The Classroom as a Dashboard: Co-designing Wearable Cognitive Augmentation for K-12 Teachers

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, PA, 15213
kjholste@cs.cmu.edu

Gena Hong
Carnegie Mellon University
Pittsburgh, PA, 15213
honggena@gmail.com

Mera Tegene
Carnegie Mellon University
Pittsburgh, PA, 15213
mtegene@cs.cmu.edu

Bruce M. McLaren
Carnegie Mellon University
Pittsburgh, PA, 15213
bmclaren@cs.cmu.edu

Vincent Aleven
Carnegie Mellon University
Pittsburgh, PA, 15213
aleven@cs.cmu.edu

ABSTRACT

When used in classrooms, personalized learning software allows students to work at their own pace, while freeing up the teacher to spend more time working one-on-one with students. Yet such personalized classrooms also pose unique challenges for teachers, who are tasked with monitoring classes working on divergent activities, and prioritizing help-giving in the face of limited time. This paper reports on the co-design, implementation, and evaluation of a wearable classroom orchestration tool for K-12 teachers: mixed-reality smart glasses that augment teachers' real-time perceptions of their students' learning, metacognition, and behavior, while students work with personalized learning software. The main contributions are: (1) the first exploration of the use of smart glasses to support orchestration of personalized classrooms, yielding design findings that may inform future work on real-time orchestration tools; (2) Replay Enactments: a new prototyping method for real-time orchestration tools; and (3) an in-lab evaluation and classroom pilot using a prototype of teacher smart glasses (Lumilo), with early findings suggesting that Lumilo can direct teachers' time to students who may need it most.

CCS CONCEPTS

- **Human-centered computing ~ User centered design**
- **Applied computing ~ Computer-assisted instruction**

KEYWORDS

Real-time analytics, personalized classrooms**,** co-design, orchestration, awareness, prototyping, K-12, mixed-reality

ACM Reference format:

K. Holstein, G. Hong, M. Tegene, B.M. McLaren, and V. Aleven. 2018. The Classroom as a Dashboard: Co-designing Wearable Cognitive Augmentation for K-12 Teachers. In *Proceedings of the International Conference on Learning Analytics and Knowledge, Sydney, Australia, March 2018 (LAK'18)*, 10 pages. DOI: 10.1145/3170358.3170377

1 INTRODUCTION

In recent years, there has been increasing interest in personalized classroom models within K-12 education [25]. In personalized classrooms, students progress along individualized learning pathways, while the teacher's role is transformed from that of a lecturer at the front of the class to that of a facilitator of students' self-paced learning [28]. To support this kind of highly personalized instruction, schools are increasingly using personalized learning software for use in their classrooms [25].

One form of personalized learning software, intelligent tutoring systems (ITSs) [17], allows students to work at their own pace while providing detailed, step-by-step guidance through complex learning activities. An advantage of such systems, when used in classrooms, is that they free up the teacher to spend more time working one-on-one with students (e.g., [12, 28]). However, they also present teachers with unique challenges, as teachers are tasked with monitoring classrooms that are likely working on a broad range of divergent educational activities at any given time [3, 21, 27]. Thus, there is a great need for usable real-time orchestration tools that can support teachers in monitoring personalized classrooms, and effectively allocating help and attention across students, in the face of limited time [3, 18].

Prior work in Learning Analytics and Human-Computer Interaction has adopted user-centered and participatory approaches to the design of real-time awareness tools for teachers working in personalized classrooms (e.g., [1, 18, 21]). However, most of this work has focused on designing tools for *university-level* instructors. Our own recent work has focused on better understanding K-12 teachers' real-time information needs in personalized classrooms – using the notion of "teacher superpowers" as a probe to elicit key needs and desires that real-time analytics might address [11]. In parallel, recent design and ethnographic work has begun to investigate the potential of emerging wearable technologies for teacher support (e.g., [11,26,29]). Such technologies hold great promise to enhance teacher awareness, while allowing teachers to keep their heads up and eyes focused on their classroom – acknowledging the highly active role teachers play in personalized classrooms [12, 26, 28].

While prior work suggests that teachers may prefer wearables over handheld devices for use in personalized classroom contexts [11, 26], this work has not involved the user-centered design and evaluation of an actual wearable orchestration tool. Building on findings from our prior user-centered design research with K-12 teachers, in which teachers suggested the idea of having smart glasses that could augment their real-time perceptions of students' learning, metacognition, and behavior [11], this paper presents the first exploration, to our knowledge, of the use of smart glasses to support teachers in personalized classrooms. Working with sixteen K-12 math teachers, we have iteratively designed and developed Lumilo: mixed-reality smart glasses that support teachers in orchestrating personalized classrooms. We focus on classrooms in which students work with ITSs – leveraging these systems as "classroom sensors", that can generate rich, actionable analytics to support teachers' real-time decision-making. In particular, we use automated, real-time detectors of student learning and behavior within ITSs, to provide teachers with several of the "superpowers" identified in [11], discussed in the next section.

The structure of this paper is as follows: we first present our iterative co-design process with K-12 teachers. As part of this process, we introduce Replay Enactments (REs): a new prototyping method for real-time orchestration tools, which builds upon prior prototyping methods from both Learning Analytics [20, 21] and Human-Computer Interaction [23]. We then present Lumilo: a pair of mixed-reality smart glasses designed to support K-12 teachers in orchestrating personalized class sessions. Using REs, we find early evidence that teachers using Lumilo spend significantly more time attending to students who would otherwise learn less from the educational software alone. While prior work suggests that real-time awareness tools can successfully direct teachers' attention to students or groups who are currently *low performing* (e.g., [16]), the present analyses represent the first evaluation of relationships with student *learning gains* (albeit in simulated classrooms). Finally, based on our findings from prototyping sessions and a classroom pilot, we present design opportunities for future teacher support tools, and directions for future research.

2 BACKGROUND

2.1 Intelligent Tutoring Systems in-the-wild

Intelligent tutoring systems (ITSs) are a class of advanced learning technologies that provide students with step-by-step guidance during complex problem-solving practice and other learning activities. These systems continuously adapt instruction to students' current 'state' (a set of measured variables, which may include moment-by-moment estimates of student knowledge, metacognitive skills, affective states, and more) [6]. Several meta-reviews have indicated that ITSs can enhance student learning, compared with other learning technologies or traditional classroom instruction (e.g., [17]). However, ethnographic studies have revealed that, in K-12 classroom settings, teachers and students often use ITSs in ways not originally anticipated by ITS designers (e.g., [11, 12, 28]). For example, Schofield et al. found that rather than replacing the teacher, a key benefit of using such AI tutors in the classroom may be that they free teachers to provide more individualized help while students work with the tutor. Although students tended to perceive that teachers provide better one-on-one help than an ITS, they also preferred ITS class sessions over more traditional sessions – in part because of this shift in teacher-student interactions [28].

2.2 Intelligent tutors as teachers' aides

Recently, some work has begun to explore the value ITSs might provide to *teachers* in K-12 classrooms, and to investigate teachers' needs and desires for real-time support in ITS classrooms. However, the design of effective support tools for teachers working in these contexts remains a largely open, challenging design problem [27]. In a series of user-centered design interviews with middle school math teachers, we previously conducted a broad exploration of teachers' needs in K-12 classrooms that use ITSs [11]. For example, in a generative card sorting exercise, we asked teachers what "superpowers" they would want during ITS class sessions, to help them do their jobs. Several of the superpower ideas that teachers generated centered on abilities to perceive information about individual students' learning and behavior, in real-time. For example, we found that all interviewed teachers wanted to be able to instantly see when a student is "stuck" (even if that student is not raising her/his hand), to instantly detect when a student is off-task or otherwise misusing the software, and to be able to see students' step-by-step reasoning, unfolding in real-time.

During this card sorting exercise, teachers also generated the idea of being able to see this information "floating over students' heads", directly within the physical classroom environment (cf. [29]). In a follow-up series of concept generation and validation [9, 23] studies with teachers, we found that teachers were particularly receptive to awareness tool designs that allowed them to keep their heads up, and their attention focused on the classroom. Teachers emphasized that some of the most useful real-time information comes from reading student body language and other cues that would not be captured by a dashboard alone. They gravitated towards the idea of wearing eyeglasses that could provide them with a *private* view of actionable information about their students in real-time, embedded throughout the classroom environment (e.g., through state indicators displayed directly above students' heads) (cf. [29]). While these "teacher smart glasses" would have many of the same advantages as ambient and distributed classroom awareness tools for teachers, such as [1] and [3], they would not reveal sensitive student data for the whole class to see (cf. [3, 13]) – a risk that several teachers referred to as a "deal-breaker" for use in middle school classrooms [11].

Finally, similar to earlier findings by Martinez-Maldonado et al. in the context of collaborative, multi-tabletop classrooms [16], we found preliminary evidence that teacher awareness of student struggle in ITS classrooms may be limited. Although teachers reported focusing their attention on students whom they thought needed help the most, teacher time allocation during ITS class sessions was not significantly related to either students' prior domain knowledge or learning gains [12]. These findings suggest that there is room for improvement via a real-time support tool.

3 METHODS

Building on the early design findings discussed above, we wanted to get a better sense of what real-time information about student learning and behavior would be most helpful to K-12 teachers during personalized class sessions. In addition to validating teacher desires for real-time support, as uncovered in [11], we wished to understand how teachers would envision *using* such information to inform their real-time decision-making during a class session. We also wanted to explore the idea of "teacher smart glasses" further, to understand their unique affordances.

To these ends, we conducted a series of iterative design studies with a total of 16 middle school math teachers, from 9 schools and 6 school districts in Pittsburgh and surrounding areas. All participating teachers had previously used an adaptive learning technology in their classrooms, and 12 of 16 teachers had previously used an ITS as a regular component of their teaching.

Table 1. Demographic information for schools

School	Region	Free/Reduced Price Lunch (proxy for poverty rate)	# of teach-ers	# teachers with < 2 years' experience
A	Rural	34%	4	0
B	Suburban	23%	1	0
C	Urban	36%	4	1
D	Urban	67%	1	0
E	Urban	63%	2	0
F	Suburban	99%	1	0
G	Suburban	78%	1	1
H	Suburban	71%	1	1
I	Urban	87%	1	1

4 LOWER-FIDELITY PROTOTYPING

4.1 Early prototyping and storyboarding

To further explore teachers needs and desires for real-time support, prior to developing specific prototypes, we first conducted storyboarding and lo-fi prototyping [9] sessions with a series of 3 middle school math teachers from schools A, C, and F. For all studies, two researchers (the 1st and 2nd authors) visited middle schools and worked with teachers in their own classrooms.

In each session, a teacher sat down at her/his desk, in front of a computer screen showing a full-screen image of a classroom full of students working with personalized learning software. A researcher asked the teacher to put on a pair of plastic eyeglass frames, which the teacher was asked to pretend were "smart glasses". As soon as the teacher put on these glasses, a researcher pressed a button on the computer, triggering additional layers of information to appear on top of the image (simulating the experience of using smart glasses). Floating text labels appeared over students' heads, alerting teachers to current detected states, based on requests for "superpowers" reported in [11]. For example, by scanning the classroom, teachers could see that certain students were currently struggling in the software, potentially off-task, or frequently making careless errors. In addition, two class-level analytics displays popped up along the classroom's whiteboard, visible only through the smart glasses – based on teachers' expressed desires for real-time class-level information, identified in [11]. One of these displays showed a list of skills that multiple students in the class had practiced but few had mastered, and the other showed a sorted list of common errors that multiple students in the class had recently exhibited.

The image displayed a single moment during a class session, and the teacher was asked to think aloud while imagining how they might (or might not) act on the information they were seeing through their "smart glasses", if this was a real class session. In the process, teachers were encouraged to remark on any information that was *visible* in the image but not so *useful*, or information that was *not visible* but might be useful to have, to guide their real-time decision-making. For example, although teachers expressed a desire to see when students are frequently making "careless errors" in the "superpowers" exercise [11], all teachers interviewed in the present study were uncomfortable with, and skeptical of the idea that a computer could make accurate judgments about student motivation. In line with prior design findings (e.g., [11, 16]), all teachers expressed a desire to see *positive* information about individual students, not just *negative* information. In particular, all teachers wanted to be able to see when students have been performing particularly well in the software recently. Teachers found this valuable for several reasons, including but not limited to: motivating themselves (since seeing nothing but negative alerts might be discouraging), motivating students (by identifying and praising students who have been doing well lately), and identifying students who may be *underchallenged* by the software.

To facilitate brainstorming, teachers were given a large, printed copy of the same classroom image shown on the screen, but with blank rectangles in place of the individual student labels and classroom analytics displays. Throughout each session, teachers could use these blank spaces to sketch out new ideas for real-time information that could be displayed about individual students or the whole class. Each time a teacher generated an idea for new information, a researcher would press the teacher to provide examples of how she or he envisioned using that information during a real class session. We found that this process of generating hypothetical usage examples often led teachers to refine their ideas, as they realized that more, or different, information may be needed to make certain decisions. The ideas that a teacher generated during one study were ultimately incorporated into the version of the experience prototype [9] that we would show to the next teacher.

At opportune moments throughout each study, researchers also probed teacher reactions to particular classroom scenarios involving the use of smart glasses, using storyboards that were prepared before the study. We took a participatory storyboarding approach [9], in which the final panel or two of a storyboard was often left blank. This allowed teachers to fill in the details of how they would imagine a classroom scenario progressing, or what decisions they might make in a particular scenario, rather than relying entirely on a researcher-envisioned sequence of events.

During the first session, we found that it was challenging for teachers to imagine the actual experience of using mixed-reality smart glasses in the classroom. So, for the second and third prototyping sessions, we transitioned from prototyping purely in Photoshop towards mixing in an experience prototyping phase using real mixed-reality smart glasses (although with Wizard of Oz'd, static analytics). We used the Microsoft HoloLens [10], which enabled us to place readily-available, default assets at fixed spatial positions throughout a teacher's classroom. When teachers then returned to the sketching and storyboarding exercises, they could ground their responses in this experience.

4.2 Iterative mid-fidelity experience prototyping

Given that we had received many positive reactions to the concept of teacher smart glasses in early prototyping and storyboarding sessions – and had begun to get a more detailed sense of teachers' real-time information needs, grounded in the sorts of in-the-moment decisions that this information might inform – we moved to mid-fidelity prototyping sessions. We next conducted prototyping sessions with a series of 5 math teachers, from schools C, E, G, H, and I. As before, for each of these studies, two researchers (the first and second authors) visited middle schools in Pittsburgh and surrounding areas and worked with teachers in their own classrooms. Each study lasted 90 minutes: the teacher wore the HoloLens during an hour-long experience prototyping [9] phase, while experimenting with different configurations of analytics displays and thinking aloud about likely use-cases. This was followed by a 30-minute semi-structured post-interview in which teachers had the opportunity to reflect and provide more detailed design feedback. For these and subsequent studies, we focused on the context of middle school classrooms using tutoring software for equation- solving.

To quickly prototype design alternatives, we used HoloSketch [22], a HoloLens application for rapid prototyping of mixed-reality experiences. Using HoloSketch, we were able to position 2D assets – including mocked-up displays of student-level and class-level analytics – throughout a teacher's classroom space. For example, when a teacher put the HoloLens on, the teacher could see indicator symbols (like those shown in Figure 1) floating over

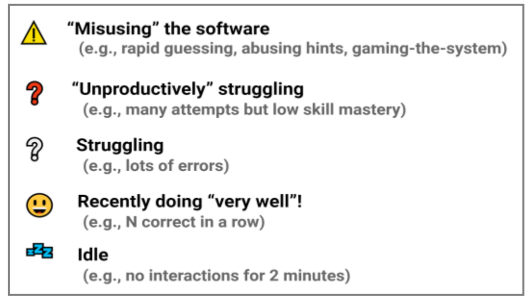

Figure 1. Consistently requested real-time indicators.

empty student seats, and class-level analytics displays appearing as "wall decorations", which the teacher could reposition.

In the first study session, we included all indicator symbols and analytics displays that teachers had consistently requested to this point in early prototyping and storyboarding sessions. Then, in-between prototyping sessions, we rapidly iterated on the design of individual student-level indicators and class-level displays and incorporated new ideas that teachers had generated during the previous session. Since these 2D assets were synchronized with the HoloLens app, we were also able to make modifications in real-time, based on teachers' live design feedback, by editing assets on a laptop as a teacher viewed them through the HoloLens.

At least one instance of each indicator idea that teachers had generated (e.g., a "Zzz" symbol to indicate that a student had been idle for a while) was displayed, positioned over empty student seats, and a set of class-level dashboards were shown at the front of the classroom. As before, these displays were static, presenting a frozen moment in time. Throughout the prototyping sessions, the teacher had the opportunity to move about her/his classroom. The teacher was asked to think-aloud, imagining what *actions* she or he might take in a real class, in response to each indicator, and what other information might help in making these decisions.

In between sessions, we reflected on our areas of highest uncertainty. For each open question, we mocked up several design alternatives. Then, towards the end of each session, teachers were brought to the back of their classroom, where (in mixed-reality) we had arranged an immersive "gallery" of these new design alternatives. Teachers had the opportunity to reposition these information displays, and experiment by decorating their classrooms with different combinations of displays, while thinking aloud and providing design feedback. Based on this feedback, we iterated on the designs prior to the next prototyping session, providing opportunities to validate previous teachers' expressed needs and design ideas.

4.3 Highlighted design findings

A PhD student (the first author of this paper) and two masters students then worked through transcriptions of approximately 12 hours of video and audio recorded experience prototyping studies,

Figure 2. Design mock-ups that emerged during lower-fidelity prototyping sessions. Left: Teacher's default view of the class. Each student has an indicator display floating above her/his head, and two class-level analytics displays are positioned at the front of the class. One shows skills practiced by many students but mastered by few; the other shows errors recently exhibited by many students. Right: Deep-dive screens shown if a teacher 'clicks' on an indicator.

across 8 teachers, to synthesize findings using two standard techniques from Contextual Design: Interpretation Sessions and Affinity Diagramming [9]. Interpretation Sessions are aimed at helping design teams develop a shared understanding of collected interview and think-aloud data, by collaboratively extracting quotes representing key issues and insights. Affinity Diagramming is a widely-used design method, aimed at summarizing patterns across participants' responses, by iteratively clustering these quotes based on content similarity, into a hierarchy of increasingly abstract, emerging themes [9, 11].

We conducted several Interpretation Sessions, and the resulting 655 quotes were iteratively synthesized into 77 level-1 categories, 23 level-2 categories, 10 level-3 categories, and 7 level-4 categories. Key findings (level-4) are highlighted below.

Student-level indicators

Five major categories of student learning states and behaviors emerged from these co-design sessions, as shown in Figure 1. Teachers strongly preferred to keep these indicators simple – displaying a single graphical symbol above each student's head (as in Figure 2, left), to avoid information overload. However, it was important to teachers that they could access brief elaborations on-demand (e.g., by gazing at an indicator, as in Figure 3, left), which could aid in understanding why an indicator was appearing for a student at a particular time.

Sequences of student states can be information-rich

In addition to seeing indicators reflecting a student's current "state" teachers highlighted the usefulness of seeing detected states *preceding* the current state. For example, if a student is currently "idle" or "misusing the software" in some way, it can be useful to know whether that student was *also* recently struggling. Teachers would then interpret the prior struggle as a potential *cause* of the current behavior and respond accordingly.

The classroom as a dashboard

Teachers remarked that it felt natural to reference information displays that were distributed throughout their physical classroom spaces. In the absence of a dashboard, teachers were used to monitoring their students by scanning the physical classroom (e.g., reading student body language), and "patrolling" rows of student seats, to catch glances of students' screens. One teacher remarked, "I would also use their body language to judge the situation, but the initial [alert] would help, so I know to go over there." Teachers also revealed that they *already* used their classrooms as distributed information displays. For example, during a typical class session, teachers would often leave notes and images for themselves on boards or projected displays, to reference throughout the session.

Need for *selective* shared awareness

All participating teachers noted that the analytics they found most useful in informing their real-time decision-making tended to be ones they would not be comfortable sharing with students. Teachers expected that these analytics could do more harm than good, by promoting unhealthy social comparison and competition among students (cf. [1, 13]). As one teacher put it, "In middle school, kids don't know what they don't know [but] kids care *so* much about how they're seen by others … [they] don't want to look stupid or feel stupid." However, teachers also noted they would want a mechanism to *selectively* share particular analytics during class. Five out of eight teachers suggested it would be useful to customize the visibility of analytics displays on a class-by-class basis. All of these teachers anticipated an interaction effect in which real-time analytics might *motivate* higher-achieving classes by promoting competition, but *demotivate* lower-achieving classes.

Ground automated inferences in 'raw' student artifacts

Much of the appeal of the glasses lay in their potential to offload the task of noticing key events in the classroom, via automated inferences. However, teachers also emphasized the importance of having access to "raw" student-generated artifacts in a familiar format. For example, the mock-ups of "deep-dive" screens shown in Figure 2 display individual students' greatest "areas of struggle". For each area, raw examples of errors that the student has recently exhibited are also shown. Showing these example errors is crucial not only in helping the teacher perform further diagnosis, but also in supporting teacher trust (cf. [15]) or enabling the teacher to "override" the system's judgments if needed.

Enable teachers to "peek" at student solution paths

In addition to presenting teachers with summaries of a student's main areas of struggle, teachers generally wanted to be able to see a live feed of a student's work within their current activity (perhaps annotated, as in Figure 2). Although we had anticipated that teachers would prefer to simply walk over to a student and observe that student's screen directly, teachers noted that approaching students can cause them to alter their behavior, reducing the diagnostic usefulness of direct observations (cf. [12]).

"Invisible hand raises"

Although most of teachers' ideas focused on ways real-time analytics could help *them* regulate students' learning, some teachers emphasized the importance of also creating opportunities to develop student help-seeking skills [2], often generating the idea of giving students an "Ask the teacher" button in the software, that would trigger a "raised hand" symbol within the glasses. Teachers expected that, by providing students with a way to request help that was not easily visible to other students, more students would feel comfortable asking (cf. [28, 29]). Otherwise, as one teacher put it, "for a number of students in my class, unless *I* [walk over], they are *never* going to say anything."

5 DEVELOPMENT OF A HIGHER-FIDELITY PROTOTYPE

All prototyping sessions until this point had relied upon Wizard of Oz'ing analytics, presented "frozen" at a single time slice. We next began prototyping the experience of using smart glasses to monitor a class session unfolding over time, using real student data and analytics. Based on findings from lower-fidelity prototyping, we created a mixed-reality application called Lumilo, developed with Unity3D for the Microsoft HoloLens [10], and capable of interfacing with a broad range of ITSs.

Using a newly-extended version of the CTAT/TutorShop architecture for ITS authoring and deployment, we developed an initial set of automated detectors of student learning and behavior, leveraging pre-existing student modeling techniques [6] to provide teachers with each of the key real-time indicators identified in the previous section. When embedded in the tutoring software, the real-time analytics generated by these detectors would then be streamed to the TutorShop learning management system, and finally to Lumilo, where they would update mixed-reality displays in the teacher's glasses. These displays consist of three main types: student-level indicators, student-level "deep-dive" screens, and class-level summaries (as shown in Figures 2 and 3). Student-level indicators and class-level summaries are always visible to the teacher by default – with student-level indicators appearing above corresponding students' heads (based on teacher-configurable seating charts), and with class-level summaries appearing at teacher-configurable locations throughout a classroom. If a teacher gazed at a particular student's indicator, a brief elaboration about the currently displayed indicator symbol would be displayed. For example, if a student was detected as recently struggling in the software, a teacher could glance at that student's indicator to reveal how long this alert had been active, and whether the student seemed to be avoiding using the software's built-in hints. If no indicators were currently active for a student, a circular outline would be displayed above that student's head (as illustrated in Figure 2).

If a teacher clicked on a student's indicator (either by using a small handheld clicker, or by making a tapping gesture in mid-air), the teacher would see "deep-dive" screens for that student, containing more detailed information about a student's path through their current problem, and any consistent areas of struggle that student might be exhibiting. The "current problem" deep-dive screen illustrated in Figure 2 displays an annotated live feed of a student's work on their current problem. Each problem step is annotated with the number of hint requests (in yellow) and incorrect attempts (in red) that a student had made on that step. The deep-dive screens also allowed teachers to view *recently active* alerts, as shown beside the student's name in Figure 3.

To support future design explorations, we engineered the initial prototype of Lumilo in a heavily-modular fashion, so as to be able to rapidly iterate on the design in-between future prototyping sessions, and even to make small adjustments *within* a single prototyping session, based on live teacher feedback. For example, alternative detector algorithms intended to measure the same teacher-identified construct (such as "unproductive struggle") could be swapped in and out during a session, and thus tested in parallel. All detectors included in our initial prototyping sessions were drawn from the Educational Data Mining, Artificial Intelligence in Education, and Learning Analytics literatures – where many automated detectors of student learning and behavior have been introduced, based on students' interactions within the software (e.g., [2, 4, 6, 14]). For example, in order to drive a real-time indicator of "system misuse", we explored combinations of the Help Model [2] and a detector of unproductive "gaming-the-system" behaviors [6]. Similarly, to drive an indicator of "unproductive struggle", we explored the use of simpler methods such as Beck and Gong's detector of "wheel-spinning" [4], as well as more sophisticated methods (e.g., "predictive stability" [14]). Each detector was implemented in a parameterized fashion, so that aspects of their behavior (e.g., alert thresholds) could be adjusted during and between sessions.

We also developed a new logging library for Lumilo, which automatically logs teacher actions during class sessions to DataShop, a major educational data repository [16]. For example, Lumilo can record time-stamped logs of a teacher's physical

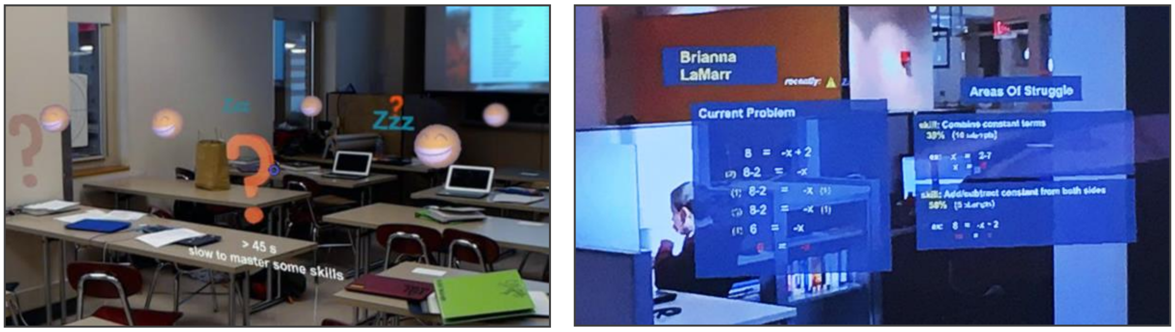

Figure 3. Point-of-view screenshots of teachers using Lumilo. LEFT: Teacher's view of student indicators, immediately following a pilot study in a live classroom. RIGHT: Teacher's view of deep-dive screens.

proximity to a given student at a given time, the target of a teacher's gaze, and all teacher interactions with the tool.

6 PROTOTYPING REAL-TIME ANALYTICS USING REPLAY ENACTMENTS

6.1 Methods

We next conducted a series of higher fidelity, iterative experience prototyping sessions, with a total of 10 math teachers, from 5 schools (schools A, B, C, D, and E) and 3 school districts in Pittsburgh and surrounding areas. As before, all participating teachers had previously used at least one adaptive educational technology in their classrooms, and 7 out of 10 teachers had previously used an ITS as a regular component of their teaching.

6.2 Replay Enactments

In order to prototype the experience of using Lumilo in a classroom, we developed a new prototyping method for real-time teacher support tools: Replay Enactments. Much like other recently proposed prototyping methods in Learning Analytics, such as the simulation methods presented in [19, 20] and [21], Replay Enactments (REs) involve replaying log data from students' interactions within educational software, to prototype real-time analytics and visualizations with teachers. However, in the spirit of recent HCI methods for prototyping radically new experiences (e.g., User Enactments (UEs) [23]), REs build on these approaches by emphasizing *embodied role-playing* in physical classroom spaces. In our own work, we have found that pushing teachers to role-play what they might actually say to a particular student at a given time often leads teachers to insights about ways an orchestration tool could be improved. In addition, we have found that asking teachers to role-play while actually navigating throughout a physical classroom space helps to create an illusion of "actually being there" while also providing early insight into potential effects of classroom layout (cf. [12]). In contrast to UEs, REs prototype an experience using authentic data and algorithms, evolving over time. Doing so allows for earlier observation of the interplay between human and machine judgments (cf. [7]), such as ways a system's false positives and negatives may impact the *experience* of using a data-driven intelligent system.

In each Replay Enactment, we brought teachers into a computer lab at our university's campus. At each empty seat in the lab, we had placed a name tag with a fabricated student name. On the corresponding computer screen, we had logged into the tutoring software, under the given student's name. Using Lumilo, we had positioned holograms throughout the computer lab so that indicators, associated with corresponding student accounts in the software, would appear over "student" heads. Class-level analytics displays were also positioned along the walls of the computer lab.

Using a newly-developed log replay system, we were able to replay log data from an entire classroom of students, using datasets collected from a classroom study in which middle school students used Lynnette, an ITS for equation solving. When a researcher pressed a button in a web-based "controller" interface, the entire class sprung to life, replaying a 40-minute class session from beginning to end, at actual speed. The teacher wore Lumilo during this simulation phase and was asked to pretend that this was an actual class session, and to think aloud while moving throughout the room. If the teacher thought they might focus attention on a particular student at a particular time, the teacher was asked to talk to the student as if they were actually there. Teachers often became quite immersed in this task. One teacher remarked, about halfway through a session, "You know what? I'm acting like they're really here now … I'm thinking that I'm gonna tell them something, and [the indicator] is gonna change."

We ran separate REs with a total of 5 teachers. Each of these sessions began with a 35-minute training and familiarization phase during which the teacher could acclimate to using the system, followed by a 40-minute simulation phase, and concluding with a 15-minute post-interview, to elicit additional design feedback. To prototype the experience of using Lumilo under a broad range of classroom dynamics, we selected one dataset from a "remedial" middle school math class, one dataset from an "advanced" class, and one dataset from an "average" class (based on the tiering system used in the schools from which these datasets were drawn). We then randomly assigned datasets to study sessions, so that the remedial and average classes were simulated for two teachers each, and the advanced class was simulated for one teacher. To account for potential influences of classroom layout, different computer labs – with a range of spatial layouts – were used across sessions.

During these REs, we elicited teacher feedback not only on Lumilo's interactions and the visual presentation of analytics, but also on the particular choices of analytics used to drive Lumilo's real-time indicators. During the training and familiarization phase of each session, teachers were provided with definitions for each indicator symbol, which included brief summaries of a detector's structure, the main features it relies upon, and the default values of any free parameters (such as alert thresholds). Within the simulation phase of each session, teachers frequently monitored students' raw activity in the software – either by approaching a computer terminal, or by opening that student's deep-dive. In doing so, they observed ways in which particular detectors may have been over- or under-sensitive (or were perhaps overlooking key features of student thinking and behavior entirely).

In-between RE sessions, and sometimes within a single RE, we would often iterate on the detectors and alert policies driving the real-time indicators, based on teachers' feedback. For example, over time, the definition of the "struggling" indicator evolved to include not only a threshold on a student's recent error rate, but also automated detection of student *hint avoidance* [6], as well as whether a student had been making good use of the software's hints, yet remained "stuck" – with the corresponding "question mark" symbol glowing gradually brighter, the longer the student was stuck. By the final two sessions, teacher observations of over- or under-sensitivity, or semantic mismatches, had become rare.

Other design features that entered the prototype during this iterative process, based on teachers' feedback, included the ability to set visual "timers" on an individual student by clicking-and-holding on the student's indicator. Teachers found this useful as a reminder to check back with a student – for example, if that student appears to be struggling *currently*, but it is unclear to the teacher whether the student might overcome this struggle on their own within the next several minutes. In addition, we found that teachers saw great value in the ability to monitor individual students' activities, while either walking or physically attending to a student seated across the classroom. As such, we enhanced Lumilo so that a teacher could have the "deep-dive" screen "tag along" with them as they walked (instead of hanging in space near the given student and visible only when looking in that direction). Finally, to give teachers' "eyes in the back of their heads" (cf. [11]), we enabled teachers to configure ambient, spatial sound notifications. For example, if a student was misusing the software, a teacher could privately perceive a soft notification, as if it were emanating from that student's location in the classroom.

6.3 Evaluating effects of Lumilo on teacher attention allocation

Prior to piloting Lumilo in live K-12 classrooms, we wanted to better understand its potential effects on teacher behavior. In particular, we wanted to investigate whether and how Lumilo might influence teacher time allocation (cf. [16]) across students of varying *prior domain knowledge* and *learning rates* within the software, compared with business-as-usual (i.e., without an orchestration tool). We ran an additional series of 6 Replay Enactments, across which Lumilo's design was held constant. For each session, replay data from a 40-minute class session was used. The replay data was randomly selected from a pool of 5 "average" and "remedial" classes. An "average" class was replayed in 4 sessions, and a "remedial" class was replayed in the remaining two. Advanced classes were omitted from the selection pool, given little between-student variance in test scores. To minimize potential effects of names or seating positions, replayed students were randomly assigned names and positions in each session.

Table 2. Relationships between teacher time allocation (in seconds) and students' prior knowledge and learning. * $p < 0.05$, ** $p< 0.01$, *** $p < 0.001$

class type	using Lumilo?	pretest	post | pre	sample size (teachers, classrooms, avg. class size)
Live	**No**	6.29	-5.49	(4, 7, 16)
RE	**Yes**	**- 4.66***	**- 21.19****	(6, 3, 15)
Live	**Yes**	**-73.75*****	5.34	(1, 1, 15)

In Lumilo, the indicators positioned above students' heads double as proximity sensors within a physical space. Using these mixed-reality sensors, a teacher's allocation of time to a given student was measured as the *cumulative* time (in seconds) that she or he spent within a 4 ft radius of that student. If a teacher was within range of multiple students, time was accumulated only for the *nearest* student. We used hierarchical linear modeling (HLM) to predict teachers' time allocation across replayed "students" as a function of either students' prior domain knowledge (measured by a pretest in the original class session) or students' learning during the class (measured by a posttest, controlling for pretest). As is the case in a typical classroom study, teachers did not have access to pre- or post-test data, and this data was not used by Lumilo. Using 2-level models, with students nested in classrooms, provided a better fit than 1-level or more complex models. Standardized coefficients for student-level variables are provided in row 2 of Table 2. As shown, teachers using Lumilo in REs spent significantly more time attending to "students" with relatively lower pretest scores, or posttest scores (controlling for pretest).

By contrast, in an in-vivo classroom study that we previously ran with 4 teachers across 7 real middle school classrooms, students worked with Lynnette while teachers monitored and helped their students (without access to a real-time awareness tool). Performing the same analysis as above, but this time with data from this classroom study (with time allocation recorded via manual classroom coding), we again found that 2-level models provided the best fit. Coefficients for these models are provided in Table 2 (row 1). Although all teachers reported attempting to devote most of their time to students whom they expected would struggle most with the material, we found no significant relationships between students' pre- or posttest scores and teacher time allocation. We take this contrast as evidence that Lumilo may aid teachers in focusing on and helping students with lower prior knowledge. Early results from a 1-hour pilot study, with one teacher using Lumilo in a real classroom (Table 2, row 3) were consistent with this finding.

More importantly, we take these results as preliminary evidence that Lumilo may successfully aid teachers in identifying those students who *would have* gone on to exhibit the lowest *learning* in a real classroom session – potentially representing a subset of students who benefit the least from working with the tutoring software alone, and who may stand to benefit the most from a teacher's help. Since Replay Enactments remove the possibility of a causal arrow from teacher behavior to students' learning within the software, this method enables us to investigate counterfactuals such as the above, for different forms of teacher augmentation. Conversely, classroom studies – although costly to run – allow investigation of effects of a tool in the context of many competing influences on a teacher's attention and judgment.

6.4 Highlighted design findings

As before, we conducted Interpretation Sessions and Affinity Diagramming, based on approximately 18.5 hours of audio/video recorded think-aloud data and design feedback from Replay Enactments with 10 teachers, along with design feedback collected immediately following the classroom pilot. The resulting 486 quotes were iteratively synthesized into 43 level-1 categories, 26 level-2 categories, 13 level-3 categories, and 5 level-4 categories. Key findings from this synthesis (level-4 categories) are highlighted below, representing directions for future work:

Value of continuous, real-time feedback on *instruction*.
Although Lumilo did not provide direct feedback to teachers about potential *effects* of their instructional interventions on student learning and behavior, teachers frequently inferred causality by monitoring changes in student and class state, following an intervention. In fact, teachers were often tempted to do so even during Replay Enactments, in which no students were actually present. In the middle of one Replay Enactment, a teacher remarked: "You know what? I'm acting like [the students] are really here now. … I'm thinking that I'm gonna tell them something and [the indicator] is gonna change". Teachers emphasized that receiving more direct, live feedback about the effects of their instructional interventions could help them adjust their instruction on-the-spot, and perhaps even improve over time (particularly if this feedback were *constructive*).

When many students need help, on different topics, *at the same time*, choice can be anxiety-inducing.
During Replay Enactments, teachers realized that when they were made more aware of student struggle, they became more aware of their limited ability to *actually help* all of their students. The main way teachers proposed addressing this was through dynamically adjustable alert thresholds, which could help them better focus their attention during times when many students may need help at once, or in otherwise chaotic classroom environments. As one teacher put it, "I'm going to be able to handle different [numbers of alerts] in different classes … I'd want to be able to control that."

Action recommendations *in addition to* awareness support.
As we moved to higher-fidelity prototyping, teachers consistently noted that it would be helpful to have explicit recommendations from the system, to help prioritize among students and/or to decide how best to help a student. For example, one teacher suggested that it would sometimes be helpful to receive recommendations for "conversation starters," (e.g., to help a teacher avoid providing "too much" scaffolding). It is clear from earlier design explorations, however, that such a system would require careful design, to respect teachers' autonomy [11].

Automated support for dynamic, adaptive peer-matching.
Teachers noted that it would be useful to receive live support from Lumilo in adaptively assigning students to serve as *peer tutors* for others, throughout the course of a class session (cf. [8, 24, 28]). Such dynamic peer-matching would enable teachers to offload some help sessions to students: "I would let them go, for a while, so I could focus my attention elsewhere in the room." In turn, teachers envisioned devoting time to students who might benefit more from a teacher's assistance than from peer tutoring.

Trade-offs between *interpretability* and *accuracy.*
Although teachers had expressed a preference for simpler, more interpretable analytics in lower-fidelity prototyping sessions, it became clear during higher-fidelity prototyping sessions that the strength of this preference was heavily dependent on the underlying construct that an indicator purported to measure. For example, when it came to detection of "system misuse" it was important to teachers that they could easily understand (and explain to students) precisely the patterns of behavior that had led to this classification. By contrast, teachers were more open to the use of "black box" algorithms for detecting potential "unproductive struggle" if this meant alerting them to these students *earlier* (after which, teachers could apply their own judgment, using other available information).

7 DISCUSSION AND FUTURE WORK

In this paper, we report on the iterative co-design, development, and evaluation of Lumilo: mixed-reality smart glasses that augment K-12 teachers' real-time perceptions of their students' learning, metacognition, and behavior as students work with personalized learning software. While prior work has explored the use of smart glasses to facilitate live instructor feedback in university lecture contexts [29], the current work represents the first exploration of the affordances of smart glasses to support teachers in orchestrating personalized, self-paced classroom sessions. Longer term, we envision generalizing wearable cognitive augmentation such as Lumilo to a broader range of personalized learning environments. Advances in multimodal learning analytics may reduce Lumilo's dependence on data streams from educational software, particularly if combined with advances in student modeling [6]. Furthermore, it may soon become possible to implement Lumilo within a lighter-weight pair of spatially aware smart glasses than the HoloLens [5].

We also present Replay Enactments (REs): a new prototyping method for real-time orchestration tools, combining prior methods from Learning Analytics [19, 20, 21] and Human-Computer Interaction [23]. REs facilitate investigation into the effects of particular tool designs, prior to deploying these in live classrooms, by immersing teachers in embodied simulation exercises using previously collected interaction logs from real students. For example, in our studies, REs provided early insight into potential *affective* consequences of enhanced awareness. We present results from a classroom study, revealing that, although K-12 teachers report focusing their attention on students whom they expect need help most, teacher time allocation (in the absence of a support tool) was not significantly related to students' prior knowledge or learning (cf. [12, 16]). Through REs we find that in simulated classrooms, Lumilo directs teachers' time towards students who would otherwise exhibit lower learning within the ITS. A single-session pilot provides early evidence that, in a live classroom, Lumilo may have the same effect. These students may represent those learners for whom the support provided by the software alone is least effective, and who would benefit most from a teacher's help.

Finally, our findings provide novel insights into teachers' needs and desires for real-time support in K-12 classrooms using ITSs, many of which we expect will generalize to a broader range of personalized learning environments. In the next phase of this work, we will continue to pilot Lumilo with teachers and students in live middle school classrooms. As we move into live classrooms we plan to increasingly inform the design of Lumilo with insights from classroom usage data. There is a scarcity of empirical knowledge about the effects real-time analytics might have on teacher-student interactions, and ultimately, on student learning [12, 27]. Even less is known about the effects particular *design features* of orchestration tools may have [27]. In upcoming studies, we plan to use Lumilo as a research tool, towards understanding how human and AI instruction can best be *combined,* to achieve outcomes greater than either can achieve alone.

ACKNOWLEDGMENTS

This work was supported by NSF Award #1530726, and by IES Grant R305B150008 to CMU. Opinions do not represent the views of NSF, IES or the U.S. ED. Special thanks to Jonathan Sewall, Octav Popescu, Peter Schaldenbrand, Zac Yu, Mary Beth Kery, Jodi Forlizzi, Lu Chen, and all participating students and teachers.

REFERENCES

[1] Alavi, H.S., Dillenbourg, P. and Kaplan, F. 2009. Distributed Awareness for Class Orchestration. *EC-TEL* (2009), 211–225.
[2] Aleven, V., Roll, I., McLaren, B.M. and Koedinger, K.R. 2016. Help Helps, but only so Much: Research on Help Seeking with Intelligent Tutoring Systems. *IJAIED*. 26, 1 (2016), 205–223.
[3] Alphen, E. Van 2016. Lernanto : Using an Ambient Display During Differentiated Instruction. *CHI*. (2016), 2334–2340.
[4] Beck, J.E. and Gong, Y. 2013. Wheel-spinning: Students who fail to master a skill. *AIED* (2013), 431–440.
[5] Border, J., Haddick, J. and Osterhout, R. 2016. See-through near-eye display glasses with a light transmissive wedge shaped illumination system. US20160187654 A1. 2016.
[6] Desmarais, M.C. and Baker, R.S.J.D. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *UMUAI*. 22, (2012), 9–38.
[7] Dove, G., Halskov, K., Forlizzi, J. and Zimmerman, J. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. *CHI* (2017), 278–288.
[8] Diana, N., Eagle, M., Stamper, J., Grover, S., Bienkowski, M., & Basu, S. (2017). Automatic Peer Tutor Matching: Data-Driven Methods to Enable New Opportunities for Help. In *EDM* (2017), 372-373.
[9] Hanington, B.M. and Martin, B. 2012. *Universal methods of design : 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers.
[10] HoloLens - Microsoft: 2017. *https://www.microsoft.com/en-us/hololens*.
[11] Holstein, K., McLaren, B.M. and Aleven, V. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. *LAK* (2017), 257–266.
[12] Holstein, K., McLaren, B.M. and Aleven, V. 2017. SPACLE: investigating learning across virtual and physical spaces using spatial replays. *LAK* (2017), 358–367.
[13] Jivet, I., Scheffel, M., Drachsler, H. and Specht, M. 2017. Awareness Is Not Enough: Pitfalls of Learning Analytics Dashboards in the Educational Practice. *EC-TEL* (2017), 82–96.
[14] Käser, T. and Gross, M. 2016. When to stop ? - Towards Universal Instructional Policies. *LAK* (2016), 289–298.
[15] Kay, J. (2000). Stereotypes, student models and scrutability. In *ITS* (2000), 19-30.
[16] Koedinger, K., Liu, R., Stamper, J., Thille, C. and Pavlik, P. 2017. Community based educational data repositories and analysis tools. *LAK* (2017), 524–525.
[17] Kulik, J.A. and Fletcher, J.D. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *RER*. 86, 1 (2016), 42–78.
[18] Martinez-Maldonado, R., Clayphan, A., Yacef, K. and Kay, J. 2015. MTFeedback: Providing Notifications to Enhance Teacher Awareness of Small Group Work in the Classroom. *IEEE TLT*. 8, 2 (2015), 187–200.
[19] Martinez Maldonado, R., Kay, J., Yacef, K., & Schwendimann, B. (2012). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In *ITS* (2012), 482-492.
[20] Martinez-Maldonado, R., Pardo, A., Mirriahi, N., Yacef, K., Kay, J. and Clayphan, A. 2016. LATUX: An iterative workflow for designing, validating and deploying learning analytics visualisations. *JLA*. 2, 3 (2016), 9–39.
[21] Mavrikis, M., Gutierrez-Santos, S. and Poulovassilis, A. 2016. Design and Evaluation of Teacher Assistance Tools for Exploratory Learning Environments. *LAK* (2016), 168–172.
[22] Mixed Reality Design Labs: 2017. *https://github.com/Microsoft/MRDesignLabs/*.
[23] Odom, W., Zimmerman, J., Davidoff, S., Forlizzi, J., Dey, A. and Lee, M.K. 2012. A Fieldwork of the Future with User Enactments. *DIS* (2012), 338–347.
[24] Olsen, J. (2017). Orchestrating Combined Collaborative and Individual Learning in the Classroom (Doctoral dissertation).
[25] Pane, J., Steiner, E., Baird, M., Hamilton, L. and Pane, J. 2017. *Informing Progress: Insights on Personalized Learning Implementation and Effects*. RAND..
[26] Quintana, R., Quintana, C., Madeira, C. and Slotta, J.D. 2016. Keeping Watch: Exploring Wearable Technology Designs for K-12 Teachers. *CHI* (2016), 2272–2278.
[27] Rodríguez-Triana, M.J., Prieto, L.P., Vozniuk, A., Boroujeni, M.S., Schwendimann, B.A., Holzer, A. and Gillet, D. 2017. Monitoring, Awareness and Reflection in Blended Technology Enhanced Learning: A Systematic Review. *IJTEL*. 9, 2–3 (2017), 126–150.
[28] Schofield, J.W., Eurich-Fulcer, R. and Britt, C.L. 1994. Teachers, Computer Tutors, and Teaching: The Artificially Intelligent Tutor as an Agent for Classroom Change. *AERJ*. 31, 3 (1994), 579–607.
[29] Zarraonandia, T., Aedo, I., Díaz, P. and Montero, A. 2013. An augmented lecture feedback system to support learner and teacher communication. *BJET*. 44, 4 (2013), 616–628.