition and self-regulated learning in student-centered leaning environments. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of student-centered learning environments* (2nd ed., pp. 171–197). New York: Routledge. https://doi.org/10.4324/9780203813799

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human–Computer Studies, 68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education, 26*(1), 350–364. https://doi.org/10.1007/s40593-015-0057-9

Bosch, N., D'Mello, S., & Mills, C. (2013). What emotions do novices experience during their first computer programming learning session? In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 11–20). Lecture Notes in Computer Science, vol. 7926. Springer. https://doi.org/10.1007/978-3-642-39112-5_2

Botelho, A. F., Baker, R. S., Ocumpaugh, J., & Heffernan, N. T. (2018). Studying affect dynamics and chronometry using sensor-free detectors. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th International Conference on Educational Data Mining* (EDM2018), 16–20 July 2018, Buffalo, NY, USA (pp. 157–166). International Educational Data Mining Society. https://eric.ed.gov/?id=ED593106

Brunner, E., & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers, 42*(1), 1–52. https://doi.org/10.1007/s003620000039

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., ... Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology, 100*(3), 525–534. https://doi.org/10.1037/0022-0663.100.3.525

Chevrier, M., Muis, K. R., Trevors, G. J., Pekrun, R., & Sinatra, G. M. (2019). Exploring the antecedents and consequences of epistemic emotions. *Learning and Instruction, 63*, 101209. https://doi.org/10.1016/j.learninstruc.2019.05.006

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., ... Lindstädt, S. (2018). Finding traces of self-regulated learning in activity streams. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 191–200). New York: ACM. https://doi.org/10.1145/3170358.3170381

Cooper, P. A. (1993). Paradigm shifts in designed instruction: From behaviorism to cognitivism to constructivism. *Educational Technology*, *33*(5), 12–19. https://www.jstor.org/stable/44428049

Cowen, A., Sauter, D., Tracy, J. L., & Keltner, D. (2019). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, *20*(1), 69–90. https://doi.org/10.1177%2F1529100619850176

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29*(3), 241–250. https://doi.org/10.1080/1358165042000283101

de Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

Di Leo, I., Muis, K. R., Singh, C. A., & Psaradellis, C. (2019). Curiosity… Confusion? Frustration! The role and sequencing of emotions during mathematics problem solving. *Contemporary Educational Psychology, 58*, 121–137. https://doi.org/10.1016/j.cedpsych.2019.03.001

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

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105*(4), 1082. https://doi.org/10.1037/a0032674

D'Mello, S., & Graesser, A. (2014a). Confusion. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 289–310). New York: Routledge. https://doi.org/10.4324/9780203148211.ch15

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

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

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review, 1*(1), 3–14. https://doi.org/https://doi.org/10.1016/j.edurev.2005.11.001

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. https://doi.org/10.1027/1016-9040.13.4.277

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906. https://doi.org/10.1037/0003-066X.34.10.906

Hargrove, R. A., & Nietfeld, J. L. (2015). The impact of metacognitive instruction on creative problem solving. *The Journal of Experimental Education, 83*(3), 291–318. https://doi.org/10.1080/00220973.2013.876604

Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., Moore, A. L., & Biswas, G. (2018). Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay. (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (AIED '18), 27–30 June 2018, London, UK (pp. 198–211). Lecture Notes in Computer Science, vol. 10947. Springer. https://doi.org/10.1007/978-3-319-93843-1_15

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2014). Analyzing the temporal evolution of students' behaviours in open-ended learning environments. *Metacognition and Learning, 9*(2), 187–215. https://doi.org/10.1007/s11409-014-9112-4

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2017). Integrating model-driven and data-driven techniques for analyzing learning behaviours in open-ended learning environments. *IEEE Transactions on Learning Technologies, 10*(2), 140–153. https://doi.org/10.1109/TLT.2015.2513387

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208. https://iaied.org/journal/997

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

Lehman, B., D'Mello, S., & Graesser, A. (2013). Who benefits from confusion induction during learning? An individual differences cluster analysis. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 51–60). Lecture Notes in Computer Science, vol. 7926. Springer. https://doi.org/10.1007/978-3-642-39112-5_6

Lehman, B., & Graesser, A. (2014). Impact of agent role on confusion induction and learning. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (ITS 2014), 5–9 June 2014, Honolulu, HI, USA (pp. 45–54). Lecture Notes in Computer Science, vol. 8474. New York: Springer. https://doi.org/10.1007/978-3-319-07221-0_6

Lehman, B., & Graesser, A. (2015). To resolve or not to resolve? That is the big question about confusion. In C. Conati, N. Heffernan, A. Mitrovic, & M. Verdejo (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (AIED '15), 22–26 June 2015, Madrid, Spain (pp. 216–225). Lecture Notes in Computer Science, vol. 9112. Springer. https://doi.org/10.1007/978-3-319-19773-9_22

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. S. (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* (EDM 2013), 6–9 July 2013, Memphis, TN, USA (pp. 114–120). International Educational Data Mining Society. http://educationaldatamining.org/EDM2013/proceedings/paper_34.pdf

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behaviour Research Methods, 52*, 464–488. https://doi.org/10.3758/s13428-019-01246-w

Mandler, G. (1990). Interruption (discrepancy) theory: Review and extensions. In S. Fisher & C. L. Cooper (Eds.), *On the move: The psychology of change and transition* (pp. 13–32). Chichester: Wiley.

Meany Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behaviour Analysis, 40*(3), 501–514. https://doi.org/10.1901/jaba.2007.40-501

Muis, K. R., Chevrier, M., & Singh, C. A. (2018). The role of epistemic emotions in personal epistemology and self-regulated learning. *Educational Psychologist, 53*(3), 165–184. https://doi.org/10.1080/00461520.2017.1421465

Muis, K. R., Pekrun, R., Sinatra, G. M., Azevedo, R., Trevors, G., Meier, E., ... Heddy, B. C. (2015). The curious case of climate change: Testing a theoretical model of epistemic beliefs, epistemic emotions, and complex learning. *Learning and Instruction, 39*, 168–183. https://doi.org/10.1016/j.learninstruc.2015.06.003

Muis, K. R., Psaradellis, C., Lajoie, S. P., Di Leo, I., & Chevrier, M. (2015). The role of epistemic emotions in mathematics problem solving. *Contemporary Educational Psychology, 42*, 172–185. https://doi.org/10.1016/j.cedpsych.2015.06.003

Munzar, B., Muis, K. R., Denton, C. A., & Losenno, K. (2020). Elementary students' cognitive and affective responses to impasses during mathematics problem solving. *Journal of Educational Psychology*, *113*(1), 104–124. https://doi.org/http://doi.org/10.1037/edu0000460

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

ISSN 1929-7750 (online).

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12). https://doi.org/10.18637/jss.v050.i12

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. B. (2015). *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual*. New York, NY/Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences. https://www.upenn.edu/learninganalytics/ryanbaker/BROMP.pdf

Ocumpaugh, J., Baker, R. S., Rodrigo, M. M., Salvi, A., Van Velsen, M., Aghababyan, A., ... Martin, T. (2015). HART: The human affect recording tool. *Proceedings of the 33rd Annual International Conference on the Design of Communication* (SIGDOC '15), 16–17 July 2015, Limerick, Ireland (Article 24). New York: ACM. https://doi.org/10.1145/2775441.2775480

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning, 13*(2), 179–212. https://doi.org/10.1007/s11409-018-9183-8

Pekrun, R., & Stephens, E. J. (2012). Academic emotions. In K. Harris, S. Graham, T. Urdan, S. Graham, & J. Royer (Eds.), *APA educational psychology handbook: Vol. 2. Individual differences and cultural and contextual factors* (pp. 3–31). Washington, DC: American Psychological Association. https://doi.org/10.1037/13274-000

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Planalp, S., DeFrancisco, V. L., & Rutherford, D. (1996). Varieties of cues to emotion in naturally occurring settings. *Cognition and Emotion, 10*(2), 137–153. https://doi.org/10.1080/026999396380303

Rodrigo, M. M. T., Baker, R. S. J. D., D'Mello, S., Gonzalez, M. C. T., Lagud, M. C. V., Lim, S. A. L.,... Viehland, N. J. B. (2008). Comparing learners' affect while using an Intelligent Tutoring System and a simulation problem solving game. In B. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (ITS 2008), 23–27 June 2008, Montreal, QC, Canada (pp. 40–49). Lecture Notes in Computer Science, vol. 5091. Springer. https://doi.org/10.1007/978-3-540-69132-7_9

Segedy, J. R., Biswas, G., & Sulcer, B. (2014). A model-based behaviour analysis approach for open-ended environments. *Journal of Educational Technology & Society, 17*(1). https://doi.org/https://www.jstor.org/stable/26896707

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2015). Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics, 2*(1), 13–48. https://doi.org/10.18608/jla.2015.21.3

Silvia, P. J. (2010). Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts, 4*(2), 75–80. https://doi.org/10.1037/a0017081

van der Stel, M., & Veenman, M. V. (2014). Metacognitive skills and intellectual ability of young adolescents: A longitudinal study from a developmental perspective. *European Journal of Psychology of Education, 29*(1), 117–137. https://doi.org/10.1007/s10212-013-0190-5

Vogl, E., Pekrun, R., Murayama, K., Loderer, K., & Schubert, S. (2019). Surprise, curiosity, and confusion promote knowledge exploration: Evidence for robust effects of epistemic emotions. *Frontiers in Psychology, 10*, Article 2474. https://doi.org/10.3389/fpsyg.2019.02474

Vogl, E., Pekrun, R., Murayama, K., & Loderer, K. (2020). Surprised–curious–confused: Epistemic emotions and knowledge exploration. *Emotion, 20*(4), 625–641. https://doi.org/10.1037/emo0000578

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *The Journal of Educational Research, 84*(1), 30–43. https://www.jstor.org/stable/40539680

Wilcox, R. R. (2011). Introduction to robust estimation and hypothesis testing (3rd ed.). Oxford, UK: Elsevier.

Wilcox, R. R., & Tian, T. S. (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics, 38*(7), 1359–1368. https://doi.org/10.1080/02664763.2010.498507

Yang, D., Kraut, R., & Rose, C. P. (2016). Exploring the effect of student confusion in massive open online courses. *Journal of Educational Data Mining, 8*(1), 52–83. https://doi.org/10.5281/zenodo.3554605

Appendix

An example of a multiple-choice question:
Question: Light from the sun comes to the earth and its energy is absorbed by the atmosphere. What is the relation between this absorbed light energy and heat energy absorbed by the Earth?
a. Absorbed light energy increases the amount of absorbed heat energy.
b. Absorbed light energy decreases the amount of absorbed heat energy.
c. Absorbed light energy does not change the amount of absorbed heat energy.
d. Absorbed light energy is not related to absorbed heat energy.
(Correct answer is a.)

An example of a short answer question:
Question: We know that deforestation (cutting down a large number of trees) increases the Earth's absorbed heat energy. Explain, step-by-step, how deforestation increases the Earth's absorbed heat energy.
Step 1: Deforestation reduces the number of trees on the Earth, so more deforestation would decrease vegetation.
Step 2: When vegetation decreases, _____
Step 3: _____