

Curiosity Notebook: The Design of a Research Platform for Learning by Teaching

KEN JEN LEE*, University of Waterloo, Canada

APOORVA CHAUHAN*, Concordia University of Edmonton, Canada

JOSLIN GOH, University of Waterloo, Canada

ELIZABETH NILSEN, University of Waterloo, Canada

EDITH LAW, University of Waterloo, Canada

While learning by teaching is a popular pedagogical technique, it is a learning phenomenon that is difficult to study due to variability in the tutor-tutee pairings and learning environments. In this paper, we introduce the Curiosity Notebook, a web-based research infrastructure for studying learning by teaching via the use of a teachable agent. We describe and provide rationale for the set of features that are essential for such a research infrastructure, outline how these features have evolved over two design iterations of the Curiosity Notebook and through two studies—a 4-week field study with 12 elementary school students interacting with a NAO robot and an hour-long online observational study with 41 university students interacting with an agent—demonstrate the utility of our platform for making observations of learning-by-teaching phenomena in diverse learning environments. Based on these findings, we conclude the paper by reflecting on our design evolution and envisioning future iterations of the Curiosity Notebook.

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**.

Additional Key Words and Phrases: teachable agents; conversational interfaces; learning by teaching; research infrastructure; educational technology

ACM Reference Format:

Ken Jen Lee, Apoorva Chauhan, Joslin Goh, Elizabeth Nilsen, and Edith Law. 2021. Curiosity Notebook: The Design of a Research Platform for Learning by Teaching. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 394 (October 2021), 26 pages. <https://doi.org/10.1145/3479538>

1 INTRODUCTION

Learning by teaching is a popular and well-studied pedagogical technique, shown to produce the protégé effect [14]—students learn more effectively by teaching others. When teaching, students synthesize and structure materials, become aware of their own learning process, and expend more effort to learn. Learning by teaching has been shown to benefit the tutor academically and psychologically [70]. These benefits include improving inquiry skills [3], transferring skills towards related domains [70], improving attitudes towards the material [18] and school [70], and bolstering self-confidence [80]. Further, certain activities during learning by teaching have been shown to enhance such benefits, including providing explanations, collaboratively building knowledge with the tutee, and receiving feedback about the tutee’s learning [23, 64, 65]. Learning by teaching has

*Both authors contributed equally to this research.

Authors’ addresses: Ken Jen Lee, kenjen.lee@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada; Apoorva Chauhan, apoorvathechauhan@gmail.com, Concordia University of Edmonton, Edmonton, Alberta, Canada; Joslin Goh, jtcgoh@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada; Elizabeth Nilsen, enilsen@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada; Edith Law, edith.law@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

This is the author’s version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the ACM on Human-Computer Interaction*, <https://doi.org/10.1145/3479538>.

found its use in a wide variety of learning settings, at different education levels (e.g., university [37], grade [16, 52], middle and high school [73]) and involving different content areas (e.g., linguistics, applied sciences [37], mathematics [57] and sciences [15, 52]). In education research, learning by teaching is closely related or synonymous to other terms, such as peer tutoring [23, 56, 70], cooperative learning [23], and peer-assisted learning [35].

Despite extensive research, our understanding of the specific conditions that make learning by teaching effective is limited because there are often substantial confounds introduced by the variability in the tutor-tutee pairings. One recommendation, put forth by Roscoe and Chi in a large survey on tutor learning [73], is to “develop teachable agents to test hypotheses about specific tutor behaviours” by systematically manipulating the teachable agent’s characteristics and question asking behaviour. The use of teachable agents as tutees allow researchers to overcome some challenges faced in conducting studies with human tutors and tutees, including the ability to precisely control the characteristics of the tutee so that participants are exposed to the same experimental stimuli and removing the potential risk of human tutees being disadvantaged in the process [56, 57, 73].

While many agent-mediated learning-by-teaching environments have been developed by researchers, most are not designed with the degree of flexibility required for reuse across a wide range of learning contexts [15, 57, 63, 64]. Our work seeks to fill this gap by introducing a learning-by-teaching platform called the Curiosity Notebook. Our premise is that such a research platform must be highly configurable to enable the investigation of learning by teaching in a large number of settings. This is important because, as Roscoe and Chi pointed out, tutor learning is not uniform across various settings and the reasons behind the variability are not well understood [73]. They also encouraged future researchers to analyze and compare tutor learning processes across settings [73]. Having a configurable research platform that supports a variety of settings and agent configurations is an important step towards facilitating comparisons across studies, by minimizing confounding factors and interactions [73]. However, such configurability is often not found because most platforms are built to investigate learning-by-teaching phenomena in very specific learning contexts. Moreover, designing a highly configurable platform requires a significant amount of resources and is often deemed unnecessary for these narrowly focused studies. Closest to our work is Betty’s Brain, where students teach a virtual agent (i.e., an avatar called Betty) about causal relationships through concept maps. Betty’s Brain has been used to study hierarchical reasoning [16], self-regulated learning [9, 60], collaborative teaching [24], the role of feedback [53, 76, 81], and scaffolded learning of multiple representations [6], to name a few; this variety suggests a certain level of configurability allowed by the platform for researchers to adjust according to each study’s specific needs. Our work goes beyond prior work by providing configurability for agent’s embodiment (e.g., text-based chatbots, voice agents or physical robots), coordinated group-based teaching (e.g., explicitly implemented turn-taking mechanisms that enable students to teach in groups of varying sizes), and learning task and materials (e.g., grade school vs. university students), all of which are discussed in this paper.

In this paper, we outline the set of configurable features in Section 3, informed by prior literature, that are useful for studying learning by teaching and essential for allowing better cross-study comparisons to be made. We provide a detailed description of the platform’s design evolution in Section 4, as informed by observations of two deployments—elementary school children participating in groups to teach physical robots in person, and university students working individually to teach a chatbot online. These deployments serve two purposes: first, they inform potential design changes that could be made for subsequent iterations of the platform; second, they serve to demonstrate the platform’s versatility as a research infrastructure in collecting useful observations for understanding learning-by-teaching phenomena in a wide range of learning contexts. In Section

5, we describe a number of observations collected using Curiosity Notebook in these deployments, as evidence for the utility of the platform and the importance of the configurability allowed by it, and conclude in Section 6 with a discussion of the envisioned design improvements for future iterations of the platform.

2 BACKGROUND

2.1 Learning by teaching

As noted above, learning by teaching produces the *protégé effect* [14]—students who are asked to teach others demonstrate more proficiency in synthesizing and structuring materials, become aware of their own learning process, and expend additional effort within the learning process. The benefit of learning by teaching comes from the additional knowledge-building activities that are inherent within the process of providing explanations for others, which in turn promotes long-term understanding and consolidation of information [26–28, 42]. Learning by teaching also allows students to better assess their own understanding of concepts and identify the gaps in their knowledge, thereby enhancing self-learning by repairing these gaps [50]. For instance, prior work has shown that students exhibit greater gains in knowledge when actively teaching, compared to just preparing to teach [42].

The benefits from learning by teaching also come from learners' motivations. Benware et al. [8] hypothesized that individuals who learn material with the purpose of teaching take a more active learning role and have higher levels of intrinsic motivation. An intrinsically motivated learner engages with learning due to personal enjoyment and interests, instead of external rewards and incentives (e.g., higher test scores). According to the cognitive evaluation theory, intrinsic motivation is a culmination of self-perception of one's own autonomy (the feeling that actions can affect the surroundings) and competence (the feeling of having mastered a topic) [54]. Benware et al., in [8], explained that learning for the purpose of teaching makes the learning process purposeful, as it increases autonomy which, in turn, enhances intrinsic motivation. Their study showed that, compared to participants who learn new material to take a test, those who learned to teach were more interested in the material and enjoyed the study more, supporting the hypothesis that learning to teach increases intrinsic motivation.

As mentioned in the introduction, learning by teaching also has various other benefits, including psychological benefits. A meta-analysis by Robinson and Schofield [70] found that tutoring other students increases the tutor's sense of belonging to their school and with other students. Moreover, tutors who taught materials from one domain of mathematics performed better in other domains of mathematics as well, showing improvements in tutors' transferring of skills. Other studies found improvements in the tutors' self-concept (defined as the "organization of qualities that the individual attributes to himself" [45]) [30], self-awareness and self-direction [79].

The degree to which programs involving learning by teaching succeed is influenced by numerous program characteristics, e.g., how much instruction the tutors received, age difference between tutor and tutee, how the tutors and tutees are chosen for the study. Moreover, many studies fail to tease apart the effects of teaching from other effects like collaborative learning (defined as when "two or more people learn or attempt to learn something together" [21]) [70]. Developing *teachable agents* as a method of systematically studying learning by teaching, as suggested by Roscoe and Chi [73], can help overcome some of these limitations. In particular, with teachable agents as tutees, researchers have full control over the tutees' characteristics and how the tutor and tutee interact. This means that confounding factors and other closely related effects (e.g., collaborative learning) can be controlled or accounted for. More specifically, teachable agents can allow researchers to study

students' responses to learning by teaching, including differential responses as the characteristics and behaviour of the agent are manipulated.

2.2 Teachable Conversational Agents

Within educational settings, conversational agents have adopted numerous roles. As experts, agents improve information acquisition; as motivators, agents increase self-efficacy; and as mentors, agents improve both learning and motivation [7]. Tutor agents have been successful in tasks including writing algebraic expressions [39], learning about urban ecosystems [36], and foreign language learning [33].

Only a handful of systems position the agent as a less intelligent or knowledgeable "peer" that students teach. These agents are referred to as teachable agents. Using teachable agents to research learning by teaching has various benefits, including 1) enabling precisely determined control conditions not possible with human tutees, 2) reducing the risk of any human tutees from being harmed in the process, and, importantly, 3) facilitating the collection of detailed interaction data between the teacher and the agent [57]. SimStudent, for example, is a simulated learner used to study student-tutor learning in mathematics problem solving [57]. It was built with the primary purpose of collecting detailed interaction data when students are teaching the teachable agent and investigating the cognitive and social theoretic underpinnings of when, how and why learning by teaching is effective. Students teach SimStudent by posing algebraic problems to it, providing feedback and help while SimStudent attempts to solve the problems, and quizzing SimStudent to gauge learning. SimStudent was found to be effective for learning procedural, but not conceptual, skills. Further, it was more effective when students taught it correctly and made good use of the feedback and quizzing features. Another example is the Betty's Brain system [9, 53], a teachable agent learning environment in which students read articles, then teach, query and quiz a virtual agent ("Betty") about causal relationships (e.g., burning fossil fuels increases CO₂) in science by manipulating concept maps. It is designed to help students develop structured networks of knowledge, take responsibility and make decisions about learning, and develop reflection or meta-cognitive skills [9], thereby preparing them for future learning.

Other researchers have explored the use of teachable agents that involve physical robots [13, 41, 82, 83, 89] and showed the importance of incorporating recursive feedback within the design of effective agents (embodied or otherwise) for learning by teaching. Recursive feedback is when teachers observe their tutees use knowledge they taught. Okita et al. [64, 65] found that participants who observed their tutees interacting with examiners learned better than participants who only performed learning by teaching. In another study, high-school students who observed the agents they taught compete with other agents learned more than high-school students who taught agents and proceed to compete against other agents themselves [64]. It turns out that this effect is also apparent when the participants are playing the role of the examiner (instead of having a separate entity play the examiner) [53].

While teachable agents provide many benefits over human tutees, such work is labour-intensive, requiring meticulous design and implementation (i.e., coding). As such, having a research platform reusable across many learning-by-teaching studies would be significantly beneficial.

3 ESSENTIAL FEATURES OF A RESEARCH PLATFORM FOR LEARNING BY TEACHING

Our goal is to create a research platform for investigating learning-by-teaching phenomena; though, certainly, there would be educational, non-research applications as well. As discussed before, a platform needs to be highly configurable to allow the investigation of learning by teaching across diverse settings. Having a single platform support multiple settings is crucial for enhancing

understanding of the complex interactions in learning by teaching and for allowing cross-study comparisons (as encouraged by [73]). Below, we discuss the configurable features (CF), their significance, and the type of research questions they support.

CF#1 Agent Characteristics. Our platform enables systematic modulation of agent characteristics hypothesized to be relevant to learning (e.g., types of question asked, amount and accuracy of content demonstrated when tested). Tutee characteristics is a commonly studied and manipulated dimension within relevant literature, and is an integral part of learning-by-teaching experiences [9, 62, 66]. Using an agent allows for precise control over tutee characteristics and behaviours not possible with human tutees [57, 73]. For instance, modifying teachable agents to have them provide feedback to their teacher and ask authentic questions improves the student tutor's engagement with the material [66]. Other studied or discussed characteristics include the degree to which agents should take control of the conversation [62] and the agents' self-regulation behaviour [9]. Similar to how the variability in human tutees behaviours [73] limits generalizability of findings, the variability in agent characteristics across studies makes it hard for meaningful comparisons to be made. Thus, there is an important need for a research platform that allows for careful and granular modifications of tutee characteristics. Having configurable agent characteristics would also allow systematic analyses and verification of findings from human-to-human learning-by-teaching studies. An example is Roscoe's study that found that human tutee's questions were significant predictors of human tutors' knowledge building and deep understanding [71]. Roscoe suggested future studies to investigate how the absolute and relative levels of tutor and tutee expertise affect the amount of knowledge building; something that is much easier to investigate using agents since it is possible to fix the level of expertise in an agent tutee, making this variable consistent across tutors [71]. By providing functionalities to configure agent characteristics, our platform can support new experiments that investigate the effects of tutor-tutee pairings on learning by teaching.

CF#2 Quantification of Teaching Strategies. Our platform provides students with choices of teaching activities (e.g., teaching vs. testing), in order to allow for a quantitative characterization of their teaching strategies. This is essential as teaching strategies could be analyzed to better understand what and why certain teaching activities diminish or enhance benefits related to learning by teaching. For instance, Munshi et al. [60] found that students who were high performers employed different patterns of teaching activities than low performers within the Betty's Brain platform. High performing students read relevant articles after quizzing Betty more frequently than low performing students. Using the same platform, Wagster et al. [87] analyzed how high- and low-performing students employ the six available teaching activities while teaching the agent. They found that high-performing students incorporated more activities related to information seeking and self-monitoring than low-performing students. Beyond Betty's Brain, however, there is a lack of research on the effects of teaching activities and strategies on the learning of the tutor teaching an agent tutee; as such, our understanding of these effects is limited to tasks for learning causal relationships between concepts, which Betty's Brain supports. An important configurable feature of our platform, therefore, is a set of user interface controls (e.g., buttons) that student tutors can invoke to initiate a range of different teaching conversations with the agent. These user interface controls enable researchers to observe what teaching activities the student tutors would choose to engage in and in what order, thus providing rich data for capturing learning-by-teaching behaviour. This can help towards understanding not only *what*, but *how* and *why* tutors learn from teaching, which have been identified as notable gaps within the extant research studies [73].

CF#3 Scalable Learning Task and Material. Our platform supports learning tasks that can be scaled in complexity to different age groups (e.g., usable by elementary school as well as university students). Existing learning-by-teaching research involving teachable agents (i.e., agent tutees) are mostly for younger children (usually in elementary school) [9, 24, 52, 53, 60, 87]. Research

on learning by teaching for older students (e.g., university students), on the other hand, mostly investigated human tutor/tutee contexts without the use of teachable agents [2, 19, 55, 71, 72]. Research have also shown that the learning process itself evolves with age. For example, although the associative learning process might remain constant in children and adults, executively controlled learning (i.e., learning that requires engagement on the learner's part) increases in frequency with older ages [48]. Metacognitive skills also develop with age; adults are more capable of such skills than children, including taking more initiative to verbally rehearse during memory tasks [44], improve on their retrieval skills given previous failed attempts [29], and employ knowledge acquisition strategies [47]. As such, while learning by teaching has been found to benefit both children and adults alike, some important questions remain unanswered: 1) are teachable agents as effective as human-to-human learning by teaching for older students? and 2) how is the effectiveness of teachable agents for young children similar or different for older students? Having a single platform that supports multiple age groups and learning task complexities allows for a direct comparison between these age groups, and eventually allowing learning by teaching to benefit students of different ages through informed and targeted configurations.

CF#4 Coordinated Group-Based Teaching. Our platform supports students teaching individually, in pairs or in larger groups, explicitly controlling the turn-taking process in order to provide equal access to teaching opportunities. Collaborative learning, defined as “the grouping and pairing of learners for the purpose of achieving a learning goal”, has many benefits [49], which include encouraging critical thinking and increasing students' motivation, and promoting active student participation and personalized learning [49]. In teachable agents contexts, however, the only studies comparing individual versus collaborative learning by teaching, are those using Betty's Brain. Namely, Emara et al. [24] found that when paired to teach Betty together, students learned more and created more accurate concept maps than when working individually. This effect emerged because working in pairs allowed students to discuss errors in quizzes and their concept maps, which in turn encouraged more systematic approaches to learning and teaching the agent [24]. While the findings are significant, many questions remain about the inner working of learning by collaboratively teaching teachable agents. In particular, it is not well understood how a group's dynamics affect each tutor's learning. Within collaborative learning literature, researchers found that its effectiveness is affected by students' personality (e.g., extroverted vs. introverted), cognitive style [58], motivation levels [69], and the group's contribution equity [77]. For instance, Shah et al. [77] found that students who perceived themselves as being less competent in the subject were dominated in discussions within dyads, and as a result, were negatively affected in their learning process. These results raise questions of how teachable agent systems should be designed to manage group dynamics, e.g., encourage or enforce equal engagement and contribution among tutors. By providing functionalities that coordinate group-based teaching, our platform enables new questions about collaborative learning by teaching to be answered.

CF#5 Flexible Agent Embodiments. Our platform supports teaching conversations with text-based, voice-based or physical agents. Agent embodiments, beyond simple text interfaces (e.g., auditory speech, digital or physical animated forms), have been found to improve performance in tasks such as tutoring, where continuous interactions with agents are held [4, 59, 78]. However, agent embodiment is nuanced and rarely investigated in a thorough manner; these nuances are often listed as limitations in studies [40, 78, 84, 86]. Although out of this paper's scope, supporting flexible agent embodiment would allow for fair and detailed comparisons of the effects of embodiment on teaching and learning performance. For instance, one can investigate how agent embodiment affects rapport and how increased/decreased rapport would impact learning-by-teaching behaviours and outcomes.

To summarize, the five configurable features above are integral for researching learning by teaching with teachable agents. However, as discussed above, many questions remain unexplored for each configurable feature, and most studies use different learning-by-teaching platforms, making it hard to compare and synthesize generalizable findings that can be easily applied to the education sector. Although prior work has explored a few of the configurable features in depth, the effects of CF#3 and CF#5 remain unexplored. As such, we set out to build Curiosity Notebook to support all five configurable features, and introduce its design in the next section.

4 CURIOSITY NOTEBOOK: A RESEARCH PLATFORM FOR LEARNING BY TEACHING

In this paper, we introduce the Curiosity Notebook—a new research platform for learning by teaching—and describe how its design evolved over two years of development. We also demonstrate how the platform supports each configurable feature (e.g., CF#3) discussed in Section 3.

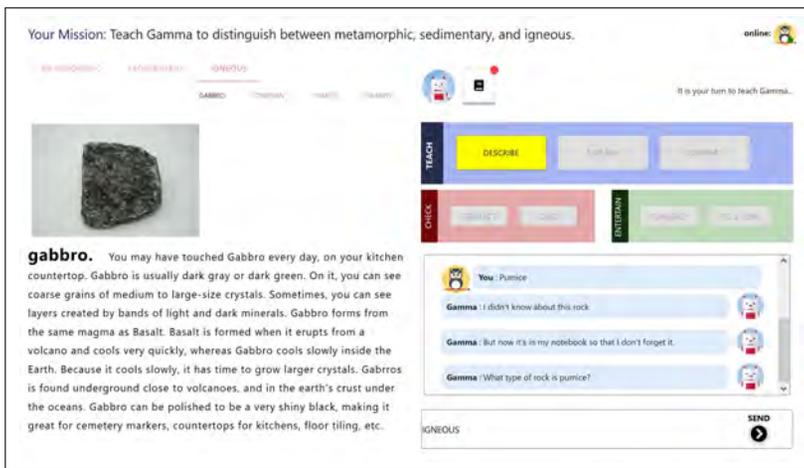


Fig. 1. Curiosity Notebook (second version) with reading panel (left) and teaching panel (right).

At a high-level, the Curiosity Notebook provides a web interface that students use to read articles and teach an agent to perform a task based on the reading. In our design thus far, we have focused exclusively on how the Curiosity Notebook enable the teaching of classification tasks—how to classify objects, e.g., animals as mammals, insects, and reptiles; rocks as sedimentary, igneous and metamorphic (rocks have also been used as a topic in previous studies on agents for education [11]); or paintings as pointillism, realism and impressionism. However, the task could take many forms; for example, students could read an article about programming and teach the agent how to correct mistakes in a piece of code, or students could read an article about grammar rules and teach the agent how to identify part of speech in a sentence. The main idea is that the platform supports the process of students teaching factual knowledge (obtained from reading) in a way that builds up and transforms into a particular skill (e.g., the ability to apply rules to classify objects) in the agent. For classification, the agent would need to be able to identify features of an object, identify the categories that the object can belong to, and map features to categories. The platform, as described in detail below, is designed to facilitate this text-to-rule translation. Shown in Figure 1, the main interface consists of a reading panel (left), and the teaching panel (right), which provides functionalities that enable students to communicate with the agent and assess the progress of the agent's learning.

4.1 Deployments

During our iterative design process, we deployed two versions of the Curiosity Notebook, each to a vastly different user population and learning setting. Before describing Curiosity Notebook’s design evolution, we first describe the details about the two deployments.

The deployments served two purposes—first, to inform the iterative design changes in the platform, and second, to demonstrate the platform’s versatility as a research infrastructure. For instance, the two deployments described below demonstrate how the platform can support users from CF#3 different age groups (elementary school children vs. university students), learning contexts (in-person vs. online), CF#4 group configurations (group-based vs individual teaching) and with CF#5 different agent embodiments (e.g., robot vs. chatbot). The differences between the deployments are summarized in Table 1.

Table 1. Comparison of deployment characteristics.

	Deployment 1	Deployment 2
Setting	In-person within school	Online
Length	4 weeks, 1.5 hours per week	Single session, 1 hour
Education level (CF#3)	Elementary school (Grades 4 & 5)	University
Number of students per agent (CF#4)	3	1
Agent embodiment (CF#5)	Physical Robot (NAO)	Chatbot

Deployment 1: Elementary School Students, In-Person. For the Curiosity Notebook’s first version, we conducted a 4-week exploratory study with 12 fourth- and fifth-grade students (7M/5F) at a local school [51]. Enrollment was on a first come, first served basis. No monetary compensation was provided; instead, students were given a “Certified Robot Teacher” certificate as a token of appreciation. The study was conducted in an after-school club, which ran once a week for 1.5 hours each. Four NAO robots were used in each session; to personalize the experience, each robot had a name tag hung around their neck with a gender-neutral name (i.e., Alpha, Beta, Gamma and Delta). Students formed groups of three, and taught the robot about a different topic (i.e., animals, rocks, paintings) each week, then all topics during the last week. Each student was given a Chromebook, and sat together with their group members facing the robot, which was positioned in a sitting posture in front of the students on the table (as shown in Figure 2). Each group of students was joined by a student researcher, who observed the group and resolved issues (if any) with the platform. We provided physical artifacts (as shown in Figure 2) for each classification task, namely animal figurines, rocks and minerals, and postcards of different styles of paintings from NYC Metropolitan Museum. During the session, we piloted a variety of surveys, iteratively re-designed the platform, made detailed observations, and interviewed students about their learning-by-teaching experience.

Deployment 2: University Students, Online. After the first deployment with elementary students, we revised the design of the Curiosity Notebook, and deployed it in a second observational study with university students. After a small pilot with 11 participants, we recruited 41 participants (22M, 19F; ages 18 to 29, $Mdn = 21.0$). All studies were conducted over Zoom in the presence of



Fig. 2. Students teaching a NAO robot (left) using the Curiosity Notebook and physical artifacts (right).

a researcher and were about 90-minutes long. Participants were given an \$20 CAD Amazon gift card as remuneration. Participants were asked to read articles about rocks and teach the agent (named Gamma) how to classify rocks as metamorphic, sedimentary, or igneous. We recorded all interactions (such as article clicks, button clicks, notebook checks, text logs of chat data between participant and the agent, and the amount of time a participant spent teaching Gamma) on the Curiosity Notebook as well as participants' responses to the administered pre-study and post-study questionnaires (submitted via Google Forms). The pre-study and post-study questionnaires measured participants' attitudes towards the agent (via likeability and perceived intelligence subscales of the Godspeed questionnaire [5]), and towards the teaching task (via Pick-a-Mood pictorial self-report scale [20] for self and agent). The Interest/Enjoyment, Pressure/Tensions and Effort/Importance subscales from the Intrinsic Motivation Inventory (IMI) [75] were used for a deeper understanding of participants' subjective experiences with the teaching task. The Academic Motivation Scale (AMS) was used to measure participants' amount and type of motivations towards the experiment [85]. Lastly, open-ended questions about participants' perceptions of the agent and the Curiosity Notebook were added for additional qualitative insights.

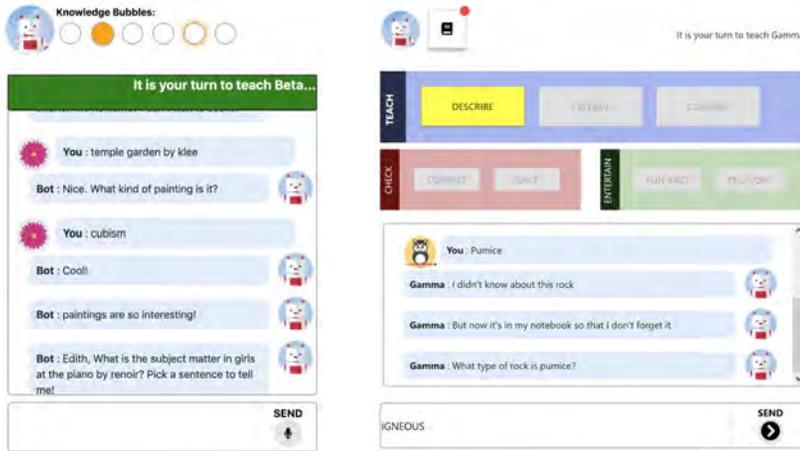
Beyond this paper, our platform has been used in a number of studies investigating the effects of agent characteristics (CF#1) on learning by teaching—to investigate agents with different humour styles in relation to student effort in teaching [12], and to study physically embodied agents that are capable of sensing group dynamics and verbally encouraging equal contributions among student tutors (which is also relevant to CF#4 and CF#5) [68]. The flexibility of the Curiosity Notebook platform (in the form of the five CFs) not only allows for its deployment in an expanded number of settings consisting of varying characteristics, but also allow researchers to expeditiously switch between these settings.

4.2 Design Evolution

We deployed the first version of the Curiosity Notebook to the first exploratory study with elementary school students, revised the design based on our observations and findings, then deployed the second version to the second exploratory study with university students. There were a number of features that evolved substantially between the two versions. We describe here these platform features and their design evolution.

4.2.1 Teaching Conversations. In the first version of the Curiosity Notebook, the agent begins each teaching conversation by asking the students to show it a physical artifact (e.g., “Can you pick an animal and tell me its name? I can’t wait to see it!”). After selecting an object, the teaching

conversation proceeds with the agent highlighting one of the “knowledge bubbles” and asking a series of 4 or 5 questions about the corresponding feature, as shown in Figure 3(a).



(a) First Version: knowledge bubbles indicating which features have been taught.

(b) Second Version: buttons for initiating teaching conversations and robot’s notebook for tracking learned knowledge

Fig. 3. Design Iterations of the Teaching Panel

The agent is designed to ask a mix of low- and high-level questions. Low-level thinking questions include questions about features (e.g., “What does the skin of mammals look like?”), examples (e.g., “Can you give me an example of a cubism painting?”), and facts (e.g., “Select a sentence to tell me about frogs and how they lay eggs.”). High-level thinking questions include why questions (e.g., “Why is a snake a reptile?”), synthesis questions (e.g., “Do all reptiles look the same?”), and questions that prompt students to repeat/rephrase or explain the meaning of a word, e.g., “Can you help me understand what you just said better?” These questions reflect different categories of cognitive operations posited by Gallagher and Aschner [34]. While low-level questions involve cognitive memory operations via the reproduction of factual knowledge directly from the articles, high-level questions involve convergent thinking via the discovery and learning of concepts (which take the form of classification rules in the context of classification tasks) from analyzing and integrating knowledge derived from the articles [34]; convergent thinking questions have been shown to be beneficial to students’ question-asking skills [1]. The system automatically generates a variety of questions by filling in predefined sentence templates with names of objects, features, and categories that the students are currently teaching. The questions are sequenced such that lower-level thinking questions always precede higher-level thinking questions, allowing students to first learn (through teaching) factual knowledge before concepts relating to classification rules. Some randomness was introduced for ordering questions to prevent the conversation from being too mechanical. The teachable agent occasionally seeks feedback from students about its learning, by asking questions about its general intelligence (e.g., “Am I smart?”), its learning progress (e.g., “Am I learning?”, “Do you think I know more now than before?”), or how well it might perform if tested (“Will I do well in a test?”).

After 4-5 rounds of questioning and answering, a “knowledge bubble” is filled and students are rewarded with confetti on the screen letting them know that the agent has “learned” that feature,

and the next feature is randomly chosen for students to teach. Between teaching conversations, students can also choose to test the agent’s knowledge. The testing interface shows a set of images representing objects (e.g., images of paintings) to be classified. Students can click on an image, and the agent will attempt to classify it (e.g., saying “I think it is an impressionist painting” or “I don’t know”). The system provides corrective feedback—an overlay over each image will show a green checkmark if the agent is correct, and a red “x” otherwise. This testing feature builds on prior research, which shows that student tutors learned more when observing their tutees use the knowledge they were taught either through some form of quiz or test [53, 64, 65], or through competition with other tutees [64].

A key change in the second version’s design is a set of controls that allows participants to choose specific actions to take next; this, in turn, allows us to more quantitatively capture the variety of teaching strategies employed (CF#2). In lieu of the knowledge bubbles, the teaching panel (as shown in Figure 3(b)) now contains, at the top, seven “conversation-type” buttons grouped into three categories, supporting *Teach*, *Check* and *Entertain* actions. When a button is clicked, the system locks the agent (named Gamma) and the participant into a conversation that involves 4-6 rounds of question and answering. The student tutor can engage in the conversation through free-form text, or by highlighting specific sentences in the text to teach the agent.

Table 2. *Explain* conversation in the second version of the Curiosity Notebook

Gamma:	It’s good to understand better why rocks look the way they do.
Gamma:	Can you pick a new rock and tell me what it’s called please?
User:	Shale
Gamma:	I don’t know about this rock.
Gamma:	But now it’s in my notebook so that I don’t forget it.
Gamma:	What category does shale belong to?
User:	Sedimentary
Gamma:	Can you tell me why shale looks the way it does?
Gamma:	Can you pick a sentence in the articles for me please?
User:	As more sediments get deposited, the particles underneath become tightly packed; eventually, they become a dense, solid rock.
Gamma:	Mmm.. okay I see.
Gamma:	Could you explain that more clearly?
User:	With time, sediments get deposited over each other, forming a dense solid rock.
Gamma:	Alright.
Gamma:	I’m really interested in rocks.
Gamma:	You can now select a new button to keep teaching me.

Each button initiates a teaching conversation with the agent. *Teach* actions include *Describe*, *Explain* and *Compare*. Upon clicking the *Describe* button, the agent will ask participants to identify an object’s category (e.g., “Pumice is an igneous rock.”) and feature (e.g., “Pumice is often white.”). This conversation allows students to better grasp factual knowledge, which is an essential step in developing competence in the topic taught [22]. The *Explain* conversation prompts for an explanation for why an object has a particular feature (e.g., “Pumice is often white because of the silica in the lava; without the silica, Pumice could be black.”). A sample *Explain* teaching conversation is shown in Table 2. Providing explanations is an important teaching activity that has proven to improve learning significantly, especially while learning the material [50]. The *Compare* conversation allows the participant to discuss similarities or differences between two objects. Ziegler and Stern found that when encouraged to make comparisons between conceptually

different material, students benefited from larger long-term learning gains [90]. In terms of cognitive operations, *Describe* supports cognitive memory while *Explain* and *Compare* support convergent thinking [34].

Check actions include *Correct* and *Quiz*. The *Correct* button allows the student to correct facts that they deemed the agent to have learned incorrectly (due to them teaching incorrectly; the agent in this version is not configured to inject errors automatically). This feature is designed to encourage metacognitive skills among student tutors. Specifically, by monitoring and identifying mistakes in the agent's understanding, student tutors may also identify similar flaws in their own understanding, since the agent learns directly from them. This is referred to as comprehension-monitoring, and has been found to be an essential part of knowledge building in learning by teaching [71]. To probe the current performance of the agent, students can click on the *Quiz* button, and select an object to test the agent on (Figure 4). The agent will classify the object correctly or incorrectly, based on its current knowledge model. As with the first version, the quiz feature encourages learning by providing feedback to the student tutors about the agent's learning progress.

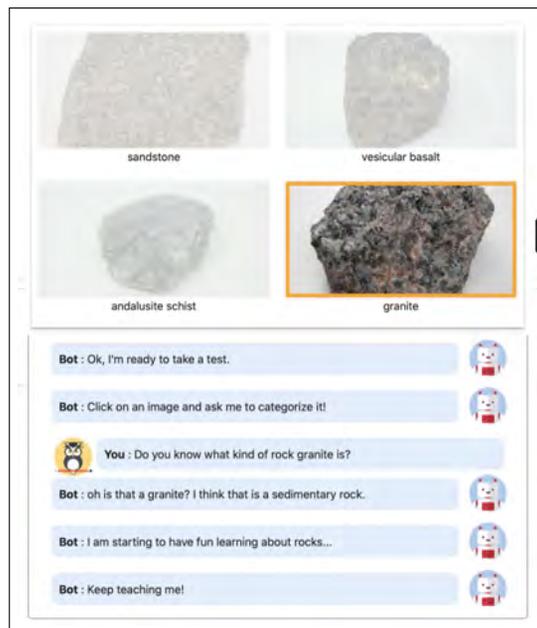


Fig. 4. *Quiz* conversation in the second version of the Curiosity Notebook

Finally, participants can choose to entertain the agent through the *Entertain* actions. The *Fun Fact* button allows the participant to tell a fun fact about an object (e.g., “Obsidian is used in heart surgery”) and provide a reason for why the fact is interesting. The *Tell Joke* button allows the participant to tell the agent a joke. Powell and Andersen found that the use of humour in learning settings, if not excessive, can benefit learning by increasing students' attention and motivation [67].

Together, these seven conversations are designed to support flexible teaching strategies on the side of the student tutors, and quantifiable teaching strategies that can be analyzed on the side of the researchers (CF#2). A final key difference between the first and second version of the Curiosity Notebook is the dialog system that drives the conversation. In the first version, the question templates were randomly drawn; whereas in the second version, each conversation is

controlled by a state machine, of which configurations are stored in a JSON file that specifies the conversation's flow. This setup enables the conversations to be easily modified without having to change any part of the code (CF#1).

4.2.2 Coordinated Turn-Taking. Our platform is highly configurable and supports a wide variety of learning-by-teaching scenarios. Students can teach the agent individually or in groups of arbitrary size, and their group placement can be configured by teachers or researchers through a command line interface in the first version, and through an admin interface in the second version (CF#4). If a student is placed in a group and their group members are present, their view of the system is synchronized—that is, if one student navigates to another interface (e.g., teaching vs. testing), all students will be automatically brought to the same screen. During the first deployment, between each after-school session, there were design modifications to the turn-taking mechanisms. Initially, our platform gave students complete freedom to choose what and when to teach the robot. This setup was too open-ended, and students had great difficulty narrowing down what content mattered and dividing the teaching task. Subsequently, the agent was redesigned to control turn taking—namely, it determines which group member is online and active, asks the student who has participated the least number of turns to teach next. When a student is stuck (e.g., picked a sentence unrelated to what the agent is asking about), the agent will also delegate the task to the next student, asking them to help. Overall, the turn taking mechanism enabled children, placed in different-sized groups within the same classroom, to work together on simultaneously teaching different robots.

4.2.3 Agent Embodiment Logic. The Curiosity Notebook supports a clean separation between agent logic and embodiment, thereby allowing the teachable agent to take on different types of embodiment (CF#5). This is accomplished by keeping the teachable agent's logic—e.g., how it learns, how it exhibits emotions, and what it says—inside the Curiosity Notebook web application, and having an external program (e.g., a python script) ping the database for chat messages that the physical robot should say out loud. Each chat message is associated with an emotion tag (e.g., curious), which can be used to control the movements/gestures of the robot (e.g., rubbing its head or chin) to convey that emotion. Similarly, the external program can push sensing events to the Curiosity Notebook. The NAO robot, for example, has tactile sensors on its head, hands and feet, which can serve as alternative ways for students to provide feedback to the robot (e.g., patting its head when it answers a quiz question correctly). If the Curiosity Notebook is used without any physical embodiment for the agent, the agent takes the form of a text-based conversational agent. This functionality enabled us to set up the agent to be a physical robot in the first deployment, and a chatbot (i.e., text-based conversational agent) in the second deployment. In other words, supporting flexible agent embodiments allows for studies that use different embodiments to be deployed either simultaneously or in quick succession without having to change the platform's core logic.

4.2.4 Interfaces for Monitoring the Agent's Knowledge. As shown earlier in Figure 3, in the first version of the Curiosity Notebook, the state of the agent's learning is represented by knowledge bubbles, each representing a feature that is relevant for the classification task at hand. For example, a feature relevant for distinguishing mammals and reptiles would be whether the animal lays eggs. A knowledge bubble becomes filled when an agent has mastered/learned the associated feature, otherwise it remains empty. In our first deployment, we found that although filled knowledge bubbles were reflective of the agent's knowledge, students could not gauge from the bubbles what content the agent learned and how different actions affected the agent's knowledge. The knowledge bubbles also had the unintentional effect of encouraging students to rush through teaching in order

to fill up all the knowledge bubbles as quickly as possible, and reducing their focus on learning the material.

To address these issues, in the second version, the knowledge bubbles were replaced with the agent’s notebook (Figure 5), which records everything that the agent has been taught. The agent’s notebook is informative in two ways; it tells students what the agent knows, and it gives students clues as to what each button does. For example, sharing a joke with the agent does not affect its knowledge, but telling a fun fact does. Throughout teaching, students can access the agent’s notebook using the *Notebook* button (next to the agent’s avatar in Figure 3(b)) at any time. In the initial testing, we found that users who were not familiar with the interface often forgot that the agent’s notebook existed. Thus, extra logic was built to make the *Notebook* button pulsate for approximately two seconds every time a new note has been taken by the agent. Additionally, the agent is programmed to tell the user through chat that it has made a new entry in its notebook.

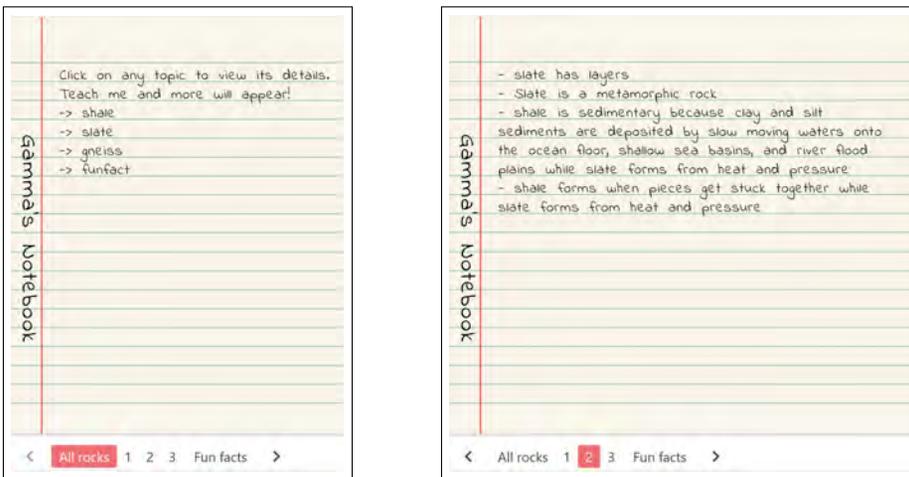


Fig. 5. Table of content page of the agent’s notebook (left) and the page containing notes for Slate (right), in the second version of the Curiosity Notebook

The notebook starts out empty initially, with the first page serving as a table of content. Upon learning a fact about a rock, a new page is created for that rock with the added note, and a new index to the page is added to the table of content. The last page of the notebook is reserved for showing all the fun facts. The cursive Indie Flower font¹ was chosen for the notebook texts to represent the agent’s handwriting. The information that the agent learns is stored in a relational database. For instance, for rock classification, the rock-to-category table stores what the agent has learned about rocks and their categories. The rock-to-feature table stores what the agent has learned about a given rock’s features. When a fact is taught, a *note* is generated (e.g., “Schist is a Metamorphic rock”, “Conglomerate has sand or pebbles”) and automatically linked to a rock-to-category mapping, or one or more rock-to-feature mappings. Additionally, if users provide an explanation in *Explain*, the explanation is added to the note as well (e.g., “Granite has large crystals because the cooling process is slow”). When two rocks from the same category are compared in *Compare*, the note associates these rocks to the same feature (e.g., “Schist has layers and Gneiss has layers”). If the rocks are from different categories, then the note associates each rock to a different feature (e.g., “Quartzite could be white because ... while Sandstone has layers”). Finally, fun facts (from *Fun Fact*) are stored

¹<https://fonts.google.com/specimen/Indie+Flower>

along with the reasons that the users provided the agent as to why they found it interesting. An example is “There is Gneiss in Canada that date back 4 billion years ago! (Reason: It is fascinating to know that rocks more than 4 billion years old can be found in this country)”.

Table 3. *Correct* conversation flow in the second version of the Curiosity Notebook

Gamma:	I got something wrong?
Gamma:	I know 'gneiss', 'shale', 'slate'
Gamma:	Ok, tell me which rock this is about?
User:	Gneiss
Gamma:	Oh, gneiss
Gamma:	Select which notebook entry you want to correct!
User:	I think that 'Gneiss is an igneous rock' is wrong.
Gamma:	So, what kind of a rock do you think gneiss is then?
User:	Metamorphic
Gamma:	Thanks for clearing this up!
Gamma:	Thanks for that information. Now you can select another button to keep teaching me!

Students can click on the *Correct* button to update a fact that was previously taught to the agent. A sample conversation is shown in Table 3—the agent will first ask what object the student wants to focus on, then present the student with a list of learned facts about that object (entries in the agent’s notebook) to choose from, and finally, use questions to elicit a specific kind of correction.

In the first version’s design, teaching is considered done when all the knowledge bubbles were filled. With the notebook, on the other hand, there is no simplistic idea of knowledge mastery. The users are told to teach the agent for 40-minutes in the way that they like, and are free to interpret if and how much the notebook is indicative of their competence as a teacher. Having the notebook also gives users more transparency into how the agent learns, what it has been taught, and what it has yet to learn. As a more explicit representation of the agent’s knowledge (compared to the knowledge bubbles in the first version), the platform better supports metacognitive skills like identifying gaps and mistakes in the agent (and the student tutor’s own) understanding. Finally, the notebook feature presents a future opportunity to simulate the behaviour of different types of agent learners, e.g., an attentive agent learner might take notes of everything, whereas an inattentive learner might only occasionally take notes.

4.2.5 Agent’s Learning Mechanism. In the first version, the agent is simplistic—it does not understand students’ responses to questions, and always pretends to learn what students have taught. Initially, the agent is unable to answer any test questions correctly; its ability to answer test questions increases with the number of features it has learned (i.e., number of completed teaching conversations). The second version improves upon this: the agent answers test questions according to the set of features it has been taught about the tested entity. As such, the agent is not able to categorize entities that it has not been taught before (and will inform the student tutor as such). It will also categorize an entity incorrectly if it has been taught the wrong information. As discussed previously, the second version’s agent knowledge is represented using the agent’s notebook in the frontend, and using the entity-to-category and entity-to-feature mappings in the backend. This allows the agent to answer test questions according to its knowledge, and allows the student tutors to accurately gauge the agent’s learning.

4.2.6 Admin Interfaces. As we envision the Curiosity Notebook may be used by researchers to conduct studies and, eventually, by teachers to organize learning by teaching activities for their

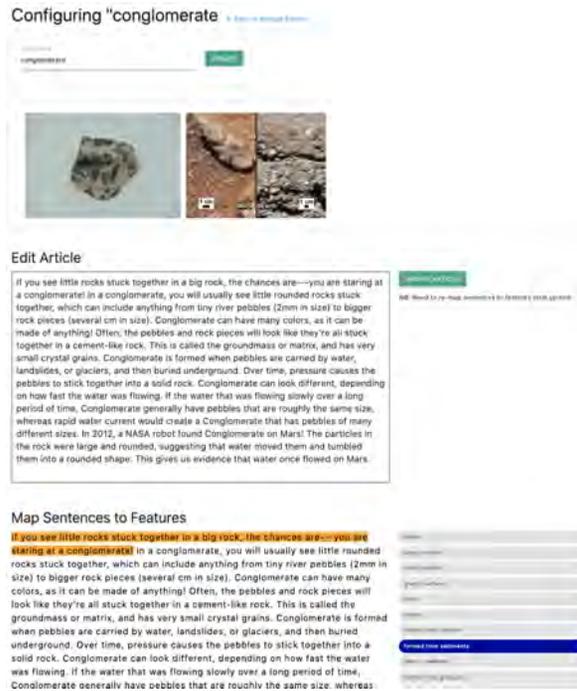


Fig. 6. Admin interface for mapping sentences to relational facts, in the second version of the Curiosity Notebook. Right image courtesy NASA/JPL-Caltech/MSSS and PSI [61]

class, the platform provides a set of web-based administrative tools for adding/removing users, updating user information, assigning users to groups, as well as configuring classification tasks and materials (e.g., articles, images). There is, for example, a semi-automatic way of mapping sentences to relational facts (e.g., Gabbro is an igneous rock, Gabbro has holes), as shown in Figure 6. For features, the system automatically maps sentences to features by scanning for a set of keywords or phrases (e.g., holes, bubble, bubbles, porous, cavities) that are synonymous to a feature (e.g., “has holes”), and these mappings are subsequently manually verified, and corrected if needed, by the human user organizing the learning-by-teaching session. These admin features allow researchers and teachers to easily configure the learning material (CF#3) and tutor groups (CF#4), thus allowing teachers and experimenters to quickly adapt to unexpected changes during learning-by-teaching sessions, e.g., having to modify group assignments due to team dynamics problems or absent group members, or having to adapt materials on the fly.

As a research platform, the Curiosity Notebook provides researchers with the ability to configure experiments. Functionalities include the ability to add/remove experiments, add/remove conditions to/from experiments, and assign users to a specific experiment and condition. The platform also provides functionalities for researchers to configure the verbal behaviour of the agent, and associate different verbal behaviour with different experimental conditions (CF#1), and deploy these agents, each with a distinctly different verbal behaviour, simultaneously to different participants. Following, we describe findings from two deployments in different settings that were made possible due to Curiosity Notebook’s support of the five CFs.

5 UTILITY OF CURIOSITY NOTEBOOK AS A RESEARCH PLATFORM FOR LEARNING BY TEACHING

In this section, we describe findings from the two deployments (detailed in Section 4.1), as a way to demonstrate how the configurability of the Curiosity Notebook makes it a useful research platform for learning by teaching. Our two deployments, which took place in contrasting learning environments—in one deployment, elementary school students taught a physical NAO robot in groups; in the second deployment, university students taught a text-based conversational agent individually—provide evidence for the versatility and utility of our platform. The ability of our platform to adapt learning materials (CF#3), group sizes (CF#4) and agent embodiment (CF#5) enabled us to quickly configure the system to deploy to vastly different learning settings. In this section, we describe, in some details, the utility of certain configurable features and what they enabled us to observe during the two deployments. For observations from the first deployment, we describe the results in aggregate as much of the observations were gathered informally from children. Due to the informal nature of the after-school club during which the sessions were run, children worked mostly with the same group of children, but were sometimes reassigned to a different group due to absences and personal preferences. For the second deployment which involved individual sessions, participants are denoted here as p_1, \dots, p_n , respectively.

The Impact of Agent Characteristics on Perception and Learning-by-Teaching Behaviour.

Our findings suggest that different students perceived the same agent characteristics (e.g., programmed personality) in both positive or negative ways. In the first deployment, some elementary school students attributed the behaviour of the agent as indicative of it being a good learner; students said that the agent is a good learner because of its attentiveness (e.g., “because it pays very close attention”, “because he/she ...sits in one spot and doesn’t get distracted”), curiosity (e.g., “because it’s curious”, “because Delta asks questions, just like a human student”), and its ability and eagerness to learn (e.g., “because he got everything right”, “because he’s always ready to learn”). However, other students had more negative perceptions, mentioning the rationale of “talking too much” as the reason for the agent not being a good student. Likewise, in the second exploratory study, most participants perceived Gamma as “eager” (p9, p11, p13, p15, p20, p22, p23, p31), “enthusiastic” (p13), “positive” (p15, p26, p40), “pleasant” (p16, p18), “cheerful” (p17, p24, p25), and “excited” (p16, p17, p36), and a good student because it takes notes. On the other hand, some participants found Gamma to be disingenuous, e.g., “over enthusiastic” (p33), “a bit over the top” (p35), “fake” (p3, p5), “artificial” (p6) and “repetitive” (p3). One participant (p21) suggested that a more natural version of Gamma would be one where “she would stop paying attention if you don’t use an entertainment button, and lose focus and maybe jots down the wrong answer.”

How students perceived the agent also critically affected how they perceived themselves as teachers. One interesting observation from the first deployment was that what the agent said, as well as the degree to which the agent was reported to be a “good” student, seemed to be associated with students’ perceptions of their own competence as teachers. That is, the majority of the students in the study saw themselves as good teachers and attributed their success at teaching to not only the learning progress of the robot (e.g., “because my robot has learned a lot”, “because we got all the bubbles for animals”), but also to the positive feedback they received from the robot (e.g., “because the robot told me so”, “because Delta always says good choice”). In contrast, university students in the second deployment, who had more access to information about what the robot has actually learned, seemed to rely more on accuracy information to judge their own competence as teachers. Participants who thought they taught well gave reasons including: 1) the notes written in the agent’s notebook were all correct (p0, p6), 2) the agent answered all quiz questions correctly

(p1, p7, p19, p20, p22, p24, p38, p40), and 3) they taught the agent all 12 rocks (p11, p13-16). None of the participants, except for one (p20), mentioned the agent's conversational feedback helped them see themselves as a good teacher, even though both the first and second deployment agents were designed to have the same level of enthusiasm. Given this observation, it may be beneficial for future versions to modify the verbal repertoire of the teachable agent to include specific feedback to the students about their teaching.

Configurability of agent characteristics (CF#1) therefore can allow us to observe how the agent, as a tutee, can affect student's self-perception and subsequently their learning and teaching behaviour.

Coordination of Group-Based Teaching. The configurable feature CF#4 allows us to create student tutor groups of varied sizes. During the first deployment, students taught the agent in groups of 3. This (i.e., having many students tutor the same agent tutee) is a somewhat atypical scenario in human-to-human peer teaching, but has been studied in human-to-machine teaching [10, 41]. We observed that some students took the initiative to offer help to their teammates when it was not their turn to teach, while others were impatient at having to wait. Interestingly, the amount of attention that the robot gives to each student tutor seems to also affect students' perceptions of their own teaching ability; one student said "Student X teaches way better because the robot chooses X more." Together, these observations suggest the benefit of having a more idiosyncratic approach to managing group-based teaching that takes into account each student's ability to work in a team and their unique need for attention from the agent.

Quantification of Teaching Strategies. Most importantly, our platform enabled us to observe, both qualitatively and quantitatively, *how* students go about teaching the agent (CF#2).

In the first deployment, we observed different groups of students demonstrating different teaching strategies, for example, some groups filled as many knowledge bubbles as possible before they tested the robot's knowledge; whereas other groups tested the robot often, e.g., after each filled bubble. Since the agent completely controls the conversation through a predefined sequence of questions, several students found this restriction as limiting their teaching process and wanted to be able to control how they teach more proactively. With the introduction of teaching buttons, the second deployment provides us with a much bigger opportunity to understand, quantitatively, the teaching behaviour of the students.

For the second deployment, we performed a cluster analysis on the data from 40 participants (p10 was removed due to completing too few teaching conversations) to investigate whether participants' behaviour based on the amount of time they spent on different teaching activities reflected different teaching styles. There are 8 features in total—one for each of the 7 buttons, plus the button used to view the agent's notebook. From these 8 features, 3 were removed upon further inspection; specifically, three buttons were used by participants less than three times on average throughout the entire teaching session: *Compare* ($M = 2.83$, $SD = 1.69$), *Correct* ($M = 1.1$, $SD = 1.19$) and *Tell Joke* ($M = 1.03$, $SD = 1.00$). In contrast, the 5 other buttons are used more frequently: *Describe* ($M = 14.38$, $SD = 9.40$), *Explain* ($M = 7.08$, $SD = 4.49$), *Quiz* ($M = 6.03$, $SD = 3.07$) and *Fun Fact* ($M = 4.55$, $SD = 2.99$). For each of these 5 teaching activities invoked by clicking button X, we create a feature to represent the relative attention that participants paid to that teaching activity, i.e., the number of times a participant clicked X divided by the total number of button clicked over the entire session. Having 5 features is an appropriate amount according to Formann, who suggested no more than k features for a sample size of 2^k when performing cluster analysis (i.e., $2^5 < 40$) [31]. The correlation between these 5 features were checked to ensure all features were distinctive during clustering [38]. Hierarchical clustering was used as the clustering method; it

was chosen given its use in past studies on tutors' behaviours on learning-by-teaching platforms [46]. The average silhouette width method [43, 74] was used to analyze the quality of clustering. Its value ranges $[-1, 1]$ and measures how close each point in their cluster is to other clusters, with a higher value indicating better quality clusters. Linear models were used to investigate the effects of other factors (demographics, perception of the agent, etc.) on button click rates. Stepwise selection was performed to select the variables that best explain the variability of each rate. The fit of the models were verified through visual inspection of the QQ-plots [88].

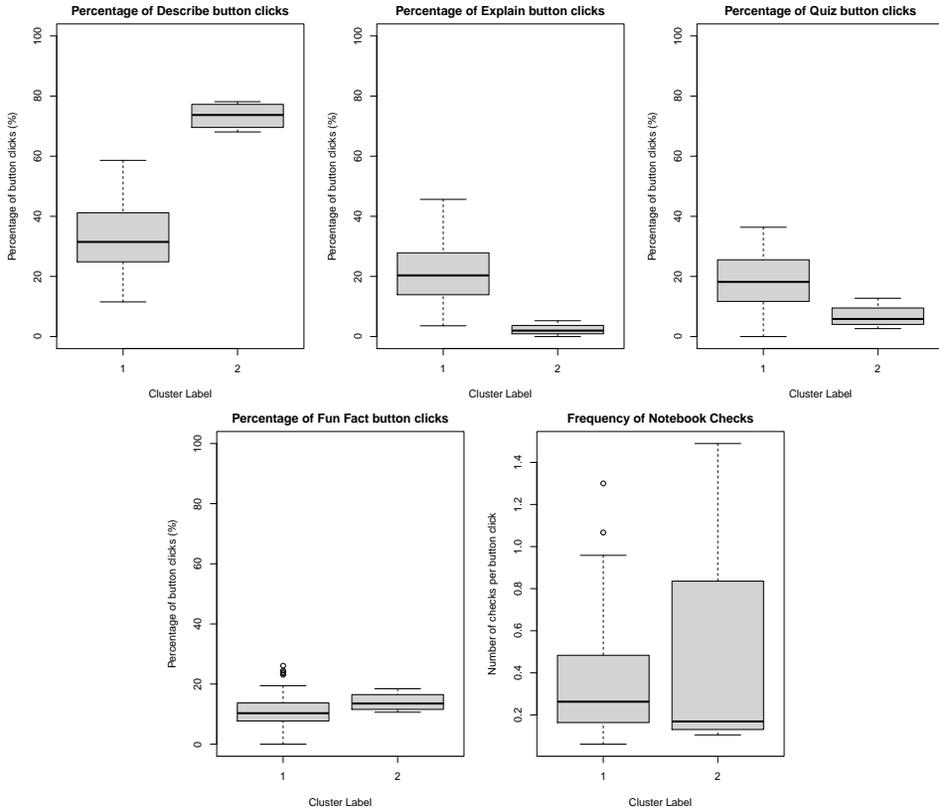


Fig. 7. Comparison of clusters on features used for clustering (deployment 2).

Two clusters were found ($N_1 = 36$, $N_2 = 4$), referred to as C1 and C2. These clusters achieve an average silhouette index of 0.36 with all participants having positive index values, indicating that all participants fitted well within their assigned cluster. As shown in Figure 7, participants in C1 have higher percentages of using *Explain* ($Mdn_1 = 20\%$, $Mdn_2 = 2\%$) and *Quiz* ($Mdn_1 = 18\%$, $Mdn_2 = 6\%$). They also check the agent's notebook more frequently ($Mdn_1 = 26\%$, $Mdn_2 = 17\%$). Participants in C2, on the other hand, have a much higher percentage use of *Describe* ($Mdn_1 = 31\%$, $Mdn_2 = 74\%$), and a slightly higher percentage use of *Fun Fact* ($Mdn_1 = 10\%$, $Mdn_2 = 14\%$).

Taking these results together, C1 participants can be interpreted as teachers who are more engaged than C2 participants. On average, 87% of button clicks by C2 are *Describe* and *Fun Fact* (which requires regurgitation, instead of synthesis, of facts), compared to only 42% by C1. C2's teaching strategy is best described by p31, "I mainly described things (rocks) to Gamma rather than explain". On the flipside, C1 participants focus more on other aspects of teaching, like providing explanations

via *Explain*, probing agent performance and gaining crucial feedback via *Quiz* and tracking the agent's learning by checking the notebook. Fitting linear models on features beyond the final set of clustering features also revealed other significant differences. For instance, C2 participants ($Mdn_2 = 68$) have, on average, significantly more notes than C1 participants ($Mdn_1 = 46$) with a large effect size $r = .45$ ($\beta = 36.28, t(37) = 5.86, p < .01$), as the *Describe* conversation is less time-consuming. This clustering is also in line with Felder and Silverman [25], which discussed two different teaching styles—Concrete, which involves teaching in a repetitive manner, as exemplified by C2 participants; and Global, which is associated with a more diverse and creative teaching repertoire, as exemplified by C1 participants [25].

We were also interested in the characteristics of the two clusters, and thus examined participants' individual responses to the questionnaires. The AMS *IM-to know* scores show that C1 participants are more intrinsically motivated to know ($Mdn_1 = 5$) than C2 participants ($Mdn_2 = 3.5$). A Mann-Whitney test shows significant difference, $U(N_1 = 36, N_2 = 4) = 131, p = .01$, and the r -value shows an effect size of .42, which is a large effect as suggested in [32]. Also, the AMS *EM-introjected regulation* scores show that C1 participants are more likely to internalize their teaching behaviour ($Mdn_1 = 4.33$) than C2 participants ($Mdn_2 = 1.67$); significance is found by a Mann-Whitney test, $U(N_1 = 36, N_2 = 4) = 124.5, p = .02$, alongside a large effect ($r = .38$). Cohen's d also showed large effect sizes for both *IM-to know* and *EM-introjected regulation* score differences (1.5 and 1.6 respectively) [17].

These aspects have all been found to be beneficial not only to teaching, but also learning. C1 scores higher on the *IM-to know* scale, which relates to the satisfaction of learning, exploring and understanding something new. Through fitting Gaussian linear models, we also found that participants who had higher *IM-to know* scores scored better in the post-study questionnaire for questions about rocks which do not have articles in the Curiosity Notebook $\beta = .42, z = .20, p = .03$. In other words, these participants scored significantly better on rocks that were not previously seen when teaching the agent; they were better at transferring classification rules learned during the teaching session to new contexts (i.e., new rocks). This suggests that participants with higher *IM-to know* could end up learning better transfer skills than others due to their tendency to engage in teaching activities involving the integration of facts (as observed in C1).

In short, the two deployments demonstrate the Curiosity Notebook's configurable features (detailed in Section 3), and, crucially, the ability to generate important insights due to that flexibility. For instance, supporting configurable agent characteristics (CF#1) provided information about the effects of the tutee agent's characteristics on students' behaviour and perceptions of themselves and the agent; supporting group-based teaching (CF#4) confirms the need for more studies (that can be carried out using Curiosity Notebook) into understanding group dynamics during learning by teaching; providing a quantification of teaching behaviours (CF#2) allowed for insights into less vs. more engaged participants' teaching behaviour and learning outcomes. Moreover, supporting configurable learning material (CF#3) allowed for each deployment to be carried out for different age groups. Lastly, supporting flexible agent embodiment (CF#5) allowed rapid pivoting from in-person studies, e.g., NAO robots in Deployment 1, to online studies, e.g., text-only agent in Deployment 2, vice versa.

6 ENVISIONING VERSION 3: A DISCUSSION

In this section, we summarize Curiosity Notebook's design and changes from the first to second version in terms of the configurable features (CF). We also discuss potential improvements for the next version.

The configurability of the platform substantially improved from the first to second version. In the first version, the agent was programmed to be an enthusiastic and quick learner, and its

implementation did not allow researchers to easily configure agent characteristics (CF#1). That is, they could not be swapped on-the-fly, nor could agents with different characteristics be deployed simultaneously (e.g., for between-subjects studies). Second, students were not given any autonomy over their teaching behaviour, such as the article taught and the way it is taught (CF#2). Third, there was no easy way for researchers to modify articles or verify/correct the mapping between content (e.g., sentences) to concepts (e.g., categories, features) (CF#3). Fourth, although the platform allowed for easy and quick allocation of students into groups of any size, more personalized methods for encouraging effective collaboration within groups should be considered (CF#4). Lastly, following CF#5, the platform enabled flexible agent embodiments. Compared to version 1, version 2 allowed for easy configuration of agent characteristics through the use of JSON configuration files (CF#1). Second, it provided students with 7 buttons that initiated 7 distinct conversations (CF#2), and the second deployment demonstrated the amount of flexibility allowed in teaching behaviour through the discovery of two distinct behaviours that affected learning. Third, new administrative interfaces allowed researchers to easily associate different agent characteristics with different experimental conditions (CF#1), configure the material (CF#3) and tutor grouping (CF#4).

These improvements allowed us to conduct a number of studies (described elsewhere in [12, 68]) where the agent's characteristics were the experimentally manipulated through quick configurations. It allowed us to release the second deployment easily, even though a complete overhaul of the material was required. Moreover, it allowed us to collect detailed quantitative data on how students go about teaching. For CF#2, a potential improvement is to design the platform to collect more detailed data on tutor-agent interaction, so that complex interactions, such as those mentioned by Roscoe and Chi, can be appropriately captured and analyzed [73]. In terms of CF#4, we envision the next version of the Curiosity Notebook to include configurability on how teams of students communicate with each other, and with the agent. For instance, a separate JSON file could be used to list preferred details of various aspects of group communication, such as how frequently should the agent explicitly encourage discussions between group members, how frequently should the agent refer to material taught by a group member while being taught by another, and what (if any) ways can group members interfere when a member is teaching. Further improvements can also allow for the integration of machine learning models that impact either agent characteristics (CF#1) or group communication (CF#4). In terms of CF#5, the next version could provide more standardized application programming interfaces (API) to enable any output devices to act as an agent. This would enhance flexibility by further reducing the ease of connecting various software (e.g., text-to-speech software) or hardware (e.g., robots) to the platform.

One of the less satisfying aspects about the learning-by-teaching platform so far is our ability to demonstrate learning gains in the students themselves. A potential reason is that the learning objectives of the agent is too inconspicuous—the *Quiz* action, which demonstrates how well the agent can classify objects, needs to be invoked explicitly by the student tutor. Prior research has showed that participants learn better when observing their students use the knowledge they taught [64, 65]. In our studies, we asked participants to teach the agent, but we did not tell them *why* the agent needs to acquire that knowledge. The student's sense that the knowledge taught to the agent matters could become a source of extrinsic motivation for the user, further enhancing their learning. Thus, having the teachable agent demonstrate its skills openly to students while it is being taught can be both informational (i.e., shows where the misunderstandings lie) and motivational. In version 3 of the Curiosity Notebook, we envision adding a "demonstration panel", where the agent is performing a task live (e.g., sorting objects into different categories, writing code, describing the steps to put together a recipe) while soliciting feedback and help from the students. Having the agent show its mistakes in an obvious fashion can also prompt for specific forms of teaching,

without needing explicit teaching buttons; for example, students may be inclined to teach certain rules to the agent in order to fix its mistakes or to help it reach its goal.

There are a few limitations in our design approach. The first deployment was over 4 weeks, thus allowing us to observe learning-by-teaching behaviours over time. However, the sample size was considerably small, which limited the potential for more complex analysis on participants' teaching behaviours. Additionally, since the design of the two deployments were not consistent, no direct comparison could be made between the studies. Nonetheless, the two deployments and multiple design iterations allowed us to understand the intricacies of the learning-by-teaching process and what is required for a research platform for learning by teaching to be maximally useful.

7 CONCLUSIONS

In this paper, we introduce the Curiosity Notebook, an interface that provides various teaching tools for participants to interact with a teachable agent. We iteratively designed the platform based on observations from two deployments—a field study with 12 elementary school children and an online study with 41 university students. We showed that by providing ways to configure features that were identified by prior research to be important, two learning-by-teaching studies deployed under very different contexts could be conducted on the same platform without much overhead. The Curiosity Notebook's value as a research utility was demonstrated by the important insights gained from the studies.

Moving forward, there are plans to allow the research community to fully exploit Curiosity Notebook's utility as a research platform through open sourcing. This greatly reduces the overhead of conducting such research by removing the need to build platforms from the ground up. Moreover, having a platform that many can use for their research will allow for more direct comparisons in findings between various learning-by-teaching studies. This will make it much easier for findings to be generalized or aggregated to a level suitable to be applied directly in the real world by those in the education sector.

8 ACKNOWLEDGMENTS

We thank all participants for their contributions, and acknowledge the funding from the NSERC Discovery Grant RGPIN-2015-0454 and the University of Waterloo Interdisciplinary Trailblazer Fund for making this work possible.

REFERENCES

- [1] Mehdi Alaimi, Edith Law, Kevin D. Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical Agents for Fostering Question-Asking Skills in Children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3313831.3376776>
- [2] Linda F Annis. 1983. The processes and effects of peer tutoring. *Human Learning: Journal of Practical Research & Applications* 2, 1 (1983), 39–47.
- [3] Safiye Aslan. 2017. Learning by Teaching: Can It Be Utilized to Develop Inquiry Skills?. *Journal of Education and Training Studies* 5, 12 (2017), 190–198.
- [4] Robert K Atkinson. 2002. Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology* 94, 2 (2002), 416.
- [5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2008. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (Nov. 2008), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- [6] Satabdi Basu, Gautam Biswas, and John S. Kinnebrew. 2016. Using Multiple Representations to Simultaneously Learn Computational Thinking and Middle School Science. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (Phoenix, Arizona) (AAAI'16)*. AAAI Press, California, US, 3705–3711.
- [7] Amy L Baylor and Yanghee Kim. 2005. Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education* 15, 2 (2005), 95–115.

- [8] Carl A Benware and Edward L Deci. 1984. Quality of Learning With an Active Versus Passive Motivational Set. *American Educational Research Journal* 21, 4 (Jan. 1984), 755–765. <https://doi.org/10.3102/00028312021004755>
- [9] 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 (2005), 363–392. <https://doi.org/10.1080/08839510590910200>
- [10] Cynthia Breazeal, Nick Depalma, Jeff Orkin, Sonia Chernova, and Malte Jung. 2013. Crowdsourcing Human-Robot Interaction: New Methods and System Evaluation in a Public Environment. *Journal of Human-Robot Interaction* 2 (03 2013), 82–111. <https://doi.org/10.5898/JHRI.2.1.Breazeal>
- [11] Jessy Ceha, Nalin Chhibber, Joslin Goh, Corina McDonald, Pierre-Yves Oudeyer, Dana Kulić, and Edith Law. 2019. Expression of Curiosity in Social Robots: Design, Perception, and Effects on Behaviour. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). ACM, New York, NY, USA, Article 406, 12 pages. <https://doi.org/10.1145/3290605.3300636>
- [12] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. *Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?* Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3411764.3445068>
- [13] Shruti Chandra, Raul Paradedo, Hang Yin, Pierre Dillenbourg, Rui Prada, and Ana Paiva. 2017. Affect of Robot's Competencies on Children's Perception. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil) (AAMAS '17). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1490–1492.
- [14] Catherine C. Chase, Doris B. Chin, Marily A. Oppezzo, and Daniel L. Schwartz. 2009. Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *Journal of Science Education and Technology* 18, 4 (01 Aug 2009), 334–352. <https://doi.org/10.1007/s10956-009-9180-4>
- [15] Doris B Chin, Ilsa M Dohmen, Britte H Cheng, Marily A Oppezzo, Catherine C Chase, and Daniel L Schwartz. 2010. Preparing students for future learning with teachable agents. *Educational Technology Research and Development* 58, 6 (2010), 649–669.
- [16] Doris B Chin, Ilsa M Dohmen, and Daniel L Schwartz. 2013. Young children can learn scientific reasoning with teachable agents. *IEEE Transactions on Learning Technologies* 6, 3 (2013), 248–257.
- [17] Jacob Cohen. 1992. A power primer. *Psychological bulletin* 112, 1 (1992), 155.
- [18] Peter A Cohen, James A Kulik, and Chen-Lin C Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *American educational research journal* 19, 2 (1982), 237–248.
- [19] Elaine B. Coleman, Ann L. Brown, and Inna D. Rivkin. 1997. The Effect of Instructional Explanations on Learning From Scientific Texts. *Journal of the Learning Sciences* 6, 4 (Oct. 1997), 347–365. https://doi.org/10.1207/s15327809jls0604_1
- [20] Pieter M.A. Desmet, Martijn H. Vastenburger, and Natalia Romero. 2016. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *J. of Design Research* 14, 3 (2016), 241. <https://doi.org/10.1504/jdr.2016.079751>
- [21] Pierre Dillenbourg. 1999. What do you mean by collaborative learning?
- [22] M. Suzanne Donovan and John D. Bransford. 2005. *How Students Learn: History, Mathematics, and Science in the Classroom*. The National Academies Press, Washington, DC. <https://doi.org/10.17226/10126>
- [23] David Duran. 2016. Learning-by-teaching. Evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International* 54, 5 (Feb. 2016), 476–484. <https://doi.org/10.1080/14703297.2016.1156011>
- [24] Mona Emará, Ramkumar Rajendran, Gautam Biswas, Mahmod Okasha, and Adel Alsaied Elbanna. 2018. Do Students' Learning Behaviors Differ when They Collaborate in Open-Ended Learning Environments? *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 49 (Nov. 2018), 19 pages.
- [25] Richard M Felder, Linda K Silverman, et al. 1988. Learning and teaching styles in engineering education. *Engineering education* 78, 7 (1988), 674–681.
- [26] Logan Fiorella and Richard E. Mayer. 2013. The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology* 38, 4 (Oct. 2013), 281–288. <https://doi.org/10.1016/j.cedpsych.2013.06.001>
- [27] Logan Fiorella and Richard E. Mayer. 2014. Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology* 39, 2 (April 2014), 75–85. <https://doi.org/10.1016/j.cedpsych.2014.01.001>
- [28] Logan Fiorella and Richard E. Mayer. 2015. Eight Ways to Promote Generative Learning. *Educational Psychology Review* 28, 4 (Nov. 2015), 717–741. <https://doi.org/10.1007/s10648-015-9348-9>
- [29] John H Flavell and Henry M Wellman. 1975. Metamemory. , 66 pages.
- [30] Marta Flores and David Duran. 2013. Effects of Peer Tutoring on Reading Self-Concept. *IJEP – International Journal of Educational Psychology* 2, 3 (2013), 297–324. <https://doi.org/10.4471/ijep.2013.29>
- [31] Anton K Formann. 1984. *Die latent-class-analyse: Einführung in Theorie und Anwendung*. Beltz, Weinheim, Germany.
- [32] Catherine O Fritz, Peter E Morris, and Jennifer J Richler. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* 141, 1 (2012), 2.

- [33] Luke Fryer and Rollo Carpenter. 2006. Bots as language learning tools. *Language Learning & Technology* 10, 3 (2006), 8–14.
- [34] James J Gallagher and Mary Jane Aschner. 1963. A preliminary report on analyses of classroom interaction. *Merrill-Palmer Quarterly of Behavior and Development* 9, 3 (1963), 183–194.
- [35] Kashinath Garkal, Anjali Shete, et al. 2018. Learning by teaching: Role of??? peer-assisted learning??? in medical education. *Journal of Contemporary Medical Education* 9, 1 (2018), 17–20.
- [36] David Griol and Zoraida Callejas. 2013. An Architecture to Develop Multimodal Educative Applications with Chatbots. *International Journal of Advanced Robotic Systems* 10, 3 (Jan. 2013), 175. <https://doi.org/10.5772/55791>
- [37] Joachim Grzega and Marion Schöner. 2008. The didactic model LdL (Lernen durch Lehren) as a way of preparing students for communication in a knowledge society. *Journal of Education for Teaching* 34, 3 (2008), 167–175.
- [38] Joseph F Hair, William C Black, Barry J Babin, and Rolph E Anderson. 2009. *Multivariate Data Analysis*. Pearson Education Limited, Harlow, Essex.
- [39] Neil T Heffernan and Ethan A Croteau. 2004. Web-based evaluations showing differential learning for tutorial strategies employed by the Ms. Lindquist tutor. In *International Conference on Intelligent Tutoring Systems*. Springer, Berlin, Heidelberg, 491–500.
- [40] Kate Hone, Farah Akhtar, and Martin Saffu. 2003. Affective agents to reduce user frustration: the role of agent embodiment. In *Proceedings of Human-Computer Interaction (HCI2003)* (Bath, UK). Association for Computing Machinery, New York, NY, USA, 4.
- [41] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. 2015. When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) (HRI '15). Association for Computing Machinery, New York, NY, USA, 83–90. <https://doi.org/10.1145/2696454.2696479>
- [42] Vincent Hoogerheide, Sofie M.M. Loyens, and Tamara van Gog. 2014. Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction* 33 (Oct. 2014), 108–119. <https://doi.org/10.1016/j.learninstruc.2014.04.005>
- [43] Phan Duy Hung, Nguyen Thi Thuy Lien, and Nguyen Duc Ngoc. 2019. Customer Segmentation Using Hierarchical Agglomerative Clustering. In *Proceedings of the 2019 2nd International Conference on Information Science and Systems* (Tokyo, Japan) (ICISS 2019). Association for Computing Machinery, New York, NY, USA, 33–37. <https://doi.org/10.1145/3322645.3322677>
- [44] Terrence J Keeney, Samuel R Cannizzo, and John H Flavell. 1967. Spontaneous and induced verbal rehearsal in a recall task. *Child Development* 38 (1967), 953–966.
- [45] John W Kinch. 1963. A formalized theory of the self-concept. *Amer. J. Sociology* 68, 4 (1963), 481–486.
- [46] John S Kinnebrew, James R Segedy, and Gautam Biswas. 2014. Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition and learning* 9, 2 (2014), 187–215.
- [47] Deanna Kuhn. 2000. Metacognitive development. *Current directions in psychological science* 9, 5 (2000), 178–181.
- [48] Deanna Kuhn and Maria Pease. 2006. Do children and adults learn differently? *Journal of cognition and development* 7, 3 (2006), 279–293.
- [49] Marjan Laal and Seyed Mohammad Ghodsi. 2012. Benefits of collaborative learning. *Procedia-social and behavioral sciences* 31 (2012), 486–490.
- [50] Andreas Lachner, Iris Backfisch, Vincent Hoogerheide, Tamara van Gog, and Alexander Renkl. 2020. Timing matters! Explaining between study phases enhances students' learning. *Journal of Educational Psychology* 112, 4 (May 2020), 841–853. <https://doi.org/10.1037/edu0000396>
- [51] Edith Law, Parastoo Baghaei Ravari, Nalin Chhibber, Dana Kulic, Stephanie Lin, Kevin D. Pantasdo, Jessy Ceha, Sangho Suh, and Nicole Dillen. 2020. Curiosity Notebook: A Platform for Learning by Teaching Conversational Agents. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems (Late Breaking Work) (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–8.
- [52] Krittaya Leelawong and Gautam Biswas. 2008. Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education* 18, 3 (2008), 181–208.
- [53] Krittaya Leelawong, Joan Davis, Nancy Vye, Gautam Biswas, Daniel Schwartz, Kadira Belynnne, Thomas Katzlberger, and John Bransford. 2002. The effects of feedback in supporting learning by teaching in a teachable agent environment. In *The Fifth International Conference of the Learning Sciences*. International Society of the Learning Sciences, Boulder, CO, USA, 245–252.
- [54] Lisa Legault. 2017. Self-Determination Theory. In *Encyclopedia of Personality and Individual Differences*. Springer International Publishing, Switzerland, 1–9. https://doi.org/10.1007/978-3-319-28099-8_1162-1
- [55] Ross B MacDonald. 1991. An analysis of verbal interaction in college tutorials. *Journal of Developmental Education* 15, 1 (1991), 2.

- [56] Noboru Matsuda, Wenting Weng, and Natalie Wall. 2020. The Effect of Metacognitive Scaffolding for Learning by Teaching a Teachable Agent. *International Journal of Artificial Intelligence in Education* 30 (2020), 1–37.
- [57] Noboru Matsuda, Evelyn Yarzebinski, Victoria Keiser, Rohan Raizada, William W. Cohen, Gabriel J. Stylianides, and Kenneth R. Koedinger. 2013. Cognitive Anatomy of Tutor Learning: Lessons Learned with SimStudent. *Journal of Educational Psychology* 105, 4 (2013), 1152–1163. <https://doi.org/10.1037/a0031955>
- [58] Judith E Miller et al. 1994. Group Dynamics: Understanding Group Success and Failure in Collaborative Learning. *New directions for teaching and learning* 59 (1994), 33–44.
- [59] Roxana Moreno, Richard E Mayer, Hiller A Spires, and James C Lester. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and instruction* 19, 2 (2001), 177–213.
- [60] Anabil Munshi, Ramkumar Rajendran, Allison Moore, Gautam Biswas, and Jaclyn Ocupaugh. 2018. Studying the interactions between components of self regulated learning in open ended learning environments. In *Proceedings of International Conference of the Learning Sciences, ICLS*. International Society of the Learning Sciences, IN, USA, 1691–1692.
- [61] NASA. 2012. Rock Outcrops on Mars and Earth. https://www.nasa.gov/mission_pages/msl/multimedia/pia16189.html
- [62] David Nichols. 1994. Issues in designing learning by teaching systems. In *Proceedings of the East-West International Conference on Computer Technologies in Education*, Vol. 1. International Centre for Scientific and Technical Information, Moscow, Russia, 6.
- [63] Amy Ogan, Samantha Finkelstein, Elijah Mayfield, Claudia D’Adamo, Noboru Matsuda, and Justine Cassell. 2012. "Oh dear stacy!" social interaction, elaboration, and learning with teachable agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, USA, 39–48.
- [64] Sandra Y. Okita and Daniel L. Schwartz. 2013. Learning by Teaching Human Pupils and Teachable Agents: The Importance of Recursive Feedback. *Journal of the Learning Sciences* 22, 3 (July 2013), 375–412. <https://doi.org/10.1080/10508406.2013.807263>
- [65] Sandra Y. Okita, Selen Turkay, Mihwa Kim, and Yumiko Murai. 2013. Learning by teaching with virtual peers and the effects of technological design choices on learning. *Computers & Education* 63 (April 2013), 176–196. <https://doi.org/10.1016/j.compedu.2012.12.005>
- [66] SeungWon Park and ChanMin Kim. 2015. Boosting learning-by-teaching in virtual tutoring. *Computers & Education* 82 (2015), 129–140.
- [67] John P Powell and Lee W Andresen. 1985. Humour and teaching in higher education. *Studies in Higher Education* 10, 1 (1985), 79–90.
- [68] Parastoo Baghaei Ravari, Ken Jen Lee, Edith Law, and Dana Kulic. 2021. Effects of an Adaptive Robot Encouraging Teamwork on Students’ Learning. In *2021 30th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. ACM, New York, NY, USA, 8.
- [69] Bart Rienties, Dirk Tempelaar, Piet Van den Bossche, Wim Gijsselaers, and Mien Segers. 2009. The role of academic motivation in Computer-Supported Collaborative Learning. *Computers in Human Behavior* 25, 6 (2009), 1195–1206.
- [70] Debbie R Robinson, Janet Ward Schofield, and Katrina L Steers-Wentzell. 2005. Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review* 17, 4 (2005), 327–362.
- [71] Rod D Roscoe. 2014. Self-monitoring and knowledge-building in learning by teaching. *Instructional Science* 42, 3 (2014), 327–351.
- [72] Rod D Roscoe and Michelene TH Chi. 2004. The influence of the tutee in learning by peer tutoring. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 26. Cognitive Science Society, Austin, TX, 1179–1184.
- [73] 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 (2007), 534–574. <https://doi.org/10.3102/0034654307309920>
- [74] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.
- [75] Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43, 3 (1982), 450.
- [76] James R. Segedy, John S. Kinnebrew, and Gautam Biswas. 2012. Supporting student learning using conversational agents in a teachable agent environment. In *10th International Conference of the Learning Sciences: The Future of Learning, ICLS 2012 - Proceedings*. International Society of the Learning Sciences, IN, USA, 5.
- [77] Niraj Shah, Colleen Lewis, and Roxane Cairns. 2014. Analyzing equity in collaborative learning situations: A comparative case study in elementary computer science. In *Learning and Becoming in Practice: The International Conference of the Learning Sciences*. International Society of the Learning Sciences, Boulder, CO, USA.
- [78] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. *Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent*. Association for Computing Machinery, New York, NY, USA,

- 1–13. <https://doi.org/10.1145/3173574.3173965>
- [79] Norman A Sprinthal and Jacqueline R Scott. 1989. Promoting psychological development, math achievement, and success attribution of female students through deliberate psychological education. *Journal of counseling psychology* 36, 4 (1989), 440.
- [80] Reema Srivastava and Muhammed Rashid. 2018. Who is at edge—tutors or tutees? Academic, social and emotional elevation through peer tutoring. In *Proceedings of 1st MEC TESOL Conference*. Arab World English Journal, Texas, USA, 64–77.
- [81] Jason Tan, Gautam Biswas, and Daniel L Schwartz. 2006. Feedback for Metacognitive Support in Learning by Teaching Environments. *Proceedings of the Annual Meeting of the Cognitive Science Society* 28 (2006), 828–833.
- [82] Fumihide Tanaka, Kyosuke Isshiki, Fumiki Takahashi, Manabu Uekusa, Rumiko Sei, and Kaname Hayashi. 2015. Pepper Learns Together with Children: Development of an Educational Application. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids '15)*. IEEE, Whitestone, NY, USA, 270–275. <https://doi.org/10.1109/HUMANOIDS.2015.7363546>
- [83] Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning. *Journal of Human-Robot Interaction* 1, 1 (2012), 78–95. <https://doi.org/10.5898/JHRI.1.1.Tanaka>
- [84] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. 2016. Physical vs. Virtual Agent Embodiment and Effects on Social Interaction. In *Intelligent Virtual Agents*. Springer International Publishing AG, Switzerland, 412–415. https://doi.org/10.1007/978-3-319-47665-0_44
- [85] Robert J Vallerand, Luc G Pelletier, Marc R Blais, Nathalie M Briere, Caroline Senecal, and Evelyne F Vallieres. 1992. The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and psychological measurement* 52, 4 (1992), 1003–1017.
- [86] Suzanne Vossen, Jaap Ham, and Cees Midden. 2009. Social Influence of a Persuasive Agent: The Role of Agent Embodiment and Evaluative Feedback. In *Proceedings of the 4th International Conference on Persuasive Technology (Claremont, California, USA) (Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, Article 46, 7 pages. <https://doi.org/10.1145/1541948.1542007>
- [87] John Wagster, Jason Tan, Yanna Wu, Gautam Biwas, and Dan Schwartz. 2007. Do learning by teaching environments with metacognitive support help students develop better learning behaviors?. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (Nashville, TN), Vol. 29. Cognitive Science Society, Austin, TX, 695–700.
- [88] M. B. WILK and R. GNANADESIKAN. 1968. Probability plotting methods for the analysis for the analysis of data. *Biometrika* 55, 1 (1968), 1–17. <https://doi.org/10.1093/biomet/55.1.1>
- [89] Elmira Yadollahi, Wafa Johal, Ana Paiva, and Pierre Dillenbourg. 2018. When Deictic Gestures in a Robot Can Harm Child-robot Collaboration. In *Proceedings of the ACM Conference on Interaction Design and Children (IDC '18)*. ACM, New York, NY, USA, 195–206. <https://doi.org/10.1145/3202185.3202743>
- [90] Esther Ziegler and Elsbeth Stern. 2014. Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction* 33 (2014), 131–146.

Received October 2020; revised April 2021; accepted July 2021