Students' Verbalized Metacognition during Computerized Learning

NIGEL BOSCH

University of Illinois Urbana–Champaign, USA, pnb@illinois.edu

YINGBIN ZHANG

University of Illinois Urbana–Champaign, USA, yingbin2@illinois.edu

LUC PAQUETTE

University of Illinois Urbana–Champaign, USA, lpaq@illinois.edu

RYAN S. BAKER

University of Pennsylvania, USA, rybaker@upenn.edu

JACLYN OCUMPAUGH

University of Pennsylvania, USA, ojaclyn@upenn.edu

GAUTAM BISWAS

Vanderbilt University, USA, gautam.biswas@vanderbilt.edu

Students in computerized learning environments often direct their own learning processes, which requires metacognitive awareness of what should be learned next. We investigated a novel method of measuring verbalized metacognition by applying natural language processing (NLP) to transcripts of interviews conducted in a classroom with 99 middle school students who were using a computerized learning environment. We iteratively adapted the NLP method for the linguistic characteristics of these interviews, then applied it to study three research questions regarding the relationships between verbalized metacognition and measures of 1) learning, 2) confusion, and 3) metacognitive problem-solving strategies. Verbalized metacognition was not directly related to learning, but was related to confusion and metacognitive problem-solving strategies. Results also suggested that interviews themselves may improve learning by encouraging metacognition. We discuss implications for designing computerized environments that support self-regulated learning through metacognition.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI • Applied computing~Education~Interactive learning environments • Computing methodologies~Artificial intelligence~Natural language processing

Additional Keywords and Phrases: Metacognition, Self-regulated learning, Confusion, Affect

ACM Reference Format:

Nigel Bosch, Yingbin Zhang, Luc Paquette, Ryan S. Baker, Jaclyn Ocumpaugh, and Gautam Biswas. 2021. Students' Verbalized Metacognition during Computerized Learning. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, XX pages.

1 INTRODUCTION

Metacognition, or *thinking about thinking*, is an activity that is crucial for learning [20,25,26,40,52,61]. Metacognition enables students to identify gaps in their own knowledge and devise strategies to fill those gaps [22]. In computerized learning environments, metacognition is especially important, because students are often expected to self-regulate their learning activities; that is, to decide what activities or topics to focus on at any given moment and what strategies to use for acquiring or applying knowledge. Self-regulated learning is possible when students have awareness of what they know and do not know (i.e., metacognitive knowledge) [43,53]. However, metacognition is difficult to study via typical methods like self-report questionnaires [29,58], which has led to additional measurement methods like analysis of prose [31] and sequences of activities [10]. In human–computer interaction (HCI) tasks in particular, metacognition has been assessed in terms of software interaction behaviors that are consistent with (but do not directly measure) metacognition [44,54,64], or by developing interfaces that allow users to explicitly note the metacognitive and self-regulated learning strategies they are currently using [5]. In this paper, we approach the problem of measuring metacognition during software interactions by applying natural language processing (NLP) methods to transcripts of interviews conducted with students as they used a computerized learning environment. In particular, we analyzed transcripts to count cases where students spoke about their own cognitive states or processes (i.e., *verbalized metacognition*). We explored how verbalized metacognition relates to learning, emotion, and software interaction behaviors.

Computerized learning environments offer varying degrees of flexibility in terms of possible interaction behaviors and how interactions are guided. For example, some environments implement a *mastery learning* approach [14,30,48]. In mastery learning, the learning environment automatically selects learning content for the student to complete, based on measures of what the student already knows or does not know, until the student has mastered a particular topic [9]. In contrast, many other learning environments are open-ended, offering students a great deal of flexibility in deciding what content to consume next or how to approach solving a given problem [7,35,55,62]. In open-ended environments, metacognition allows students to choose learning strategies that may help them learn more efficiently [34], since students have many choices about what actions to take, including some that are better suited to their particular needs than others [57].

Metacognitive awareness of where knowledge gaps are is especially critical for students to be able to identify actions (e.g., read about a particular topic) that will address those gaps. Consequently, metacognitive monitoring skills are also needed to recognize the existence of knowledge gaps [51]. However, simply being aware of knowledge gaps is not always enough; students must also have strategies to close those gaps and take action to do so [28]. For example, students may become confused if they encounter a mismatch between content regarding a particular concept and their own understanding of that concept. If the source of that confusion is resolved, it may be beneficial to learning [19], whereas unresolved confusion may lead to frustration and eventual disengagement from learning [17]. Resolving confusion in an open-ended self-regulated learning environment requires four metacognitive elements: recognition of feelings of confusion (i.e., *metacognitive experiences*) to identify the existence of a knowledge gap (i.e., *metacognitive knowledge*), approaches to resolve those gaps (i.e., *metacognitive strategies*), and an intention to act on those strategies [23,24,28]. Treating confusion as a metacognitive experience (as opposed to focusing on the affective experience of confusion itself) broadens the focus to the source of confusion, and metacognitive strategies needed to resolve it.

Metacognitive strategies are often observable in HCI contexts via interaction logs, the records of sequential actions users have taken in an interface. In the case of computerized learning environments, for example, we can observe whether students who struggle with a particular concept subsequently take action by seeking out more information about that concept in the learning environment [64]. However, other aspects of metacognition – students' metacognitive

experiences and metacognitive knowledge – are often less salient. The approach to measuring metacognition that we explore in this paper (i.e., automated analysis of interview transcripts) enables assessment of metacognition across three elements of metacognition (experiences, knowledge, and strategies). We combined verbalized metacognition measurement with existing measures of confusion and metacognitive strategies, which are derived from students' interaction logs in the learning environment. The interaction-based measures of metacognitive strategies also address the fourth aspect of metacognition: taking action. With these combined data, we investigated the following three research questions (RQs), which contribute to understanding of users (in this case, students) as they use an open-ended computerized learning environment.

RQ1: Does verbalized metacognition relate to students' learning?

Hypothesis: We expect that students who express more metacognitive thoughts during interviews will learn more, because they are more aware of what they need to learn and how to learn it.

Relevance to HCI: This research question focuses on the importance of metacognition, broadly conceptualized, in computerized learning environments, and thus how important it is to design such environments to support metacognition.

RQ2: Does verbalized metacognition relate to confusion, as measured via interaction logs of student behaviors?

Hypothesis: We expect that interviews coinciding with periods of student confusion will contain more verbalized metacognition from students, as they express their confusion and try to resolve it through think-alouds and conversation.

Relevance to HCI: If confused students express more metacognition during interviews, it suggests that targeted interviews (or discussion, more generally speaking) may be an effective way to resolve confusion in learning software.

RQ3: Does verbalized metacognition relate to metacognitive strategies students use in the software? And if so, which strategies?

Hypothesis: We expect that students who exhibit more verbalized metacognition will also employ metacognitive strategies more frequently, since they are more likely to be aware of their knowledge gaps and have strategies to address those gaps.

Relevance to HCI: The metacognitive strategies we extracted from interaction logs are hypothesized to measure metacognition based on *coherence analysis* [54], as described below. Aligning these metacognitive strategies with students' verbalized metacognition will provide validation that these strategies are indeed indicative of metacognition and can thus be applied in contexts where more involved measures (like interviews or surveys) are difficult to use.

Taken together, these three research questions expand our understanding of the relationships between verbalized metacognition and measures that are more common in computerized learning environments. Specifically, these questions explore the importance of students' verbalized metacognition in learning (RQ1), how verbalized metacognition relates to confusion (RQ2) and how verbalized metacognition relates to enacted metacognitive strategies (RQ3). The contributions of this paper consist of the results of a real-world classroom study that investigated these research questions, as well as the details of our novel verbalized metacognition measurement process. We provide procedures and code needed to apply our method to other HCI contexts so that verbalized metacognition can be studied more generally in think-aloud studies, voice user interfaces, or other applications. The contributions in this paper are founded on previous research on the intersection of metacognition, learning, and HCI, which we discuss next.

2 RELATED WORK

Research on metacognition began in earnest in the 1970s [25,26], resulting in a large body of research exploring it across various domains. Here, we focus specifically on research that has quantified the role of metacognition in HCI, then how it relates to confusion and software usage behaviors.

2.1 Metacognition in HCI

Metacognition is helpful, or even essential, in a wide variety of HCI tasks, such as using computer programming tools [8,39,46], writing argumentative text on the web [60], creating and assessing visualizations [33], and learning via technology [4]. Metacognition is also important in activities that commonly take place during HCI research, like participatory design, where participants need metacognitive knowledge to be able to identify what knowledge to contribute to a design [16].

In one study, Vu et al. [59] asked users to self-assess their expertise (i.e., make a metacognitive judgment) before describing how to complete tasks in word processing software. Self-assessed expert users were more accurate in their descriptions of how to complete the tasks, and they used more complex strategies when needed for difficult tasks. These findings indicate that users' metacognitive evaluation of knowledge correlates with application of that knowledge for some HCI tasks.

However, Ackerman and Goldsmith [1] found some limitations in users' metacognitive knowledge. They compared how well users read text on a computer screen versus on paper, in terms of a reading comprehension post-test. Users were also required to estimate how well they would perform on the post-test before taking it, which served as a measure of metacognitive knowledge (i.e., whether their estimates aligned with how well they actually performed). While both computer- and paper-reading users overestimated their actual post-test scores, computer users did so significantly more. Moreover, when given a fixed amount of time to complete the reading, post-test scores were similar across computer and paper conditions, but when users were allowed to read for as long as they wanted, they achieved significantly better post-test scores reading from paper. These findings suggest that metacognitive knowledge is comparatively lacking in computerized environments, and that self-regulation (in this case, deciding how long to read) may also be more difficult to calibrate for computer users, perhaps because diminished metacognitive knowledge prevents users from accurately deciding how long to spend on the task.

Users' metacognition has also been measured by means other than questionnaires in some tasks. For example, Litman and Forbes-Riley [38] measured metacognition during interactions with a conversational computerized learning environment in which students learn by verbally discussing introductory physics topics with the computer. A researcher transcribed students' speech in real-time to provide text to the learning software, and annotated the text for statements of certainty (i.e., verbalized metacognitive knowledge). In a subsequent version, automatic speech recognition provided transcriptions while a machine learning model trained on speech features (e.g., pitch, energy) classified certainty. In both versions, verbalized metacognition was positively related to how much students learned.

These previous studies have shown that metacognition is important for understanding users' interactions with computers, including during learning experiences. Our work extends this area of research by considering verbalized metacognition expressed during interviews, as opposed to questionnaires or other methods, and by exploring relationships between verbalized metacognition and users' software interactions (including those that indicate confusion and metacognitive strategies, as discussed below).

2.2 Metacognition and Confusion

Expressing confusion requires awareness of the existence of confusion (a metacognitive experience) and may incorporate the cause of that confusion (metacognitive knowledge) [21]. When students encounter a knowledge gap or conflict between their existing mental model and newly encountered information, confusion may result [18]. When students are aware of the cause of that confusion, they have an opportunity to address the cause and thus learn something new [15,17,19,37], provided that they intend to act on that opportunity [28]. Hence, promoting awareness of confusion is a common tool for improving metacognition and learning [56].

Several studies have measured confusion, or even manipulated it [37], to determine its relationships to metacognition and its role in learning. Campbell et al. [11] annotated students' computer-mediated messages to a human tutor in an computerized environment for learning about electronics and electricity. Specifically, researchers marked statements made by students about comprehension (e.g., "I understand") and confusion (e.g., "I don't understand") as measures of metacognitive experiences. Both types of metacognitive statements were negatively related to post-test scores. However, in contrast to research by Litman and Forbes-Riley [38] and Dowd et al. [21] (discussed below), Campbell et al. did not measure learning (i.e., improvement between pre- and post-tests); thus, their results are consistent with the possibility that students with low prior knowledge (and thus low pre- and post-test scores) may have engaged in more metacognitive activity by necessity.

Conversely, Dowd et al. [21] controlled for pre-test score in an analysis of students' self-reported confusion, and found that confusion was positively related to learning introductory physics. They also found that confusion was negatively related to pre-test score, which suggests that more confused students did have lower prior knowledge, but were able to learn from their confusion. Notably, students self-reported confusion, thus only capturing metacognitive experiences of confusion, where they were paying attention to their confusion and thus well-positioned to address it.

Rather than relying on self-reports, Zhang et al. [64] employed trained observers to record instances of student confusion in real-time as students interacted with a computerized learning environment for natural science topics. They found no overall relationship between confusion and learning, but did find that students used more metacognitive strategies (as measured via coherence analysis [54]) during periods of confusion, suggesting that students were aware of their confusion and knew how to take steps to address it.

In this paper, we measure learning (as opposed to only post-test scores), and incorporate assessment of verbalized metacognition in response to researcher prompts (rather than self-reports) with measures of metacognitive strategies based on coherence analysis. We thus extend previous research by determining whether students are aware of and addressing their confusion through interviews.

2.3 Metacognition Strategies in Software Usage

Coherence analysis, a measure of metacognitive strategy usage, refers to whether actions done in sequence in an interface are informed by one another [54]. For example, if a user performs some action *X* (e.g., reading a message) that provides them with knowledge they need to perform action *Y*, then the *XY* sequence is said to be coherent (as opposed to doing *Y* before *X*, for example). Coherent action sequences in computerized learning environments are evidence of students employing metacognitive strategies, a type of self-regulated learning behavior, to address issues identified via their metacognitive experiences (e.g., confusion) or metacognitive knowledge. Given the potential benefits of metacognition, several studies have measured or encouraged coherent actions and related metacognitive strategies to understand or improve computerized learning.

In two closely-related studies, Wang et al. examined the effects of software features intended to support metacognition in an environment called MindDot [62,63]. MindDot is a computerized learning environment for concept mapping, where students learn to graphically link together ideas via relationships [42]; for example, *species* [concept] → *provide* [link] → *ecological services* [concept]. In one study, researchers added a tool intended to facilitate coherent actions in MindDot by making related information more readily accessible [63], while in another study they added a feature designed to incorporate expert knowledge via a template used as a starting point for their concept map [62]. These features were designed to support metacognitive strategies for comparison by making information and links between concepts more salient. They found that use of metacognitive strategies was more predictive of learning than the quality of students' concept maps themselves.

Metacognitive strategies are also valuable for learning computer programming [8,45,46]. Prather et al. [45] added a scaffolding phase to an interface for automatic program evaluation, wherein students were required to trace through an example case for a problem before writing C++ code to solve the problem. The scaffolding phase was intended to improve students' metacognitive awareness of potential difficulties they might encounter during the problem. In a controlled experiment, students using scaffolding outperformed students in the control condition on the subsequent programming task. Furthermore, researchers qualitatively coded students' speech during the experiment (via a think-aloud protocol) and found that students in the scaffolding expressed more verbalized metacognition, including metacognitive strategies, experiences, and knowledge.

In sum, related work has shown that metacognition is a key consideration for modeling users' experiences in HCI tasks, especially those involving learning. Moreover, there are many methods for measuring metacognition, including those that focus on only one facet (e.g., metacognitive experiences of confusion) and those intended to capture more aspects of metacognition (via students' speech). These findings motivated the study in this paper, which explored an NLP method for measurement of verbalized metacognition and triangulated verbalized metacognition with learning, confusion, and software interactions.

3 DATA COLLECTION

We conducted a study with students in a U.S. middle school classroom context using a computerized learning environment called Betty's Brain [7].

3.1 Betty's Brain

Betty's Brain is a "learning by teaching" [6] concept mapping environment in which a student's goal is to construct a concept map that functions as the "brain" for a simulated student named Betty (Figure 1). Specifically, students create a causal map, where the links in the map denote cause-and-effect relationships. Students can then administer quizzes to Betty, who will answer these questions according to the concept map the student has constructed (Figure 2). Students also have access to a pedagogical agent named Mr. Davis, who can provide hints about what to do next (referred to as *feedback* below). In addition to constructing concept maps, students can read about the topic for which they are building a concept map via the "Resources" tab, take notes, and see the results of quizzes that Betty has taken. Students choose what action to perform next throughout the learning process; they thus engage in self-regulated learning, where metacognition is expected to be especially valuable [24,52].

Students' interactions with the Betty's Brain software are logged in timestamped files, which enables analysis of their interactions to infer certain aspects of their learning experience. In this study, we used a version of Betty's Brain that provides events triggered automatically by students' affective and behavioral states. Affective states included emotions

like confusion and delight, and behaviors like reading for a long period of time [32,41]. Specifically, Betty's Brain detects affective states from students' interaction log files, and triggers events based on sequences of affective states that are hypothesized to be especially relevant for learning [3,17].

In our study, we used these affect and behavior events to signal interviewers to begin an interview with a student. Interviewers were signaled via a publicly-available smartphone application called Quick Red Fox[1], which integrates with Betty's Brain events and allows users to record metadata related to each event (in this case, timestamps and which student was being interviewed). For our second research question, we examined cases where interviews were initiated by affect sequences involving confusion, where the confusion experience was either resolved (recently or not recently), ongoing, or unresolved (Table 1). Betty's Brain detects these events and stores them to a server. Quick Red Fox then reads from the same server in real-time to trigger interviews by signaling to the interviewer.

Table 1: Confusion-related affect sequences that were used to initiate some of the interviews in our study. Other interviews were initiated by affect sequences not involving confusion, behavior events, or explicit requests from students.

Affect sequence	**Expected interpretation**
Engagement → confusion → delight → engagement	The student has fully resolved the source of confusion and returned to working on the task
Confusion → delight	The student recently resolved the source of confusion, but has not yet returned to engagement with the task
Confusion → frustration	The student became frustrated after not resolving the source of confusion, but has not disengaged
Engagement → confusion → frustration → boredom	The student did not resolve the source of confusion, and disengaged instead

[1] http://www.knossys.com/deploy/QRF/

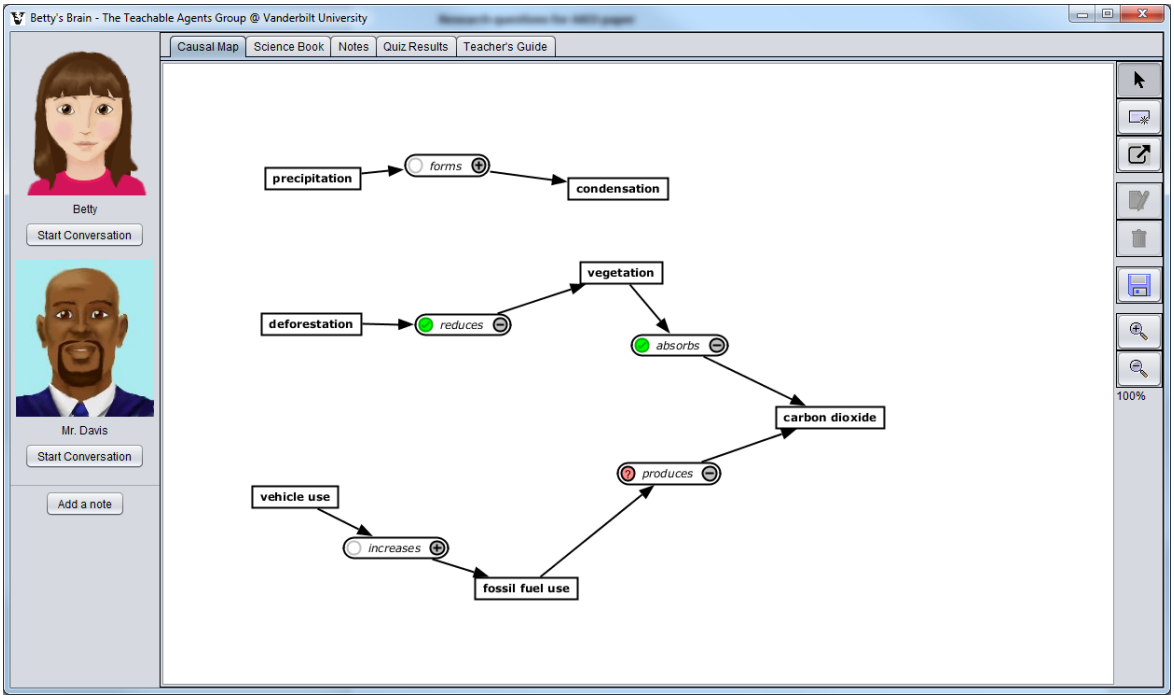

Figure 1: Screenshot of Betty's Brain showing a concept map under construction. On the left and bottom of the interface are options for students to interact with virtual agents (Betty and Mr. Davis) to either test their concept map (by administering a quiz to Betty) or to receive feedback about what to do next (from Mr. Davis).

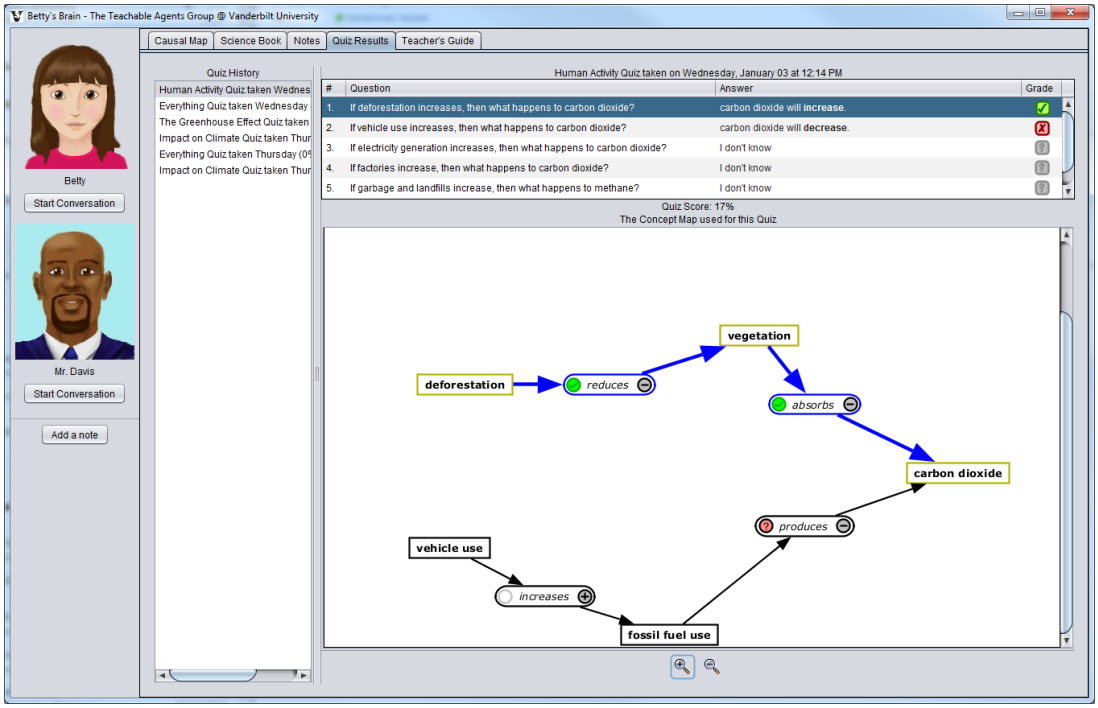

Figure 2: Screenshot of Betty's Brain showing the quiz results interface, where students can see each question in the quiz and a history of how the Betty agent responded to each question. Students can also see *why* Betty responded in a particular way for each question, according to sections of the concept map that are highlighted.

3.2 Study Participants and Procedure

We conducted our study during the 2018–2019 school year with 99 middle school students in the U.S., who participated as part of their regular classroom learning. The classroom teacher was not involved in the study apart from typical daily announcements and one classroom-wide intervention, which was needed after several students were disappointed with Betty's quiz scores. All study procedures were approved by an institutional review board before data collection began. Students participated during two data collection periods, and received training on how to use Betty's Brain before data collection began. The first data collection period lasted four days, during which students spent approximately 45–50 minutes each day using Betty's Brain to learn about climate change (visible in Figure 2). The second data collection period lasted three days, and was conducted two months after the first. The second period was similar to the first, except that students learned about thermoregulation. Students completed a pre-test measuring their prior knowledge before beginning each of these topics, and an identical post-test after completing each topic. No students achieved perfect scores on either test, and there were no apparent floor effects. We calculated the difference between pre- and post-test scores (averaged across topics) as a measure of how much students learned. Most students improved; only 7.1% experienced negative learning gains. We also calculated a normalized measure of learning, where we divided the difference between post- and pre-test scores by the amount of improvement possible (i.e., 1 – pre-test score). The normalized measure correlated almost perfectly with the difference alone ($r = .995$), however, and subsequent results were virtually identical. Hence, we proceeded with the simpler difference measure.

Two interviewers working separately conducted verbal in-person interviews with individual students throughout both periods of data collection and recorded the interviews with microphones carried by the interviewers. Interviewers primarily selected students to talk to based on the affective and behavioral events presented to them via the Quick Red Fox smartphone application described above, but also employed other criteria, including explicit requests from students who wanted to talk about their current experiences. Unstructured interview strategies were employed, and interviews lasted 36 seconds on average. Interviewers often encouraged students to talk about their specific strategies or to provide feedback about their experiences with the software, but conversational strategies to interrogate student motivations, including intrinsic interest were also used. These included questions about students' interest in science, their favorite school subjects, and their preferred reading materials. Thus, these conversations were not explicitly designed to promote students' metacognition, but metacognitive topics were prevalent in students' speech nonetheless (as shown below).

Interviews were transcribed by two researchers to enable NLP analysis. Interviews with multiple students or multiple interviewers were excluded, since the Quick Red Fox application did not support recording metadata for such interviews. In total, 493 interviews of 99 students remained after these exclusions. Each student was interviewed at least once, and 4.98 times on average (SD = 2.68), allowing us to measure verbalized metacognition for each student.

4 MEASURING METACOGNITION

We measured verbalized metacognition from interview transcripts, which was necessary for all three research questions, and measured metacognitive strategies via coherence analysis as part of studying RQ3.

4.1 NLP Analysis of Interviews

We adapted an existing open-source NLP tool for measuring metacognition from text [31]. This tool counts phrases beginning with a first-person pronoun and ending with a metacognitive indicator word, such as "considered" or "expected". The tool is intended to capture metacognitive statements across three types of metacognition (metacognitive knowledge, experiences, or strategies); as observed in previous research, verbalized metacognition appears to span all three categories [45]. The tool also subdivides metacognitive phrases into positive (confident) and negative (unconfident) categories, similar to related work that examined confidence and confusion separately [11]. However, we combined these two subcategories into an overall metacognition category, given that negative metacognitive statements were relatively uncommon and previous research shows that results are similar for both subcategories [11,31].

The NLP tool was originally designed to extract metacognitive statements from text written in online discussion forums by college students, unlike the context of our study. The vocabulary and grammar of middle school student interview transcripts differs somewhat from the expectations of the tool. For example, one student in our study said "So, actually, my last quiz she got it, he just, she just didn't know, and I've just been using questions like why, to figure out what kind of material to input." In this example, the student refers to a problem-solving strategy intended to address metacognitive knowledge gaps; however, the sentence structure is complicated. In another case, a student said "Um, like, when I play, I use the strategy of my dad..." Here the student indicates metacognitive awareness of a strategy they are using, but expressed it in a way that was not captured by the NLP method. We addressed these and other commonly-observed errors made by the tool by adjusting the dictionaries of metacognitive indicator words to suit the vocabulary of students in our study and adding regular expressions to capture common metacognitive expressions that did not conform to the expected phrasing structure. Our modified version of the NLP tool is publicly available[2].

[2] https://github.com/pnb/metacognitive_phrase_detector

We adapted the NLP tool via an iterative process consisting of five rounds of human annotation (with the same annotator for each round) and comparison to NLP predictions, all of which were performed on transcripts from the first period (four days) of data collection only (we expected that the tool would generalize well to the second period of data collection with the same students and context). We developed a written guide for annotation in which we defined metacognition and instructed the annotator to count the number of first-person metacognitive phrases observed, including contextual information (e.g., transcriptions of the interviewer's speech) to make decisions. Annotation began with exploratory rounds to determine major issues with applying the NLP tool in this context, then continued to larger rounds informed by power analysis.

Round 1: We annotated 15 interviews, which included 85 conversational turns from students. We measured agreement between the annotator and the NLP tool via linear-weighted Cohen's kappa [13], achieving kappa = .655 (where 0 indicates random guessing and 1 indicates perfect agreement). This kappa indicates "substantial" agreement [36], but we nevertheless addressed one NLP shortcoming (a regular expression needed to capture a common phrase) before the second round.

Round 2: We annotated an additional 15 interviews (63 student conversational turns), yielding kappa = .849, or "almost perfect" agreement [36]. We observed no consistent NLP errors in this round, and thus proceeded to round 3 with a larger data sample.

Round 3: Given round 2 agreement, we conducted an 80% power analysis assuming a null kappa of .7 (midpoint of the "substantial" range [36]) using the *irr* package [27] in R [47], based on which we annotated 593 additional student turns. Kappa was .530, indicating "moderate" agreement. We observed several consistent errors in this round, and added 15 additional metacognitive indicator words (e.g., "strategy", "progress"), 4 regular expression patterns, and 1 first-person pronoun ("me") that had not been included previously. Kappa on this round was .667 after these changes.

Round 4: We conducted another power analysis assuming a null kappa of .5 (midpoint of the "moderate" range), given the lower round 3 results (kappa = .667 post-adjustments), and thus annotated 285 additional student turns. Kappa was .656, indicating "substantial" agreement between the annotator and the updated NLP tool. We proceeded to round 5 without additional updates, given that agreement was substantial and the tool appeared to generalize well from round 3.

Round 5: Finally, we annotated the remainder of the first data collection period's transcripts (1,121 additional student conversational turns). Kappa on these data was .688, indicating that the changes made through the first 3 rounds generalized well to the remaining data.

After applying the finalized NLP tool to all transcripts for both data collection periods, kappa for annotated transcription turns (i.e., data from the first period of data collection) was .684. We then proceeded to measure metacognition from interaction log files.

4.2 Coherence Analysis of Interaction Logs

Betty's Brain includes metadata linking concepts to learning materials and quizzes, which enables features like explanations of Betty's quiz answers (Figure 2) and, in our case, identification of coherent actions. Coherence analysis relies on sequences of events (e.g., quiz → reading relevant material), but there may be intermediate events (e.g., quiz → concept map editing → reading relevant material). Thus, we counted a pair of actions as coherent if the second action was informed by the first **and** the second action occurred within five minutes of the first, with any number of other actions separating the two. This cutoff was in line with prior work applying coherence analysis in Betty's Brain [54,64].

We extracted the first four metacognitive strategies based on previous coherence analysis in Betty's Brain [64], and one new strategy (coherent feedback).

Coherent quiz view: Students view the results of a quiz, then perform any action to resolve an issue identified in the quiz (typically, like editing the concept map to add a missing link or reading about a concept related to an incorrect link in the map).

Coherent mark: Students mark a link in their concept map (i.e., annotate it), based on either quiz results or prompts from Mr. Davis, the virtual tutor agent. Possible marking actions include marking the link as correct or potentially incorrect, or deleting the annotation.

Coherent edit: Students edit links in the concept map based on information either gathered from a quiz or from reading about a specific concept (e.g., reading about deforestation, then adding a "reduces" link between "deforestation" and "vegetation" in the concept map).

Coherent read: Students read about a topic based on a quiz result or feedback from Mr. Davis. Such feedback is generated, for example, when students adds several incorrect concept map links in a row or when they explicitly ask Mr. Davis for help.

Coherent feedback: Students followed the feedback automatically generated by Mr. Davis who suggested that they read about a topic. This is a more specific subset of coherent read to distinguish between potential impetuses for reading.

We normalized the counts of these metacognitive strategies by dividing by the amount of time each student spent using Betty's Brain because some students spent slightly more or less time with the software than others. These coherent action frequencies (count/minute) then served as the measures of metacognitive strategies in software use for RQ3.

5 RESULTS

We report results organized by each research question. Throughout results, we rely on Spearman's *rho* correlations, given that verbalized metacognition counts were ordinal but not normally distributed.

5.1 RQ1: Verbalized Metacognition and Learning

Research question 1 asks *does verbalized metacognition relate to students' learning?* To investigate this question, we measured the correlation between verbalized metacognition (normalized by the number of interviews) and learning (student-level post-test – pre-test scores, averaged across data collection periods). We normalized verbalized metacognition by dividing the student-level metacognitive phrase counts by the number of interviews that student participated in to get the per-interview average, for this and the other research questions.

Verbalized metacognition correlated $rho = .051$ ($p = .627$) with learning, indicating no significant relationship. Our hypothesis for RQ1 was therefore not supported. We examined the normalization strategy to determine whether this might have had some effect on the results, and found that interview count itself positively correlated with learning ($rho = .254$, $p = .014$). We further explored the effect of interviews for students who expressed high (above median) versus low metacognition per interview. Interview count was positively correlated with learning for high-metacognition students ($rho = .327$, $p = .027$), but was not significantly related for low-metacognition students ($rho = .228$, $p = .119$). Thus, RQ1 findings suggest that, although verbalized metacognition per interview was not related to learning, interviews themselves may have benefitted learning, especially for students who were likely to engage in verbalized metacognition. However, it also possible that interviewers were more likely to interview successful students, though this was not their goal.

Table 2 provides illustrative examples of two interview transcripts where differences in verbalized metacognition are apparent. In one interview, a student discussed their metacognitive experiences (e.g., "it was a little bit like a dawning realization") and demonstrates metacognitive knowledge (e.g., "I already knew it was bad"), though does not describe strategies for addressing problems with the concept map. In the other, a student describes encountering a problem ("this idiot [Betty] kept getting everything wrong"), but does not appear to be aware of knowledge gaps that might be the source of this problem, and describes a strategy ("Take away everything and start from scratch") that is not informed by strong metacognitive knowledge of exactly what the specific flaws in the concept map are or strong metacognitive strategies to address those specific flaws.

Table 2: Example interview excerpts from a student exhibiting a high amount of verbalized metacognition versus a low amount.

High metacognition interview		**Low metacognition interview**	
Interviewer	You're not taking the test? Yeah OK. Has anything in the content really surprised you?	*Interviewer*	How's it going here?
Student	I was aware of it but I wasn't, like, so informed about all this.	*Student*	I've had to reset everything.
Interviewer	OK.	*Interviewer*	Why? Is the system broken, or…?
Student	So, like, all the details of how exactly it works, and…	*Student*	No, no, it's not a system problem, it's just that this idiot [Betty] kept getting everything wrong.
Interviewer	OK.	*Interviewer*	Uh oh.
Student	I wouldn't say it surprised me. I already knew it was bad, but, like, it was a little bit like a dawning realization of how, uh, bad it is.	*Student*	So, I just had to get it out. Take away everything and start from scratch.
Interviewer	Worse than you thought? So— So what you're describing is more just a surprise, and not like confusion there, nothing not making sense?		
Student	Yeah. Yes. Making it makes sense when I understood how well— reasonably, I understand better now.		
Interviewer	Yeah.		
Student	I'm a little more aware now		

5.2 RQ2: Verbalized Metacognition and Confusion

Research question 2 asks *does verbalized metacognition relate to confusion, as measured via interaction logs of student behaviors?* We investigated this research question by measuring the student-level correlations between verbalized metacognition and the proportion of interview conversational turns that were from interviews initiated for confusion-related reasons. As noted above, interviews could be initiated for affective, behavioral, and other reasons, including sequences of multiple affective states. We considered an interview confusion-related if it was initiated from any sequence of affective states involving confusion. Note that the proportion of interviews initiated by each sequence varied

substantially because we prioritized certain sequences over others, which was necessary given that some sequences were subsequences of others.

Results partially supported the hypothesis that students would verbalize more metacognition during periods of ongoing and recent confusion, versus less recent confusion experiences and unresolved confusion. In particular, the affect sequence where students had already returned to engaging with the task (*engagement* → *confusion* → *delight* → *engagement*) showed no relationship with metacognition (*rho* = .011, p = .937, 69.7% of interviews). When confusion had been resolved recently (*confusion* → *delight*) and when it had not yet been resolved (*confusion* → *frustration*), verbalized metacognition trended positive (but was not significant; *rho* = .127, p = .371, 8.9% of interviews, and *rho* = .185, p = .189, 2.0% of interviews, respectively). Conversely, when confusion went unresolved (*engagement* → *confusion* → *frustration* → *boredom*), verbalized metacognition was significantly lower (*rho* = -.282, p = .043, 8.1% of interviews). Affect sequences of interest triggered 88.7% of interviews combined. Other triggers contributed little to the number of interviews: less than 0.1% of interviews were triggered by affect sequences not considered in this paper (*frustration* → *engagement* and *sustained delight*), approximately 1% were triggered by student requests for interviews, and the rest were triggered based on one of 12 different behavior-based triggers (e.g., reading for a long period of time).

5.3 RQ3: Verbalized Metacognition and Metacognitive Strategies

Research question 3 asked *does verbalized metacognition relate to metacognitive strategies students use in the software? And if so, which strategies?* We studied this question by computing correlations between verbalized metacognition and the strategies identified via the coherence analysis described in Section 4.2.

Results in Table 3 show that verbalized metacognition was positively related to use of metacognitive strategies, as we had hypothesized, for some strategies. In particular, coherent quiz view (*rho* = .201, p = .046) and coherent mark (*rho* = .217, p = .031) were significantly positively related to verbalized metacognition, while other strategies were not. However, even insignificant correlations showed a positive trend, suggesting further support for the RQ3 hypothesis given a larger sample size; the probability of five relationships each having the same sign (either positive or negative) is 6.25% (i.e., p = 0.0625; marginally significant).

Table 3: Correlations between verbalized metacognition and metacognitive strategy use. *indicates p < .05.

Metacognitive Strategy	**Correlation (*rho*)**
Coherent quiz view	.201*
Coherent mark	.217*
Coherent edit	.047
Coherent read	.124
Coherent feedback	.116

6 DISCUSSION

Overall, results partially supported two of our three hypotheses (RQ2 and RQ3), did not support one (RQ1), and elucidated the role of verbalized metacognition in computerized learning environments. In this section, we discuss these results and their practical implications for HCI, then point toward future work to address limitations inherent to our study and offer concluding remarks.

6.1 Main Findings

Based on previous theoretical and empirical work [20,26,38,52], we expected verbalized metacognition would be positively related to learning (RQ1). We found that this was not the case. However, interviews themselves were correlated with learning, and especially so for students who expressed more verbalized metacognition. It may thus be the case that students used the interviews as an avenue for resolving metacognitive experiences like confusion. Furthermore, given that students themselves could initiate interviews by getting the attention of the interviewers, it is possible that student-initiated interviews were more likely to be for students who were metacognitively aware of knowledge gaps or sources of confusion that they wished to discuss. However, it is also possible that the interviews served as an opportunity for students to engage in self-explanation, which can benefit learning [12]. Explanation activities also benefit tutor learning in peer tutoring contexts [49], which aligns with the learning-by-teaching approach of Betty's Brain.

Our second hypothesis (RQ2) focused more specifically on the role of verbalized metacognition for confusion resolution, given previous research suggesting a coupling between metacognition and confusion [11,64]. Results did not support the hypothesis that confused students would verbalize more metacognition, but further suggested that students used interviews as an opportunity to discuss (and, hopefully, resolve) their recent and ongoing confusion experiences.

For our third research question (RQ3), we expected that our novel measure of verbalized metacognition would show convergent validity with measures of metacognition derived from logs of students' software interactions [54,64]. This hypothesis was partially supported; some – but not all – measures of metacognitive strategies correlated with verbalized metacognition. However, even non-significant correlations trended positive, as expected from our hypothesis. These findings indicate that that students' software interaction behaviors mirror their verbalizations of metacognitive strategies – i.e., that students are often consciously aware of the strategies they are implementing in the software. Note, however, that coherence analysis is intended to capture specific two facets of metacognition (namely, metacognitive strategies and action), while our verbalized metacognition measurement method is intended to capture three facets of metacognition (including metacognitive knowledge and experiences) – though not metacognitive action. Thus, we would expect some differences between verbalized metacognition and metacognition measured from interaction logs, even in a best-case scenario where both measures were free from measurement error.

In sum, we found limited support for the hypothesized link between verbalized metacognition and learning, and stronger support for the hypothesized links between verbalized metacognition and measures of confusion and metacognitive strategies measured from interaction log files. Overall, these findings indicated that interviews conducted during students' interactions with computerized learning environments do offer some insight into their metacognitive states.

6.2 Implications for HCI

Both our methodology and results offer implications for HCI research and practice. Assessing metacognition via interviews is a generalizable method with applications to research and design in situations like think-aloud protocols, where the transcripts of users' thoughts could be analyzed to assess the role of metacognition in their interactions. Similarly, in speech-oriented computer interfaces (voice user interfaces; e.g., Amazon Alexa, Google Nest), analysis of automatically-transcribed speech may provide insight into users' problem-solving processes and experiences using these interfaces.

Computerized education environments may also be improved by implementing metacognition-focused interview-like functionality. As observed in results for RQ1, completing more interviews correlated with higher learning. While

the existing Betty's Brain interface includes some support for conversations with the virtual agents (Betty and Mr. Davis), results show potential for further improving learning via additional interview-like functionality. Such functionality may be informed by the results of this paper, to specifically focus on promoting metacognitive evaluations and self-explanations from students. Note, however, that metacognitive evaluation alone is insufficient without motivation, confidence, and intention to act on those evaluations [2,28], which should also be considered during interface design.

Results from RQ2 offer possible implications for how adaptive learning software might select points in time to automatically suggest a metacognitive intervention like an interview. In particular, the automatic measures of affective state sequences incorporated in Betty's Brain showed that unresolved confusion correlated with less verbalized metacognition, and suggested (though not significantly) that ongoing and recently-resolved confusion might result in more. Since RQ1 results also indicated that interviews might be especially beneficial for highly-metacognitive students, the points in time where ongoing confusion is automatically detected offer opportunities to provide a metacognition intervention. Selecting the right time is essential, given that related research on think-aloud methods implies that an interview that interrupts students would have negative effects [50]. Conversely, an appropriately timed intervention could help resolve confusion before students transition to boredom and disengage from metacognitive activity.

RQ3 results further support the idea that interviews or similar procedures could be effective metacognitive interventions. If the link between verbalized metacognition and metacognitive strategy use is causal (i.e., if interviews are effective metacognitive prompts), interventions could focus on encouraging such strategies, perhaps describing them to users as a means of confusion resolution.

Considered together, the findings in this paper suggest that NLP analysis of interview transcripts is a promising method for gaining further understanding into users' metacognition, and that interviews that resolve metacognitive experiences of confusion may be a way to promote learning in computerized environments.

6.3 Limitations and Future Work

The implications of this study suggest several avenues for future research, including incorporating more interview-like functionality directly into Betty's Brain and experimenting with automatically timing interviews to coincide with periods of confusion. There are, however, limitations that should also be addressed in future work. The study reported in this paper was observational, and was primarily intended to yield insights for future work from analysis of interview transcripts. This limited our ability to make key causal inferences; for example, do interviews indeed improve learning via confusion resolution, or are students experiencing confusion simply more likely to learn because the confusion itself leads to learning? However, the observational design of this study allowed us to maximize statistical power (versus dividing students into conditions and comparing conditions), which was essential for this study given the difficulty of recruiting schools, teachers, and students to participate in a classroom study. Future work will test potential automated interview-based interventions informed by the results in this paper.

There are also two notable sources of unexplained variance in the measures we used. First, we only measured verbalized metacognition during interviews, though it is possible that students engaged in similar metacognitive activities during peer-talk or even self-talk. Future work could compare these other sources of verbalized metacognition by recording students constantly, rather than only during interviews. Second, the pre- and post-tests we administered to measure learning were identical, so it is possible that students learned from the test and were more prepared for the post-test because of that. This source of variance might have contributed to the lack of evidence for a relationship between verbalized metacognition and learning that we observed in RQ1, and should be explored more in future work.

The study in this paper was also limited to a single physical site and set of students, which may restrict generalization of findings to other populations. However, this allowed us to collect a large number of in-person interviews with limited researchers, which enabled detection of some fine-grained relationships like the correlations between verbalized metacognition and metacognitive strategies. Future work will also be needed to address this limitation, and may be enabled by more scalable automated interview-based interventions.

7 CONCLUSION

The research in this paper was motivated by a lack of understanding regarding how metacognition measured in the moment (and verbalized metacognition in particular) relates to learning outcomes, experiences, and usage patterns in educational software. Addressing this topic via NLP analysis of interview transcripts allowed us to triangulate metacognition with expected patterns of confusion and metacognitive strategies, which showed that verbalized metacognition is indeed linked to interaction behaviors, and that getting students to talk about their metacognitive processes may be an effective way to improve learning in computerized learning environments. Ultimately, these findings will enable the design of educational software and instructional practices that improve student outcomes by supporting their metacognitive needs.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) Award #1561676.

REFERENCES

[1] Rakefet Ackerman and Morris Goldsmith. 2011. Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied* 17, 1 (2011), 18–32. DOI:https://doi.org/10.1037/a0022086

[2] Icek Ajzen and Martin Fishbein. 2005. The influence of attitudes on behavior. In *The handbook of attitudes*, Dolores Albarracín, Blair T. Johnson and Mark P. Zanna (eds.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 173–221.

[3] Alexandra Andres, Jaclyn Ocumpaugh, Ryan S. Baker, Stefan Slater, Luc Paquette, Yang Jiang, Nigel Bosch, Anabil Munshi, Allison L. Moore, and Gautam Biswas. 2019. Affect sequences and learning in Betty's Brain. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19)*, ACM, New York, NY, 383–390. DOI:https://doi.org/10.1145/3303772.3303807

[4] Roger Azevedo and Vincent Aleven. 2013. Metacognition and learning technologies: An overview of current interdisciplinary research. In *International Handbook of Metacognition and Learning Technologies*, Roger Azevedo and Vincent Aleven (eds.). Springer, New York, NY, 1–16. DOI:https://doi.org/10.1007/978-1-4419-5546-3_1

[5] Roger Azevedo, Amy Johnson, Amber Chauncey, and Candice Burkett. 2010. Self-regulated learning with MetaTutor: Advancing the science of learning with metacognitive tools. In *New Science of Learning: Cognition, Computers and Collaboration in Education*, Myint Swe Khine and Issa M. Saleh (eds.). Springer, New York, NY, 225–247. DOI:https://doi.org/10.1007/978-1-4419-5716-0_11

[6] Gautam Biswas, Krittaya Leelawong, Daniel Schwartz, Nancy Vye, and The Teachable Agents Group at Vanderbilt. 2005. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence* 19, 3–4 (March 2005), 363–392. DOI:https://doi.org/10.1080/08839510590910200

[7] Gautam Biswas, James R. Segedy, and Kritya Bunchongchit. 2016. From design to implementation to practice - A learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education* 26, 1 (March 2016), 350–364. DOI:https://doi.org/10.1007/s40593-015-0057-9

[8] A.F. Blackwell. 1996. Metacognitive theories of visual programming: What do we think we are doing? In *Proceedings 1996 IEEE Symposium on Visual Languages*, IEEE, 240–246. DOI:https://doi.org/10.1109/VL.1996.545293

[9] Benjamin S. Bloom. 1968. Learning for mastery. *Evaluation Comment* 1, 2 (1968).

[10] Ivar Bråten and Marit S. Samuelstuen. 2007. Measuring strategic processing: Comparing task-specific self-reports to traces. *Metacognition Learning* 2, 1 (April 2007), 1–20. DOI:https://doi.org/10.1007/s11409-007-9004-y

[11] Gwendolyn E. Campbell, Natalie B. Steinhauser, Myroslava Dzikovska, Johanna D. Moore, Charles B. Callaway, and Elaine Farrow. 2009. *Metacognitive awareness versus linguistic politeness: Expressions of confusion in tutorial dialogues.* Naval Air Warfare Center Training Systems Div. Retrieved September 3, 2020 from https://apps.dtic.mil/sti/citations/ADA530033

[12] Michelene T. H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* 13, 2 (April 1989), 145–182. DOI:https://doi.org/10.1016/0364-0213(89)90002-5

[13] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, (1960), 37–46. DOI:https://doi.org/10.1177/001316446002000104

[14] Scotty D. Craig, Xiangen Hu, Arthur C. Graesser, Anna E. Bargagliotti, Allan Sterbinsky, Kyle R. Cheney, and Theresa Okwumabua. 2013. The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education* 68, (October 2013), 495–504. DOI:https://doi.org/10.1016/j.compedu.2013.06.010

[15] Scotty Craig, Art Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media* 29, 3 (2004), 241–250. DOI:https://doi.org/10.1080/1358165042000283101

[16] Betsy DiSalvo. 2016. Participatory design through a learning science lens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), Association for Computing Machinery, New York, NY, USA, 4459–4463. DOI:https://doi.org/10.1145/2858036.2858405

[17] Sidney K. D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (April 2012), 145–157. DOI:https://doi.org/10.1016/j.learninstruc.2011.10.001

[18] Sidney K. D'Mello and Arthur C. Graesser. 2014. Confusion. In *International Handbook of Emotions in Education*, Reinhard Pekrun and Lisa Linnenbrink-Garcia (eds.). New York, NY: Routledge, 289–310.

[19] Sidney K. D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29, 1 (2014), 153–170.

[20] Yehudit Judy Dori, Zemira R. Mevarech, and Dale R. Baker (Eds.). 2018. *Cognition, Metacognition, and Culture in STEM Education*. Springer, Cham, CH. Retrieved from https://doi.org/10.1007/978-3-319-66659-4

[21] Jason E. Dowd, Ives Araujo, and Eric Mazur. 2015. Making sense of confusion: Relating performance, confidence, and self-efficacy to expressions of confusion in an introductory physics class. *Phys. Rev. ST Phys. Educ. Res.* 11, 1 (March 2015), 010107. DOI:https://doi.org/10.1103/PhysRevSTPER.11.010107

[22] David Dunning. 2011. Chapter five - The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology*, James M. Olson and Mark P. Zanna (eds.). Academic Press, 247–296. DOI:https://doi.org/10.1016/B978-0-12-385522-0.00005-6

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

[24] Anastasia Efklides. 2008. Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist* 13, 4 (January 2008), 277–287. DOI:https://doi.org/10.1027/1016-9040.13.4.277

[25] John H. Flavell. 1976. Metacognitive aspects of problem solving. In *The Nature of Intelligence*, L. B. Resnick (ed.). Erlbaum, Hillsdale, NJ, 231–236.

[26] John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist* 34, 10 (1979), 906–911. DOI:https://doi.org/10.1037/0003-066X.34.10.906

[27] Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. 2019. *irr: Various coefficients off interrater reliability and agreement.*

[28] Peter M. Gollwitzer and Bernd Schaal. 1998. Metacognition in action: The importance of implementation intentions. *Personality and Social Psychology Review* 2, 2 (May 1998), 124–136. DOI:https://doi.org/10.1207/s15327957pspr0202_5

[29] George M. Harrison and Lisa M. Vallin. 2018. Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning* 13, 1 (April 2018), 15–38. DOI:https://doi.org/10.1007/s11409-017-9176-z

[30] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human earning and teaching. *Int J Artif Intell Educ* 24, 4 (2014), 470–497. DOI:https://doi.org/10.1007/s40593-014-0024-x

[31] Eddie Huang, Hannah Valdiviejas, and Nigel Bosch. 2019. I'm sure! Automatic detection of metacognition in online course discussion forums. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, IEEE, Piscataway, NJ, 241–247. DOI:https://doi.org/10.1109/ACII.2019.8925506

[32] Yang Jiang, Nigel Bosch, Ryan S. Baker, Luc Paquette, Jaclyn Ocumpaugh, Juliana Ma. Alexandra L. Andres, Allison L. Moore, and Gautam Biswas. 2018. Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In *Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED 2018)*, Springer, Cham, CH, 198–211. DOI:https://doi.org/10.1007/978-3-319-93843-1_15

[33] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), Association for Computing Machinery, New York, NY, USA, 1375–1386. DOI:https://doi.org/10.1145/3025453.3025592

[34] John S. Kinnebrew, James R. Segedy, and Gautam Biswas. 2014. Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition Learning* 9, 2 (August 2014), 187–215. DOI:https://doi.org/10.1007/s11409-014-9112-4

[35] James A. Kulik and J. D. Fletcher. 2016. Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research* 86, 1 (March 2016), 42–78. DOI:https://doi.org/10.3102/0034654315581420

[36] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. DOI:https://doi.org/10.2307/2529310

[37] Blair Lehman and Art Graesser. 2015. To resolve or not to resolve? That is the big question about confusion. In *Artificial Intelligence in Education* (Lecture Notes in Computer Science), Springer International Publishing, Cham, CH, 216–225. DOI:https://doi.org/10.1007/978-3-319-19773-9_22

[38] Diane Litman and Kate Forbes-Riley. 2013. Towards improving (meta)cognition by adapting to student uncertainty in tutorial dialogue. In *International Handbook of Metacognition and Learning Technologies*, Roger Azevedo and Vincent Aleven (eds.). Springer, New York, NY, 385–396. DOI:https://doi.org/10.1007/978-1-4419-5546-3_25

[39] Dastyni Loksa, Amy J. Ko, Will Jernigan, Alannah Oleson, Christopher J. Mendez, and Margaret M. Burnett. 2016. Programming, problem solving, and self-awareness: Effects of explicit guidance. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), Association for Computing Machinery, New York, NY, USA, 1449–1461. DOI:https://doi.org/10.1145/2858036.2858252

[40] Richard E. Mayer. 2016. The role of metacognition in STEM games and simulations. In *Using Games and Simulations for Teaching and Assessment: Key Issues*, Harold F. O'Neil, Eva L. Baker and Ray S. Perez (eds.). Routledge, 183–205.

[41] Anabil Munshi, Ramkumar Rajendran, Jaclyn Ocumpaugh, Gautam Biswas, Ryan S. Baker, and Luc Paquette. 2018. Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (UMAP '18), Association for Computing Machinery, New York, NY, USA, 131–138. DOI:https://doi.org/10.1145/3209219.3209241

[42] Joseph D. Novak and Alberto J. Cañas. 2008. *The theory underlying concept maps and how to construct and use them.* Florida Institute for Human and Machine Cognition (IHMC).

[43] Ernesto Panadero. 2017. A review of self-regulated learning: Six Models and four directions for research. *Frontiers in Psychology* 8, (2017). DOI:https://doi.org/10.3389/fpsyg.2017.00422

[44] Nancy E. Perry and Philip H. Winne. 2006. Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educ Psychol Rev* 18, 3 (September 2006), 211–228. DOI:https://doi.org/10.1007/s10648-006-9014-3

[45] James Prather, Raymond Pettit, Brett A. Becker, Paul Denny, Dastyni Loksa, Alani Peters, Zachary Albrecht, and Krista Masci. 2019. First things first: Providing metacognitive scaffolding for interpreting problem prompts. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (SIGCSE '19), Association for Computing Machinery, New York, NY, USA, 531–537. DOI:https://doi.org/10.1145/3287324.3287374

[46] James Prather, Raymond Pettit, Kayla McMurry, Alani Peters, John Homer, and Maxine Cohen. 2018. Metacognitive difficulties faced by novice programmers in automated assessment tools. In *Proceedings of the 2018 ACM Conference on International Computing Education Research* (ICER '18), Association for Computing Machinery, New York, NY, USA, 41–50. DOI:https://doi.org/10.1145/3230977.3230981

[47] R Core Team. 2013. R: A language and environment for statistical computing. (2013).

[48] Steve Ritter, Michael Yudelson, Stephen E. Fancsali, and Susan R. Berman. 2016. How mastery learning works at scale. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale* (L@S '16), ACM, New York, NY, 71–79. DOI:https://doi.org/10.1145/2876034.2876039

[49] Rod D. Roscoe and Michelene T. H. Chi. 2007. Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research* 77, 4 (December 2007), 534–574. DOI:https://doi.org/10.3102/0034654307309920

[50] Jonathan W. Schooler, Stellan Ohlsson, and Kevin Brooks. 1993. Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General* 122, 2 (1993), 166–183. DOI:https://doi.org/10.1037/0096-3445.122.2.166

[51] Gregory Schraw. 2009. A conceptual analysis of five measures of metacognitive monitoring. *Metacognition Learning* 4, 1 (April 2009), 33–45. DOI:https://doi.org/10.1007/s11409-008-9031-3

[52] Gregory Schraw, Kent J. Crippen, and Kendall Hartley. 2006. Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education* 36, 1 (March 2006), 111–139. DOI:https://doi.org/10.1007/s11165-005-3917-8

[53] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary Educational Psychology* 19, 4 (1994), 460–475.

[54] James R. Segedy, John S. Kinnebrew, and Gautam Biswas. 2015. Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics* 2, 1 (May 2015), 13–48–13–48. DOI:https://doi.org/10.18608/jla.2015.21.3

[55] Valerie J. Shute, Matthew Ventura, and Yoon Jeon Kim. 2013. Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research* 106, 6 (2013), 423–430.

[56] Kimberly D. Tanner. 2012. Promoting student metacognition. *CBE—Life Sciences Education* 11, 2 (June 2012), 113–120. DOI:https://doi.org/10.1187/cbe.12-03-0033

[57] Marcel V. J. Veenman, Laura Bavelaar, Levina De Wolf, and Marieke G. P. Van Haaren. 2014. The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences* 29, (January 2014), 123–130. DOI:https://doi.org/10.1016/j.lindif.2013.01.003

[58] Marcel V. J. Veenman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: Conceptual and methodological considerations. *Metacognition Learning* 1, 1 (April 2006), 3–14. DOI:https://doi.org/10.1007/s11409-006-6893-0

[59] Kim-Phuong L. Vu, Gerard L. Hanley, Thomas Z. Strybel, and Robert W. Proctor. 2000. Metacognitive processes in human-computer Interaction: Self-assessments of knowledge as predictors of computer expertise. *International Journal of Human–Computer Interaction* 12, 1 (May 2000), 43–71. DOI:https://doi.org/10.1207/S15327590IJHC1201_2

[60] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376732

[61] Margaret C. Wang, Geneva D. Haertel, and Herbert J. Walberg. 1990. What influences learning? A content analysis of review literature. *The Journal of Educational Research* 84, 1 (September 1990), 30–43. DOI:https://doi.org/10.1080/00220671.1990.10885988

[62] Shang Wang, Deniz Sonmez Unal, and Erin Walker. 2019. MindDot: Supporting effective cognitive behaviors in concept map-based learning environments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3290605.3300258

[63] Shang Wang, Erin Walker, and Ruth Wylie. 2017. What matters in concept mapping? Maps learners create or how they create them. In *Artificial Intelligence in Education* (Lecture Notes in Computer Science), Springer International Publishing, Cham, CH, 406–417. DOI:https://doi.org/10.1007/978-3-319-61425-0_34

[64] Yingbin Zhang, Luc Paquette, Ryan S. Baker, Jaclyn Ocumpaugh, Nigel Bosch, Anabil Munshi, and Gautam Biswas. 2020. The relationship between confusion and metacognitive strategies in Betty's Brain. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK20)*, ACM, New York, NY, 276–284. DOI:https://doi.org/10.1145/3375462.3375518