Routledge
Taylor & Francis Group

Interactive Learning Environments

ISSN: 1049-4820 (Print) 1744-5191 (Online) Journal homepage: https://www.tandfonline.com/loi/nile20

Can an orchestration system increase collaborative, productive struggle in teaching-by-eliciting classrooms?

Kurt VanLehn, Hugh Burkhardt, Salman Cheema, Seokmin Kang, Daniel Pead, Alan Schoenfeld & Jon Wetzel

To cite this article: Kurt VanLehn, Hugh Burkhardt, Salman Cheema, Seokmin Kang, Daniel Pead, Alan Schoenfeld & Jon Wetzel (2019): Can an orchestration system increase collaborative, productive struggle in teaching-by-eliciting classrooms?, Interactive Learning Environments, DOI: 10.1080/10494820.2019.1616567

To link to this article: https://doi.org/10.1080/10494820.2019.1616567

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2019.1616567

Routledge
Taylor & Francis Group

Can an orchestration system increase collaborative, productive struggle in teaching-by-eliciting classrooms?

Kurt VanLehn[a], Hugh Burkhardt[b], Salman Cheema[c], Seokmin Kang[a], Daniel Pead[b], Alan Schoenfeld[d] and Jon Wetzel[a]

[a]Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA; [b]School of Education, University of Nottingham, Nottingham, UK; [c]Microsoft, Redmond, WA, USA; [d]School of Education, University of California, Berkeley, CA, USA

ABSTRACT

Mathematics is often taught by explaining an idea, then giving students practice in applying it. Tutoring systems can increase the effectiveness of this method by monitoring the students' practice and giving feedback. However, math can also be taught by having students work collaboratively on problems that lead them to discover the idea. Here, teachers spend the bulk of their time orchestrating collaborations and supporting students in building productively on each other's contributions. Our research question is: Can tutoring technology somehow make teaching-by-eliciting more effective? Using tutoring technology, we developed an intelligent orchestration system named FACT. While students solve problems in small groups, it makes recommendations to the teacher about which groups to visit and what to say. Data from over 50 iterative development trials (study 1) suggest that FACT increased neither the collaboration nor productivity of the students' struggle compared to paper-based classes. However, the data also suggest that when there is just one teacher in the classroom, then only a few of the groups that need a visit can get one. We modified FACT to directly send students the provocative questions that it formerly sent only to teachers. A pilot test (study 2) suggests that this version may increase productive struggle, but increasing collaboration remains an unsolved problem.

ARTICLE HISTORY
Received 26 February 2019
Accepted 5 May 2019

KEYWORDS
Collaborative learning; digital media; classroom orchestration systems; tutoring systems; formative assessment

1. Introduction

Two dichotomies need to be introduced: one for teaching and one for educational technology. Although both are familiar, neither have standard names. Although perhaps a bit interesting in themselves, we introduce them just to make it easier to explain our research problem.

1.1. Tell vs. Elicit and Tutors vs. Tools

Methods for teaching math can be dichotomized as Telling vs. Eliciting. In the Telling method, teachers explain some mathematical knowledge to the students and then guide them as they practice applying it. In the Eliciting method, teachers ask questions of the student that elicit a variety of ideas from them, and then guide the students' discussions so that they converge on a consensus that is also correct. Clearly, these are the extreme ends of a continuum of teaching methods. Nonetheless, let us use Tell and Elicit to refer to this distinction in teaching methods.

CONTACT Kurt VanLehn kurt.vanlehn@asu.edu

As an illustration of this distinction, suppose students need to learn how to answer questions of the form, "The initial price is $200. After a ___% increase, it is $300" or "The initial price is $300. After a __% decrease, it is $200."

- *Teaching by Telling:* The teacher tells the students, "You subtract the two prices. That gives you the amount of change. You divide that by the initial price." The teacher then gives students practice solving such problems. Student may find parts of the explanation confusing (e.g. are negative numbers allowed as the result of the subtraction?), so instruction also takes place during the practicing.
- *Teaching by Eliciting:* The teacher asks students to write their answers to both questions. The teacher then asks different students to show their answers to the class and explain why they think their answers are correct. The teacher does not indicate which answers correct. The teacher then divides the class into groups such that students with different answers are in the same group. The teacher asks each group to come to a consensus on the correct answers and why they are correct. After discussion dies down, the teacher has every group report their conclusions to the whole class. The teacher then provokes a reasoned debate between groups that hold different opinions. By asking provocative questions, such as, "What would an increase of 0% be?" the teachers guides the discussion to converge on the correct idea.

These different methods of teaching are as old as teaching itself. For example, Socrates used teaching by Eliciting. More recently, the Eliciting method has been characterized as learner-centered teaching (Bransford, Brown, & Cocking, 2000).

Now for the second dichotomy. Educational technology used by math students can be dichotomized as Tutors vs. Tools. A Tutor has a representation of the correct performance inside it, against which it assesses the student's behavior in order to give the student advice or feedback. A Tool is a system that students use in an open-ended way. The tool does not understand what tasks the students are trying to accomplish so it has no representation of correct performance. Nonetheless, the tool's design affords opportunities for the student to learn.

As an illustration of this distinction, consider two technologies for teaching the concept mentioned in the preceding illustration.

- *Tutor*: The system presents fill-in-the-blank exercises of both the original questions and ones that require students to show their intermediate reasoning steps: "The initial prices is $200 and the final price is $300. Thus, the change in price is $___, which is ___% of the initial price." The system gives feedback and hints, based on how the students fill in the blanks.
- *Tool*: The students are given a stack of cards with percentages on them. They are given a worksheet with pairs of price changes, such as: $200→$300; $300→$200. They are asked to put the cards bearing the percentages onto the arrows. However, the cards and the price changes are cleverly designed so that the only interpretation that allows all cards to be placed on arrows is the correct interpretation. If the students have a misconception, then they can place some of the cards in accord with their misconception, but they will not be able to place all of them. This tool could be implemented with a computer instead of paper. In fact, our system implements exactly this card-on-arrow Tool.

This distinction in computer support for learning has often been noted but no standard name exists. In a book contrasting these two technologies (Lajoie & Derry, 1993), the editors use "modelers" and "non-modelers" for advocates of Tutors and Tools, respectively.

Tools are usually used to support teaching by Eliciting, and Tutors are usually used to support teaching by Telling. However, Tools can be used for teaching by Telling. For example, a teacher can explain and demonstrate the Tell procedure mentioned above, and then give students feedback

as they apply it using the cards-on-arrow Tool. Thus, even if a Tool is designed for teaching by Eliciting, it can be used for teaching by Telling.

On the other hand, an ordinary Tutor may thwart teaching by Eliciting. Ordinary feedback and hints would ruin the elicitation process. Instead of trying to come up with their own ideas, students would try to figure out what will make the Tutor give positive feedback.

1.2. Our research question

Now that the Tell/Elicit and Tutor/Tool dichotomies have been defined, our research question can be simply stated: *Can Tutor technology help teaching by Eliciting?* As just noted, the Tutor should not give feedback to students on the correctness of their performance. However, there might be something else a Tutor could do that would help the elicitation process. Our initial hypothesis was that the Tutor could help teachers with two demanding aspects of teaching by Eliciting: formative assessment and visiting.

Formative assessment (also called *assessment for learning*) refers to the analysis that teachers do in order to understand a student's reasoning and the (mis-)conceptions behind it (Black & Wiliam, 1998a; Burkhardt & Schoenfeld, 2019). Understanding how students have reasoned is necessary if teachers want, for example, to create a group whose members hold different misconceptions or to ask students with different perspectives to show their work to the whole class. Many Tutors do a type of formative assessment that is limited to determining whether a student's performance is correct or not. For teaching by Eliciting, formative assessment has to understand misconceived as well as correct performances.

Another demanding aspect of teaching by Eliciting occurs when a teacher is circulating among students who are working on a task, such as placing cards on arrows, and the teacher stops to visit a group or a student. When teaching by Eliciting, the purpose of such a visit is not to correct the students' reasoning. Instead, the visit should encourage students to articulate and discuss their ideas with each other. Instead of simply nagging students by saying, for example, "remember to discuss your ideas with each other," teachers can ask provocative questions that challenge the particular misconceptions held by these students. Such questions might reignite a discussion that has died out.

We hypothesize that Tutor technology can conduct formative assessments and display appropriate provocative questions to teachers, who would then ask them to students. This might increase the effectiveness of the visit.

1.3. Overview of this paper

We developed a system that helps teachers with formative assessment and visiting (VanLehn et al., 2016; VanLehn et al., 2018a, 2018b; Wetzel et al., 2018). We named it Formative Assessment with Computational Technology (FACT). FACT was novel and could only be tested in real classrooms, so it took many classroom trials to converge on an acceptable design and robust implementation. Although classroom observers and teachers could see a steady increase in usability, it wasn't clear to the naked eye whether FACT was improving teaching by Eliciting. Fortunately, we collected video recordings during these trials. We have been analyzing them recently to see if we could discern an impact on either students or teachers. This paper reports one of those analyses.

The bottom line is that we no longer believe that helping *expert* teachers with formative analysis and visiting will make much difference. The expert teachers seem to be doing a good job already, and there appears to be little room for the system to help them do better. Although FACT might help *novice* teachers, we have only tested it with expert teachers.

However, our analyses of the videos does suggest a way that FACT could help even expert teachers. Expert teachers can only visit one group at a time even when there are many students who need a visit. Perhaps FACT could "visit" students on the teachers' behalf.

We augmented FACT with this new capability. The first pilot test was marred by a design flaw. The second shows some promise. Both are reported here.

So our new working assumption is that the bottleneck in teaching by Eliciting is not that expert teachers need help with formative analysis and visiting. The bottleneck is that more students need visits, and yet there is just one teacher. Tutoring technology seems capable of providing such "visits."

2. The classroom challenges exemplify teaching by eliciting

The project's original goal was to see if technology could improve the effectiveness of a specific set of mathematics lessons, called Classroom Challenges (Burkhardt and Schoenfeld, in press; Mathematics Assessment Project, 2018) or CCs. In their paper-based form, the lessons are known to be highly effective (Herman et al., 2015). They exemplify a particular kind of teaching by Eliciting (Black & Wiliam, 1998a; Burkhardt & Schoenfeld, in press), wherein teachers no longer give explanations and feedback, but instead keep students engaged in solving problems. To do so, teachers should analyze the students' work, detect the line of reasoning being followed and then ask questions that push the students further along that line. Because formative assessment is so prominent in this method of teaching, the method itself is called "formative assessment" and the lessons are often called "formative assessment lessons."

The CC students solve problems that are complex and open-ended. Problem solving lasts almost a whole lesson of 90–120 min. The problems are cleverly designed so that if the students stay engaged, they usually discover mistakes they have made and converge, often as a whole class, on a correct solution. Teachers are often surprised that students can figure out the math themselves, and this causes them to change their practice to be less teacher-centered (Inverness Research, 2016). Hence, the CCs are often used for professional development (FaSMEd, 2017; Joubert & Larsen, 2014; Research for Action, 2015).

The CC students work on large posters, to which they add cards and handwriting. The posters can become so complicated and messy that formative assessment becomes difficult even for expert teachers. A typical poster might end up with 30 cards on it, with handwriting next to most of them. Moreover, teachers can only see the current state of the poster and not its history. They cannot tell, for example, whether onegroup member did all the work or whether all members participated equally. Lastly, when teachers start a visit with students, they have only a few seconds to conduct a formative assessment of the poster. These facts suggested that teachers may have difficulty doing formative assessment when they are circulating around the classroom visiting students.

3. FACT

During the initial design sessions with CC authors and teachers, these design criteria became paramount:

(1) Students should work on digital media so that FACT could see their performance.
(2) Teachers should carry a dashboard with them as they circulated, so that FACT could display advice to the teacher prior to a visit.
(3) To avoid harming the effectiveness of the paper-based CCs, the digital media should mimic the paper-based CCs as closely as possible.

Thus, FACT students edit an electronic document called a poster. Posters can have movable cards on them. Figure 1 is an example. The posters and cards look nearly identical to the paper-based versions. Although FACT can be used on laptops or tablets, students in our trials always used a tablet with an active digital stylus so that they could use their normal handwriting.

FACT students can type or draw on both cards and the poster. Using their fingers or stylus, students could pan, zoom, move cards, pin cards or resize cards. Although the posters "handed out"

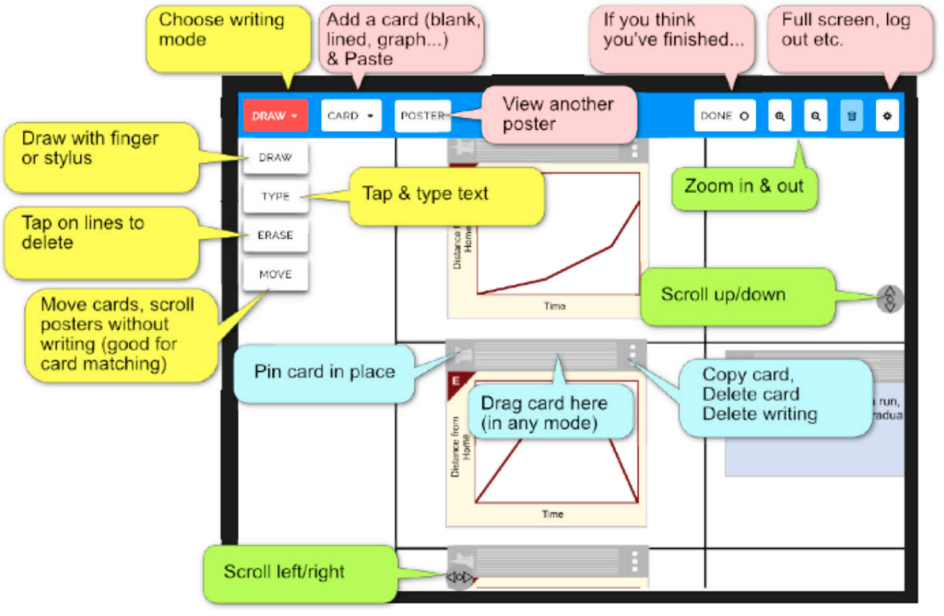

Figure 1. The FACT student screen, zoomed in to show part of a poster.

to students usually had cards such as the ones in Figure 1, students can also add blank cards or cards with blank lines, tables or *x-y* graphs.

Students can edit their own individual poster or their group's poster. When editing a group poster, all the members of the group can edit simultaneously, just as one does with a shared Google document. Each student's ink is a different color, so students and teachers can tell who has contributed what.

As students work, the teacher can monitor their work with a dashboard. The dashboard is designed to run on a tablet so it can be carried around the classroom. The dashboard shows a tile for every student in the class (see Figure 2). The tile displays the student's name and the ID number of the student's group. The top of the tile displays a progress bar indicating how much work the students have finished correctly (green) or incorrectly (red). The progress bars allow the teacher to identify groups or students that are far behind or far ahead of the rest of the class. The dashboard has several other features shown in Figure 2.

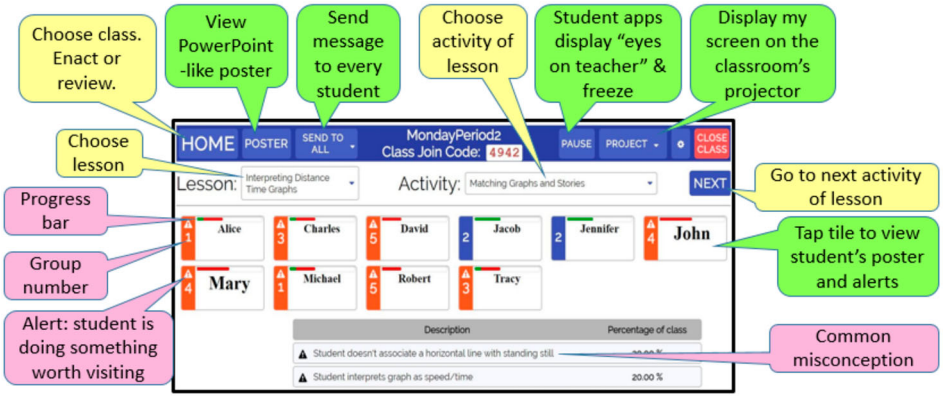

Figure 2. The teacher's dashboard.

To conduct a formative assessment of students' work, FACT has dozens of *issue detectors*. Most of them compare the students' work to expected work; these are called *product detectors*. The expected work is integrated into the detector itself, so different activities have different detectors.

FACT also has *process detectors*. These monitor the chronological pattern of students' edits. Different activities can have different process detectors. Many activities specify that students should collaborate, so there are several detectors for ways that groups fail to collaborate. For example, the "working separately" detector fires when it sees that the students in a group are simultaneously editing different parts of the poster. The "working alone" detector fires when one student in the group makes all the edits and works so rapidly that the students cannot possibly be discussing the edits as they are made. Similar collaboration detectors were quite accurate when used in a lab study (Viswanathan & VanLehn, 2018).

When a detector raises an issue, it remains active until the conditions that raised the issue no longer exist. Thus, in a class of 30 students, there can be a hundred active issues, with issues activating and deactivating every second. FACT shows a selected subset of these issues as orange bars on the teachers' dashboard (Figure 2). These are called alerts. When the teacher peeks at a student poster, FACT shows the highest priority issue in a sidebar (see Figure 3). The teacher can browse sideways to see other issues.

When an issue is displayed in the sidebar, it has one or more messages for students. The teacher can choose one and send it to the student or group. The messages appear in the students' Inbox, causing a button at the top of their screen to turn green (Figure 4). If they open their inbox, they can browse the messages they have received that are still relevant. Most messages refer to a card or set of cards, so those cards are highlighted with a green box.

4. Other orchestration systems

The CCs involve individual work, group work and whole-class discussions. The teacher must integrate workflows and ideas across all three planes of activity. "Classroom orchestration" refers to the planning and enacting of such integrated workflows (Dillenbourg & Jermann, 2010). A "classroom orchestration system" is intended to help the teacher with classroom orchestration (Prieto et al., 2011). FACT

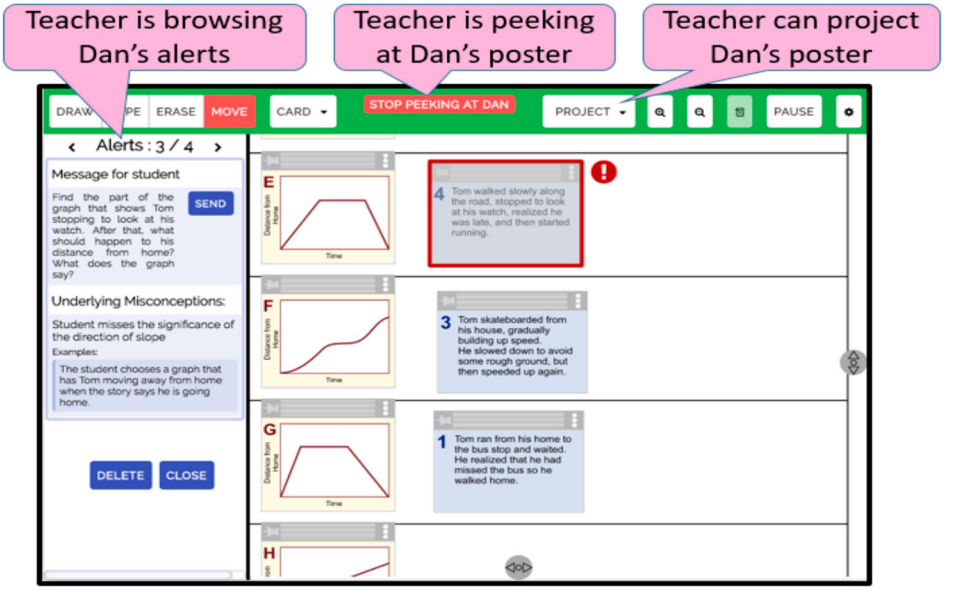

Figure 3. Teacher's view of an issue while peeking.

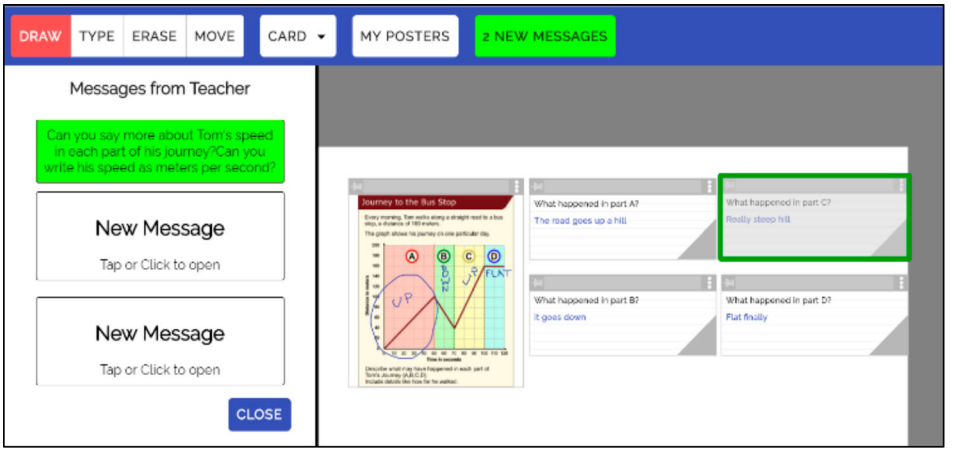

Figure 4. Student is viewing a message.

is a classroom orchestration system. This section introduces such systems, then situates FACT amongst them.

Classroom orchestration systems can be divided into several groups. One group (e.g. Alavi & Dillenbourg, 2012; Rojas, Garcia, & Kloos, 2012) is quite general but can only help the teacher with one orchestration function, such as hand raising. Most other orchestration systems attach a teacher dashboard to an existing system so that they can provide many features.

When dashboards are attached to modified operating systems or browsers (e.g. Lenovo, 2018; Netop, 2018; NetSupport, 2018; Xuetangx, 2018), the systems do not know the lesson plan nor can they access data from inside the applications. Nonetheless, their dashboards offer several orchestration features such as:

(1) Indicating which students have requested help and how long they have waited.
(2) Indicating which students or groups claim that they are done.
(3) Peeking at a student's work. That is, the dashboard displays the student's screen.
(4) Raising an alert when a student device goes off line.
(5) Pausing the class: The students' devices freeze and display "Eyes on teacher."
(6) Projecting a student's screen on the classroom's digital projector.
(7) Sending a text message to an individual student or to the whole class.

Some orchestration systems (e.g. Haklev et al., 2017; Prieto et al., 2014) provide an editor that allows the teacher to enter complex lesson plans involving multiple applications, workflows and group structures. Their dashboard features include the ones listed above plus:

(8) Advancing the class to the next activity in the lesson plan.
(9) Transmitting student work from one activity to subsequent activities.
(10) Modifying the lesson plan in the midst of enacting it.
(11) Modifying groups to deal with, e.g. students arriving late to class.

Some orchestration systems add a dashboard to a general purpose collaborative editing system (e.g. Looi, Lin, & Liu, 2008; Martinez-Maldonado et al., 2012) or a physical object system (e.g. Cuendet et al., 2013). These systems can measure the amount of editing done by each member of a group, so they can detect unbalanced participation. However, they do not understand how the

edits relate to the group's task, so they cannot detect errors nor display progress accurately. Such systems may have the dashboard features above plus:

(12) Displaying the proportion of participation (editing actions only) of each group.
(13) Raising alerts when a group's participation becomes too unbalanced.

Lastly, there are tutoring systems that have dashboards (Holstein, McLaren, & Aleven, 2018; Martinez-Maldonado, Yacef, & Kay, 2015; McLaren, Scheuer, & Miksatko, 2010; Mercier, 2016; Schwarz & Asterhan, 2011; Molenaar and Knoop-van Campen, in press). They understand how the student's edits relate to the task goals. This allows them to give feedback and hints directly to the students without involving the teacher. Their dashboards provide several more orchestration features for use during class:

(14) Accurate progress bars.
(15) Displaying the proportion of errors or error types per student.
(16) Raising alerts when a student's work has too many errors.

Although some systems in this last group are not complete orchestration systems because they do not support all three planes of activity (group, individual and whole-class), they often can assess a student's action relative to both domain norms (i.e. is the action correct?) and collaboration norms (i.e. is the group interaction co-constructive or transactive?). Thus, this class of orchestration systems could be said to be more "intelligent" than the other classes.

FACT has all 16 features listed above. It is intelligent, in the limited sense defined above. It differs from other intelligent orchestration systems primarily in that its pedagogical goal is different. Instead of giving students feedback as they practice of mathematics that they have been told, FACT helps teachers elicit mathematics from students.

5. How can success be measured?

The overall goal of FACT is to make the CCs more effective than their paper-based implementation. However, the CCs are not simple teaching-by-telling lessons whose effectiveness can be measured with conventional pre-tests and post-tests. This section introduces the problems involved in measuring the effectiveness of the CCs (and the effectiveness of teaching by Eliciting in general).

The CCs were designed to address a large set of goals, called TRU Math (Schoenfeld, 2014; Schoenfeld & Floden, 2014). Our evaluation can ignore the TRU Math goals that address the design of lessons, because the FACT and the paper-based CCs have exactly the same design, content and activities. For our evaluation, TRU Math specifies two relevant pedagogical goals: the teachers should engage in *formative assessment* (Black et al., 2004; Black & Wiliam, 1998a, 1998b) and the students should engage in *collaborative, productive struggle*. That is, students should work hard together to solve challenging, open-ended problems that afford many mathematical insights and discussions. A lesson's success is not measured by the number of correct answers it enables students to generate, but instead by the degree to which it engages the students in mathematically meaningful, productive, collaborative behavior.

Collaborative, productive struggle and formative assessment are characterizations of student and teacher behavior *in the classroom*. They are the first steps in a theory of change that has two parallel pathways:

(A) The CCs increase collaborative, productive struggle, which improves student math knowledge, identities and beliefs, which increase student mathematics learning.
(B) The CCs increase formative assessment, which improves teacher beliefs and practices, which increase student mathematics learning.

The viability of the second, professional development path was confirmed by a year-long study of the paper-based CCs with 2,690 students and 56 teachers (Herman et al., 2015). Compared to a constructed control group, the CC group achieved significant improvements on state-mandated standardized tests. The effect size was equivalent to 4.6 months of extra schooling. However, teachers used on average only 6 CCs during the year. This suggests that the main impact of the CCs was to convince teachers to change their teaching methods to include more formative assessment. These findings were consistent with interviews of teachers and students (Inverness Research, 2016). Teachers often commented that they changed their teaching because the CCs convinced them that students could learn from each other, so the teacher did not need to instruct them on every little bit of math. The CCs seem ideally suited for teacher professional development, and have often been used for such (FaSMEd, 2017; Herman et al., 2014; MDC, 2016).

Unfortunately, the existing evidence suggests that the other pathway (via student behavior) is not functioning well. Student scores on the math taught by the CCs did not improve much (Herman et al., 2015). Surveys of CC students indicated only modest changes in their math beliefs and attitudes (Inverness Research, 2016).

Because the second pathway (via teacher practices) is working, we need to improve the first pathway (via student beliefs). The first pathway has a proximal link (increasing collaborative, productive struggle) and a distal link (improving students' math beliefs, attitudes and identities). The main purpose of the FACT is to improve the first link in the first pathway. Thus, measuring collaborative, productive struggle seems the best way to evaluate FACT.

6. Study 1

A proper summative evaluation of FACT would require randomly assigning students to use either FACT or paper-based CCs. However, it would not make sense to randomly assign students in the same classroom to different treatments. Although random assignment of classrooms to treatments makes sense, it would require many more classrooms than we had access to.

Nonetheless, we wanted to test both FACT and our measures of collaborative, productive struggle, so we decided to use data that we had already collected for other purposes. Although these data allow us to use the methodology of summative evaluation, Study 1 is definitely *not* a summative evaluation.

Data for Study 1 came from our iterative development process. Before designing FACT, we chose 8 middle-school CCs to focus on. We chose CCs that were popular and sampled the whole range of CC activities. We then conducted several trials, where each trial consisted of a class enacting one of the 8 paper-based CC while we observed and video recorded it. These trials allowed us to design an initial version of FACT. While developing FACT, we conducted many more classroom trials. We were often able to conduct several trials at the same school on the same days. After analyzing the data from one set of trials and modifying FACT, we conducted another set of trials, sometimes with the same teachers and sometimes with different teachers. This process was repeated many times. This process is called iterative development, design-based research or formative evaluation. It generates copious data.

However, the data are not appropriate to use for comparing the treatment (FACT) to the baseline (paper) because many factors are neither adequately controlled nor randomly varied. Nonetheless, such a comparison is exactly what Study 1 does.

6.1. Participants

Teachers were recruited by the CC authors in England and by the Silicon Valley Mathematics Initiative in California. Teachers were in middle schools in Nottingham, UK or Silicon Valley, California. All the teachers were comfortable with technology. All were experienced CC teachers, and some had

participated in development of the paper-based CCs. Although we did not collect data on class size, the videos suggest that the classes had 20–35 students.

6.2. Procedure

A trial consists of a teacher enacting a CC with one classroom of students. From the 8 CCs that FACT supports, teachers chose one that was appropriate their class. A CC typically took 90–120 min and was usually enacted over two consecutive days. CCs w structured to spend over half the time on small group activities, with the rest of the time spent in whole class discussion or individual work. Almost all the groups consisted of just two students.

We observed both paper-based classes and FACT classes. In the Paper classes, pairs worked on a large sheet of paper (hence the name, poster). They cut out cards from another sheet of paper, arranged them on the poster and sometimes wrote explanations. When they were finished, they glued the cards down. In the FACT classes, pairs of students used electronic versions of posters and cards, as described earlier.

Every CC has its own teacher's guide. Every guide has a list of questions for teachers to ask students as a way to stimulate re-engagement and thinking. These are the provocative questions displayed in FACT's issue sidebar when teachers peek (Figure 3). For CC activities that did not have enough questions, the CC authors provided more.

Every trial was recorded by three cameras. One shoulder-mounted camera followed the teacher. Two other cameras were mounted on tall tripods. Each focused downward on the students' desk and recorded a single pair. The group's conversation was recorded by a boundary microphone on the table.

6.3. Collaboration hypothesis, measure & results

Because we have 2 rather different hypotheses, each will be presented in its own section along with the relevant measures and results. This section addresses collaboration.

Although FACT now has process detectors for collaboration, they were not present during the trials reported here. Although we hope that the collaboration detectors will eventually improve collaboration, our hypothesis is that FACT and the paper-based CCs will foster the same amount of collaboration during Study 1.

However, we feared that the traditional trials would show *more* collaboration than the FACT trials, for several reasons: Students rarely looked up from the tablet, so they lost eye contact. Their hands were near the tablet, so they may have gestured less. Their mouths were aimed at the tablet, so it may have been difficult for them to hear each other in a noisy classroom. Students often tried to refer to cards by pointing at their own tablet, which was difficult for their partner to see. Indeed, when the idea of replacing paper with a shared digital document was first suggested to the teachers and designers in the CC community, they were quite concerned that it may harm collaboration.

For each class, the 3 video streams were first synchronized and divided into 30-second segments. Only lessons with at least 30 segments (i.e. 15 min) of small group activity were included. Then, using the videos of pairs, trained human coders gave each such segment a code indicating the pair's behavior during that segment.

Our coding scheme was based on Michelene Chi's ICAP framework (Chi & Wylie, 2014). We preferred it over other schemes (Meier, Spada, & Rummel, 2007; Prieto et al., 2011; Roschelle, Dimitriadis, & Hoppe, 2013) because its categories are associated with learning gains. Our codes are shown in Table 1, along with their corresponding ICAP categories. The table rows are ordered from most desirable to least.

In order to check interrater agreement, 58% of the videos were coded by two coders. Interrater agreements (Kappa) averaged 0.79, which is considered acceptable. Disagreements on the codes assigned to segments were resolved in a meeting of the two coders and one of this paper's authors.

Table 1. Study 1 average percentage of number of segments per code.

Description (ICAP categories in parentheses)	FACT $N = 59$	Paper $N = 15$	FACT	Paper
Co-construction. Both students shared their thinking. Their contributions built upon each other. Transactivity. (Interactive)	2.8%	4.0%	3.7%	4.2%
Cooperation. The students worked simultaneously and independently on different parts of the poster. (Constructive + Constructive)	10.6%	7.1%	13.9%	7.4%
Unclear. Students explained their thinking, but the audio was not clear enough to determine whether it was one or two. (Constructive + Passive, or Interactive)	0.4%	0.9%	0.5%	0.9%
One explaining. One student explained his or her thinking (talking constructively) but the other student was either silent or merely agreeing. (Constructive + Passive)	4.0%	4.2%	5.3%	4.4%
None explaining. Students made edits, but neither explained their thinking. For example, one student might say "Let's put card B here," and the other student agrees. (Constructive + Passive)	53.8%	67.7%	70.8%	70.9%
The teacher was visiting the pair. (Passive + Passive)	3.3%	6.6%	4.3%	6.9%
The teacher was making a brief comment to the whole class, and these students were listening. (Passive + Passive)	1.1%	5.0%	1.4%	5.2%
The students were stuck and waiting for help from the teacher. (Disengaged + Disengaged)	0.6%	0%		
The students were done with the task and waiting for the teacher to give them something to do. (Disengaged + Disengaged)	11.8%	2.8%		
One student was off-task; the other worked without much talk. (Constructive + Disengaged)	4.5%	0.9%		
Both students were off-task. (Disengaged + Disengaged)	7.1%	0.8%		

We coded 15 Paper pairs and 59 FACT pairs. For each pair, we counted the number of segments per code. Because the total number of segments was different for different pairs, we converted the counts into percentages. Table 1 presents the averages of the percentages. The FACT distribution is not different from the Paper distribution (Chi-square, $p > .99$).

Notice that the last four rows, where at least one student was disengaged, show the greatest differences between the conditions. If these are summed, then 24% of the FACT segments have at least one student disengaged vs. 4.5% of the Paper segments. When these categories are eliminated, then the revised distributions are nearly identical, as shown in the rightmost 2 columns of Table 1. This suggests that the major difference between the conditions is that more students were disengaged in the FACT condition. This may be due to the iterative development of FACT. The FACT teachers may have paid less attention to their students than the Paper teachers because the FACT teachers spent more time helping with the development of FACT. Also, the tablets made it easy to erase digital ink, so FACT students would often doodle then erase their doodles almost immediately. This rarely occurred in Paper classes.

6.4. Productive struggle

As mentioned earlier, we hypothesize that productive struggle will be more frequent in FACT trials than traditional trials, so the goal of this analysis is to categorize pairs as either: (1) struggling productively, (2) struggling unproductively, or (3) not struggling. When students are not struggling, the teacher's guides and FACT's alerts suggest additional, more challenging work, such as writing out explanations. Thus, not struggling and struggling unproductively are both undesirable.

Unfortunately, the data were limited by several practical issues. Many of the student pairs said very little as they worked (the most frequent category in Table 1). Consequently, we could not use speech as a reliable indication of productive struggle. We could only use actions. Unfortunately, not all actions were clearly visible. Ultimately, only 6 Paper pairs and 18 FACT pairs could be analyzed. All tasks involved positioning cards only.

When a pair works quickly and makes few errors, it makes sense to classify it as not struggling. Figure 5 plots errors per card (horizontal) against minutes per card (vertical) of the FACT and Paper Pairs. For this analysis, an error is placing a card incorrectly initially, regardless of whether it

was later moved to a correct position. In all cases where a card was placed correctly initially, it was left there until the end of the activity. As Figure 5 indicates, 4 pairs were both fast and accurate, so they are classified as not struggling. In particular, one pair placed 30 cards correctly in 3 min, i.e. 6 s per card.

If a pair makes no errors but takes a long time per card, then it makes sense to classify it as struggling productively. As Figure 5 indicates, 2 pairs were 100% correct but slow.

Having classified 6 pairs, 14 FACT pairs and 4 Paper pairs remain to be split into struggling productively vs. unproductively. One sign of productive work is that students notice and fix their errors without any help from the teacher or FACT during the 30 s preceding the correction. Figure 6 shows the percentage of errors that each pair self-corrected. Pairs that self-corrected most of their errors should be classified as struggling productively, but exactly where to put the boundary is unclear. Table 2 reports the counts with two different boundaries: at 80% self-correct and at 40% self-corrected. Either way, more Paper pairs are productively struggling than FACT pairs, and the difference in distributions is reliable.

Another analysis of the same data can be done using errors as the unit of analysis. Caveat: Although this analysis avoids the somewhat arbitrary thresholds of the preceding analysis, it treats all errors the same regardless of whether they came from the same group or different groups. Anyway, FACT students self-corrected 33% of their 89 errors, whereas Paper students self-corrected 50% of their 12 errors. This was a reliable difference (Chi-square, $p < .001$), which suggests that productive struggle was more frequent for Paper students.

6.5. Discussion of study 1

The results of Study 1 suggest that FACT did not affect collaboration compared to paper-based classes, but FACT seems to have decreased the amount of productive struggle compared to paper-based classes. One possible explanation of the latter result is that because FACT teachers interacted with researchers during class, they paid less attention to their students. Thus, the students to took their classwork less seriously and engaged less, as suggested by Table 1.

However, the most striking result from Study 1 was just how few students in both Paper and FACT classes were engaged in collaborative, productive struggle. Fewer than 5% of the segments were

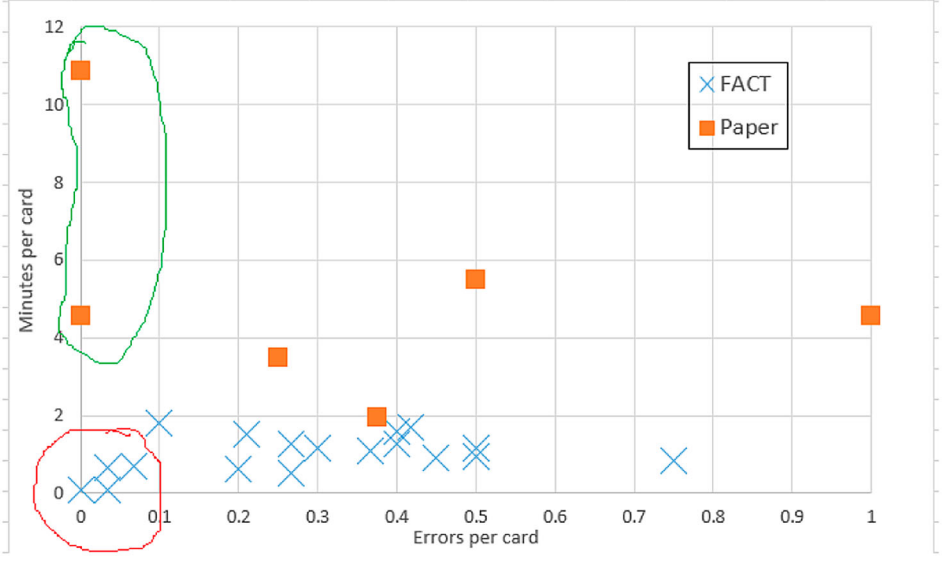

Figure 5. Red circled pairs are not struggling. Green circled pairs are struggling productively.

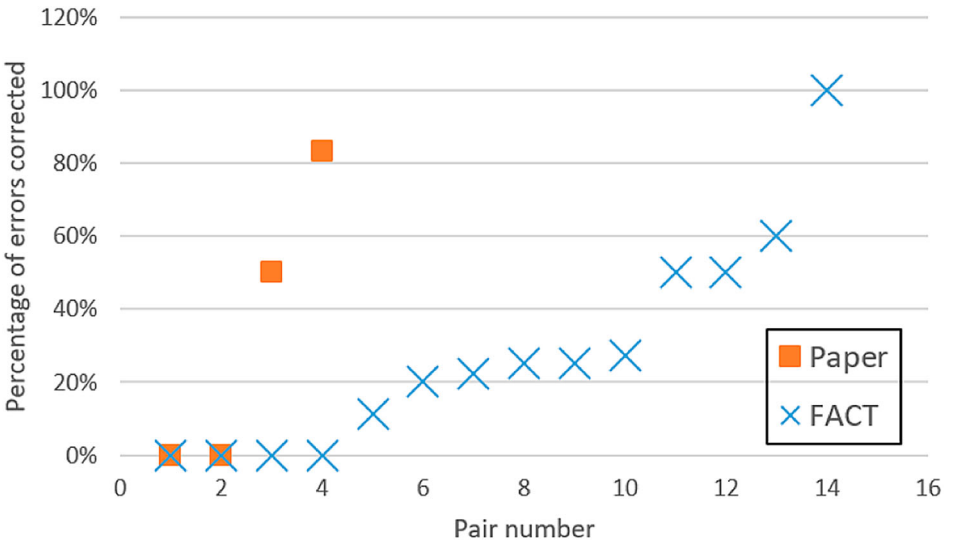

Figure 6. Distribution of self-correction proportions.

coded as co-constructive collaboration (Table 1). Approximately 70% of the engaged students were hardly talking at all. Less than half the pairs were productively struggling, even in the Paper classes (Table 2). A lack of collaborative, productive struggle is consistent with the results from early studies of the paper CCs, which showed that students' knowledge of the topics taught by the CCs was not increasing much (Herman et al., 2015) nor were their beliefs and attitudes toward math changing much (Inverness Research, 2016).

Because FACT gave advice to teachers and not students, we wondered if it had any impact on teacher behavior at all. Our analysis of teacher behavior is not complete, but so far it appears that the FACT teachers and Paper teachers behaved similarly. In particular, the mean time between visit starts (4:25 for FACT vs. 5:40 for Paper; $p = 0.182$) did not differ.

The frequency of teacher visits (about one visit every 4 or 5 min) suggests what might be going wrong. The figures above indicate that some groups are not productively struggling and almost all are not collaborating properly. Thus, when a teacher finishes one visit and is deciding whom to visit next, *almost every group in the class needs to be visited*. Even if FACT helps the teacher make an optimal choice of whom to visit, many other groups are left without a visit. FACT would need to dramatically shorten visits in order to allow the teacher to visit everyone who need a visit. Moreover, the figure of about 5 minutes per visit has been observed in other studies of classroom orchestration systems (Molenaar & Knoop-van Campen, 2017), so it may not be possible to greatly increase the frequency of teacher visits.

Thus, the pedagogical problem may not be that teachers are visiting the wrong groups or saying the wrong things. It may simply be that there aren't enough teachers in the classroom, so too few groups receive visits.

Table 2. Productive struggling classifications.

	40% threshold		80% threshold	
	FACT	Paper	FACT	Paper
Not struggling	4	0	4	0
Struggling unproductively	10	2	13	3
Struggling productively	4	4	1	3
chi-square (FACT is expected)	0.0055		0.0004	

This hypothesis prompted us to modify FACT. With some trepidation, we made it more like a traditional Tutor in that it would send messages directly to students instead of communicating only with teachers. Here's how.

As mentioned earlier, when teachers Peek at a student, they see a sidebar with an issue. The sidebar also shows provocative questions that the teachers can ask during a visit. Teachers can push a Send button to send a question directly to students. It then appears as a message in the student's inbox.

FACT was modified to, so to speak, push the Send button itself. After an activity began, it waited 5 minutes so students could get well started, then it would send students a message from their highest priority issue. It would always wait at least 2 min between sending messages to a group. We called this policy "auto-sending."

Our first trial of the auto-send capability was a disaster. It was also the first trial with the process detectors turned on. They turned out to be too aggressive. For example, one process detector repeatedly sent the messages "Are you working together with your partner?" Students rapidly learned this, and stopped opening their inbox when new messages arrived. They referred to FACT's messages as spam.

Because the school year was almost over, we did not have time to fix the process detectors. Thus, for the next trial, we simply turned them off. Auto-sending of product issue messages continued.

7. Study 2

We designed the last trial of FACT as an AB evaluation. That is, some of the classes had the full FACT system, and others had FACT with all its detectors turned off. When the detectors were off, the teachers saw alerts neither on the dashboard nor when Peeking, and FACT auto-sent no messages. Because the trial involved only a small number of classrooms, we did not expect to find statistically reliable results. We were basically just pilot testing the methodology.

7.1. Participants

The participating classrooms were in a middle-class Silicon Valley, California suburb. Two teachers were recruited by the Silicon Valley Mathematics Initiative. Both were experts at enacting the Classroom Challenges and comfortable with technology, but they had not participated in the design of FACT nor had they used FACT before.

7.2. Procedure

We asked teachers to choose Classroom Challenges that were at the right level of difficulty for their particular classes. Teacher A taught 3 classes "Solving Linear Equations", and Teacher B taught 2 classes "Increasing and Decreasing Quantities by a Percent". The class periods were 90 min long, so the whole CC could be taught in one period.

For both teachers, the first class was taught with analytics turned off. For the remaining classes, the analytics were turned on. As the video crew was setting up before a teacher's first class, the teacher was given a 10 min introduction to FACT, focusing on how to use its dashboard. As the first class was leaving and the second class was entering, the teacher was given a 3 min explanation of FACT's alerts. A FACT researcher was present in all classes to help with technology issues.

The Analysis Off classes were an 8th grade accelerated class and a 7th grade gifted class. The Analysis On classes were an 8th grade gifted class, a 7th grade accelerated class and an 8th grade Common Core class.

We used the same video recording procedure as in Study 1, except that 3 pairs were recorded in teacher A's classes and 4 pairs were recorded in teacher B's classes.

7.3. Collaboration

We used the same analytic methods as in Study 1's analysis of collaboration. Only lessons with at least 30 segments (i.e. 15 min) of small group activity were included, so one Analysis On lesson was omitted (the 8th grade Common Core class). There were 7 student pairs who worked with Analysis On, and 7 pairs who worked with Analysis Off.

Results are shown in Table 3. For comparison, the table shows also results from the FACT condition of Study 1. The Analysis On distribution did not differ from the Analysis Off distribution (Chi square, $p > .99$).

7.4. Productive struggle

For productive struggle, the hypothesis, measures and analytic methods were similar to those in Study 1. One Analysis Off pair made no errors and spent less than a minute per card, so it was classified as not struggling. One Analysis Off pair and one Analysis On pair made no errors, but spent several minutes per card, so we classified both as struggling productively. Figure 7 shows the percentage of errors self-corrected by the remaining pairs, where "self-correction" means that the students corrected an error without speaking with the teacher or reading a message from FACT during the preceding 30 s. Which ones are classified as productively struggling depends on the threshold. Table 4 shows the distributions with four different thresholds, along with statistical tests of the reliability of the differences. For most thresholds, the Analysis On pairs exhibited more productive struggle than the Analysis Off pairs.

When we ignore the pairs and just count number of errors in each condition, then the Analysis On students self-corrected 72% of their 18 errors, whereas the Analysis Off students self-corrected 38% of their 13 errors. This is a reliable difference (Chi-square, $p < 0.001$).

8. Discussion

8.1. Summary

While iteratively developing FACT, we video recorded both FACT classes and paper-based classes enacting the same lessons (called CCs). To get a rough sense of whether FACT was making a difference, we analyzed the videos. Unfortunately, the Paper classes exhibited more productive struggle than the FACT classes. Fortunately, the frequency of collaboration was not different.

Table 3. Average percentage of number of segments per code in On pairs, Off pairs, Study 1's FACT pairs.

Description (ICAP categories in parentheses)	On $N = 7$	Off $N = 7$	Study 1 $N = 59$
Co-construction. Both students shared their thinking. Their contributions built upon each other. Transactivity. (Interactive)	2.5%	7.3%	2.8%
Cooperation. The students worked simultaneously and independently on different parts of the poster. (Constructive + Constructive)	4.0%	2.1%	10.6%
Unclear. Some students explained their thinking, but the audio was not clear enough to determine whether it was one or two. (Constructive + Passive, or Interactive)	0.0%	0.0%	0.4%
One explaining. One student explained his or her thinking (talking constructively) but the other student was either silent or merely agreeing. (Constructive + Passive)	8.7%	6.7%	4.0%
None explaining. Students made edits, but neither explained their thinking. For example, one student might say "Let's put card B here," and the other student agrees. (Constructive + Passive)	60.2%	65.1%	53.8%
The teacher was visiting the pair. (Passive + Passive)	4.0%	1.5%	3.3%
The teacher was making a brief comment to the whole class, and these students were listening. (Passive + Passive)	1.3%	1.1%	1.1%
The students were stuck and waiting for help from the teacher. (Disengaged + Disengaged)	0.8%	0.0%	0.6%
The students were done with the task and waiting for the teacher to give them something else to do. (Disengaged + Disengaged)	13.9%	13.9%	11.8%
One student was off-task; the other worked without much talk. (Constructive + Disengaged)	2.2%	0.0%	4.5%
Both students were off-task. (Disengaged + Disengaged)	2.3%	2.4%	7.1%

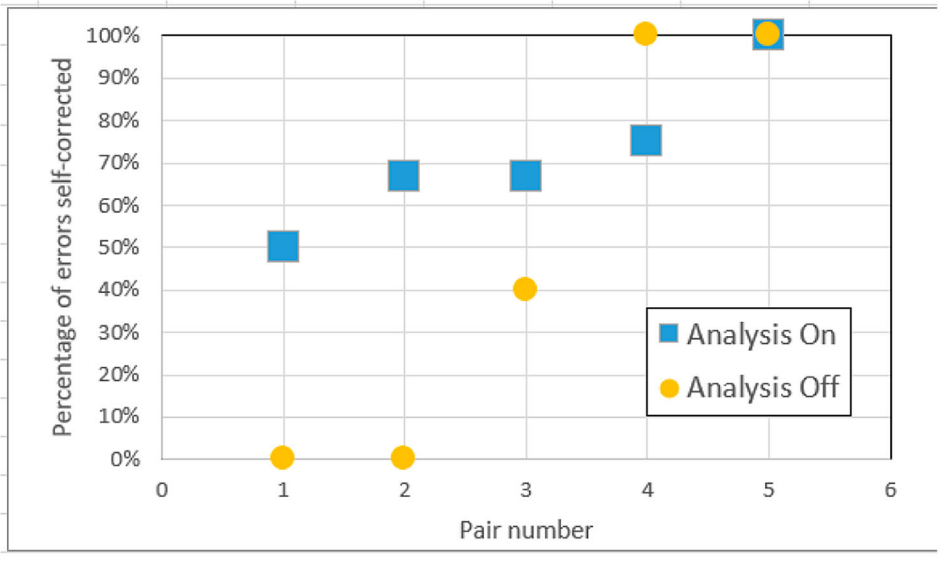

Figure 7. Percentage errors self-corrected in Study 2.

However, the video analyses suggest that more teacher visits would be helpful. Many groups were either not interacting collaboratively, not productively struggling or both. That is, the pedagogical problem in our classes may not be that the teachers were visiting the wrong groups or holding ineffective conversations during their visits. This is particularly plausible given that all the teachers in Study 1 were experts at teaching the CCs. Instead, the problem may be simply that there weren't enough visits, because there was only one teacher but almost all groups needed visits.

Thus, we modified FACT to automatically send the messages that teachers could send. We pilot tested this auto-send capability in two trials. In the first trial, FACT sent far too many messages, so students ignored them. In the second trial, we compare two versions of FACT, with its analytics turned either On or Off. Groups in the Analysis On condition more frequently struggled productively than groups in the Analysis Off condition. This is consistent with our hypothesis that the bottleneck in our classes is that more groups need to be visited, and that FACT's auto-send feature can at least partially fill the gap.

8.2. Interpretations

Although we refer to the conditions in Study 2 as Analysis On vs. Off, many other factors co-varied with the manipulation including the classes, the time of day and the familiarity of the teachers with FACT. Thus, we cannot conclude that turning the analysis on *caused* students to correct more errors. Larger, better-controlled experiments are needed.

Study 2's positive finding might be dismissed as yet another demonstration that feedback tends to increase learning (Hattie & Timperley, 2007). For example, Molenaar, Knoop-van Campen, and

Table 4. On vs. Off classification of pairs, various thresholds.

	80% threshold		60% threshold		41% threshold		39% threshold	
	On	Off	On	Off	On	Off	On	Off
Not struggling	0	1	0	1	0	1	0	1
Struggling unproductively	4	3	1	3	0	3	0	2
Struggling productively	3	3	6	3	7	3	7	4
chi-square (Off expected)	0.513		0.069		0.009		0.072	

Hasselman (2017) found that a dashboard-equipped tutoring system that gave feedback to students caused them to learn more than students without the system. If FACT's auto-sent messages are viewed as feedback, then our results are consistent with those of Molenaar et al. and many other studies.

However, FACT's messages differ from the feedback messages of most tutoring systems. FACT's messages do not indicate whether the students' work is correct or incorrect. FACT does not use a hint sequence that eventuates in telling the students the correct action to take. Instead, its product issues only offer questions to think about which, at first glance, seem to have nothing to do with the mathematical issues that caused the detector to fire (see examples of such questions in Figures 3 and 4). The pedagogical goal of FACT is different from most tutoring systems. It is not trying to decrease errors. Instead, it is trying to get students to discover errors in their thinking and thus discover some math.

Despite the preliminary success reported here, many challenges remain. FACT did not improve collaboration, probably because its collaboration detectors were never good enough to be used. It may need to monitor the speech of students in order to reliably measure co-constructive collaboration. FACT's actions are not synchronized with the teacher's actions because it does not monitor what the teacher is doing. Although FACT helps teachers when they circulate among students, it provides no help during whole-class activities. FACT has only been used with expert CC teachers. It may need modification to accommodate teachers who are just learning how to teach by Eliciting. FACT will need a new kind of intelligence in order to understand and help both teachers and students.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Science Foundation [grant number 1840051]; Bill and Melinda Gates Foundation [grant number OPP1061281].

Notes on contributors

Kurt VanLehn is the Diane and Gary Tooker Chair for Effective Education in Science, Technology, Engineering and Math in the Ira Fulton Schools of Engineering at Arizona State University. He has published over 125 peer-reviewed publications, is a fellow in the Cognitive Science Society, and is on the editorial boards of Cognition and Instruction and the International Journal of Artificial Intelligence in Education. Dr. VanLehn's research focuses on intelligent tutoring systems, classroom orchestration systems, and other intelligent interactive instructional technology.

Hugh Burkhardt was Director of the Shell Centre for Mathematical Education and Professor in the Department of Mathematics at Nottingham from 1976-92. Since then he has led a series of impact-focused research and development projects. His work has been recognised by an ISDDE Prize for Lifetime Achievement and, with Malcolm Swan, as first recipients of the Emma Castelnuovo Prize of the International Commission on Mathematical Instruction.

Salman Cheema is a research scientist at Microsoft. His 2014 Computer Science PhD was from University of Florida. He has published 8 refereed publications and secured 1 patent. He was won first prize at 3 software competitions and was awarded a Provost's Fellowship in 2007.

Seokmin Kang is a Postdoctoral Research Associate in the Ira Fulton Schools of Engineering at Arizona State University. His 2012 PhD in Human Cognition and Learning was from Columbia University. He has published 9 journal articles and 22 refereed conference papers. His 2016 article was selected a "featured article" by the Psychonomics Society.

Daniel Pead has contributed to the design of computer-based learning and teaching materials in many Shell Centre Projects since the early 1980s, becoming IT director for the Shell Centre and MARS. He research work includes the strengths and limitations of computers in assessing complex student performances.

Alan Schoenfeld is the Elizabeth and Edward Conner Professor of Education and Affiliated Professor of Mathematics at the University of California at Berkeley, and Honorary Professor in the School of Education at the University of

Nottingham. His research has been recognized with, among many honors, the Presidency of the American Educational Research Association and the Felix Klein Prize of the International Commission on Mathematical Instruction.

Jon Wetzel is an Assistant Research Scientist in the Ira Fulton Schools of Engineering at Arizona State University. His 2014 PhD from Northwestern University was in Computer Science with Specialization in Cognitive Science. He has published 7 journal articles and 13 refereed conference publications, and secured 1 patent.

References

Alavi, H., & Dillenbourg, P. (2012). An ambient awareness tool for supporting supervised collaborative problem solving. *IEEE Transactions on Learning Technologies*, *5*(3), 264–274.

Black, P., Harrison, C., Lee, C., Marshal, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8–21.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academy Press.

Burkhardt, H., & Schoenfeld, A. H. (2019). Formative assessment in mathemtaics. In R. E. Bennett, G.J. Cizek, & H.L. Andrade (Eds.), *Handbook of formative assessment in the Disciplines* (pp. 35–67). New York, NY: Routledge.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Cuendet, S., et al. (2013). Designing augmented reality for the classroom. *Computers & Education*, *68*(557-569).

Dillenbourg, P., & Jermann, P. (2010). Technology for classroom orchestration. In M. S. Khine & I. M. Saleh (Eds.), *New Science of learning: Cognition, Computers and collaboration in Education* (pp. 525–552). New York, NY: Springer.

FaSMEd. (2017). *Improving progress for lower achievers through formative assessment in science and mathematics education (FaSMEd)*. Retrieved from https://research.ncl.ac.uk/fasmed/

Haklev, S., et al. (2017). *Orchestration graphs: Enabling rich social pedagogical scenarios in MOOCs*. In *Proceedings of the Fourth (2017) ACM Conference on Learning@Scale*. Cambridge, MA.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Herman, J., et al. (2014). *Implementation and effects of LDC and MDC in Kentucky districts (CRESST Policy Brief No. 13)*. University of California, National Center for Research on Evaluation, Standards, & Student Testing (CRESST): Los Angeles, CA.

Herman, J., et al. (2015). The implementation and effects of the mathematics design collaborative (MDC): Early findings from Kentucky ninth-grade algebra 1 courses (CRESST Report 845).University of California at Los Angeles, National Center for Research on Evaluation, Standards and Student Testing: Los Angeles, p. 144.

Holstein, K., McLaren, B., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *Artificial intelligence in Education* (pp. 154–168). London: Springer.

Inverness Research. (2016). *Mathematics assessment program (MAP): Project portfolio*. Retrieved from http://inverness-research.org/mars_map/1_welcome.html

Joubert, M., & Larsen, J. (2014). *A patchwork of professional development: One teacher's experiences over a school year*. In *Proceedings of the 8th British Congress of mathematics Education*, pp. 207–214.

Lajoie, S. P., & Derry, S. J. (1993). *Computers as cognitive tools*. Hillsdale, NJ: Erlbaum.

Lenovo. (2018). *LanSchool: A classroom management system*.

Looi, C.-K., Lin, C.-P., & Liu, K.-P. (2008). Group scribbles to support knowledge building in a jigsaw method. *IEEE Transactions on Learning Technologies*, *1*(3), 157–164.

Martinez-Maldonado, R., et al. (2012). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In S. A. Cerri et al. (Eds.), *Intelligent tutoring systems: 11th international conference, ITS* 2012 (pp. 482–492). Berlin: Springer-Verlag.

Martinez-Maldonado, R., Yacef, K., & Kay, J. (2015). TSCL: A conceptual model to inform understanding of collaborative learning processes at interactive tabletops. *International Journal of Human-Computer Studies*, *83*, 62–82.

Mathematics Assessment Project. (2018). *Assessing 21st century Math*. 2018. Retrieved from http://map.mathshell.org/index.php

McLaren, B., Scheuer, O., & Miksatko, J. (2010). Supporting collaborative learning and e-discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence and Education*, *20*, 1–46.

MDC. (2016). *The mathematics design collaborative*. Retrieved from http://k12education.gatesfoundation.org/student-success/high-standards/literacy-tools/mathematics-design-collaborative/

Meier, A., Spada, H., & Rummel, N. (2007). *A rating scheme for assessing the quality of computer-supported collaboration processes*. *International Journal of Computer-Supported Collaborative Learning*, *2*, 63–86.

Mercier, E. (2016). Teacher orchestration and student learning during mathematics activities in a smart classroom. *International Journal Smart Technology and Learning*, *1*(1), 33–52.
Molenaar, I., Knoop-van Campen, C. A., & Hasselman, F. (2017). The effects of learning analytics empowered technology on students' arithmetic skill development. In *Learning analytics and knowledge LAK '17*. Vancouver, BC: ACM.
Molenaar, I., & Knoop-van Campen, C. A. (2017). *Teacher dashboards in practice: Usage and impact*. In *European conference on technology enhanced learning: EC-TEL*. Springer, pp. 125–138.
Molenaar, I., & Knoop-van Campen, C. A. (in press). How teachers make dashboard information actionable. *IEEE Trans. on Learning Technologies*, 614–615.
Netop. (2018). *Vision: A classroom management system*. Retrieved from https://www.netop.com/edu.htm
NetSupport. (2018). *School: A classroom management system*.
Prieto, L. P., et al. (2011). Orchestrating technology enhanced learning: A literature review and conceptual framework. *International Journal of Technology Enhanced Learning*, *3*(6), 583–598.
Prieto, L. P., et al. (2014). Supporting orchestration of CSCL scenarios in web-based distributed learning environments. *Computers & Education*, *73*, 9–25.
Research for Action. (2015). *MDC's influence on teaching and learning*. Philadelphia, PA: Author.
Rojas, I. G., Garcia, R. M. C., & Kloos, C. D. (2012). Enhancing orchestration of lab sessions by means of awareness mechanisms. In A. Ravenscroft et al. (Eds.), *7th European conference of technology enhanced learning* (pp. 113–125). Berlin: Springer.
Roschelle, J., Dimitriadis, Y., & Hoppe, U. (2013). *Classroom orchestration: Synthesis*. *Computers & Education*, *69*, 523–526.
Schoenfeld, A. H. (2014). *What makes for powerful classrooms, and how can we support teachers in creating them? A story of research and practice, productively intertwined Educational Researcher*, *43*(8), 404–412.
Schoenfeld, A. H., & Floden, R. (2014). *The TRU math scoring rubric*. Retrieved from http://ats.berkeley.edu/tools/TRUMathRubricAlpha.pdf
Schwarz, B. B., & Asterhan, C. (2011). E-moderation of synchronous discussions in educational settings: A nascent practice. *Journal of the Learning Sciences*, *20*(3), 395–442.
VanLehn, K. (2016). Some less obvious features of classroom orchestration systems. In L. Lin & R. K. Atkinson (Eds.), *Educational technologies: Challenges, applications and learning Outcomes* (pp. 73–94). Hauppauge, NY: Nova Scientific Publishers.
VanLehn, K., et al. (2018a). The effect of digital versus traditional orchestration on collaboration in small groups. In C. Rosé et al (Eds.), *Artificial intelligence in education: Proceedings of the 19th international conference* (pp. 369–373). Berlin: Springer.
VanLehn, K., et al. (2018b). How can FACT encourage collaboration and self-correction? In K. Millis et al. (Eds.), *Multi-disciplinary approaches to deep learning*. New York, NY: Routledge.
Viswanathan, S. A., & VanLehn, K. (2018). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies*, *11*(2), 230–242.
Wetzel, J., et al. (2018). A preliminary evaluation of the usability of an AI-infused orchestration system. In C. Rosé et al. (Eds.), *Artificial intelligence in education: Proceedings of the 19th international conference* (pp. 378–383). Berlin: Springer.
Xuetangx. (2018). *Rain Classroom*.