JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 6(2), 27–52. http://dx.doi.org/10.18608/jla.2019.62.3

Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity

Kenneth Holstein,[1] Bruce M. McLaren,[2] and Vincent Aleven[3]

Abstract

Involving stakeholders throughout the creation of new educational technologies can help ensure their usefulness and usability in real-world contexts. However, given the complexity of learning analytics (LA) systems, it can be challenging to meaningfully involve non-technical stakeholders throughout their design and development. This article presents a detailed case study of the iterative co-design of Lumilo, a wearable, real-time learning analytics tool for teachers working in AI-enhanced K–12 classrooms. In the process, we argue that the co-design of LA systems requires *new kinds of prototyping methods*. We introduce one of our own prototyping methods, REs, to address unique challenges of co-prototyping data-driven algorithmic systems such as LA tools. This work presents the first end-to-end demonstration in the literature of how non-technical stakeholders can participate throughout the whole design process for a complex LA system — from early generative phases to the selection and tuning of analytics to evaluation in real-world contexts. We conclude with a summary of methodological recommendations for future LA co-design efforts.

Notes for Practice

- The field of Learning Analytics (LA) is beginning to explore new methods and strategies for co-designing LA tools with critical stakeholders such as teachers, students, and parents. Effectively implementing LA co-design processes requires drawing upon design and prototyping methods from a range of disciplines, and in some cases creating new ones. However, demonstrations of successful co-design processes for LA tools remain very rare in the literature.
- This article provides an end-to-end demonstration and methodological recommendations for how non-technical stakeholders can meaningfully participate throughout the design of a complex LA tool. For example, when implementing co-design processes, designers of LA tools should centre initial discussions on *stakeholder needs*, rather than specific analytics, visualizations, or other technical considerations. In addition, to ensure the usefulness and usability of the resulting designs, designers should scaffold stakeholders in reflecting on how particular LA displays might inform (or fail to inform) instructional *decision-making* and *action* in the context of specific tasks and scenarios.
- The present case study illustrates the importance of carefully considering which roles to augment, which to automate, and which to leave alone, when designing LA tools for use in social contexts such as classrooms. Working closely with key stakeholders such as teachers can help designers to better understand their values, and nuances of the contexts in which LA technologies will be used. In turn, this can help in understanding where automation or augmentation via LA can help more than hurt.

Keywords

Real-time analytics, co-design, orchestration, prototyping, K–12, classrooms, AI, automation, augmentation

Submitted: 31.10.2018 — **Accepted:** 10.05.2019 — **Published:** 07.08.2019

[1] *Corresponding Author Email: kjholste@cs.cmu.edu Address: Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA.*

[2] *Email: bmclaren@cs.cmu.edu Address: Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA.*

[3] *Email: aleven@cs.cmu.edu Address: Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA.*

1. Introduction

Actively involving educational practitioners and other stakeholders throughout the design of educational technologies can help ensure their alignment with real-world needs, values, and constraints, and their usefulness and usability in actual educational contexts (Bonsignore, DiSalvo, DiSalvo, & Yip, 2017; Prieto-Alvarez, Martinez-Maldonado, & Anderson, 2018; Schuler & Namioka, 1993). Yet given the complexity of learning analytics (LA) systems, it can be challenging for stakeholders to meaningfully contribute to their design (cf. Martinez-Maldonado et al., 2016; Mavrikis, Gutierrez-Santos, Geraniou, Noss, & Poulovassilis, 2013; Prieto-Alvarez et al., 2018). For example, teachers and students may have limited data literacy or limited understanding of the potentials of state-of-the-art data gathering, processing, or visualization techniques (Martinez-Maldonado et al., 2016; Mavrikis et al., 2013). Effectively co-designing LA systems with practitioners thus requires generative design techniques and strategies for eliciting end-user desires, needs, values, and constraints, untethered by specific pre-existing technologies or users' understandings of what is currently possible. In addition, meaningfully involving non-technical stakeholders throughout the whole design process requires prototyping methods that can aid them in understanding the consequences of particular design choices in complex LA systems (e.g., how particular analytic measures, coupled with specific parameter settings and visualizations, might behave when used for different student populations), so that they can provide informed design feedback.

In this article, we present a case study of the iterative co-design, development, and evaluation of a real-time learning analytics tool for K–12 teachers who use AI learning technologies in the classroom — culminating in Lumilo, a set of mixed-reality smart glasses that tune teachers in to the rich analytics generated by these systems. This work provides the first end-to-end demonstration in the literature of how teachers can meaningfully participate at every stage of the design of a complex LA system, from initial need-finding and generative design activities to the selection and tuning of analytic measures to the evaluation of resulting technologies in live K–12 classrooms.1 In addition, this work presents the first broad investigation in the literature of teachers' needs for real-time analytics (i.e., one that is not initially tied to an existing prototype or constrained by immediate technical feasibility, such as the current availability of data or analytic measures), within or outside the context of AI-enhanced classrooms (Holstein, McLaren, & Aleven, 2017b; Rodriguez-Triana et al., 2017). While we investigate teacher needs in the context of AI-enhanced K–12 classrooms, we expect that many of our design findings reflect broader needs for classroom orchestration support, and may thus generalize to other learning contexts.

The structure of this article is as follows: we first provide a brief overview of prior work (Section 2) on the co-design of learning analytics systems, as well as the use of AI tutors to support teachers in the classroom. Following a high-level overview of our overall design process (Section 3), we then report on our initial, generative design studies with K–12 teachers (Section 4) and our iterative prototyping studies (Section 5), presenting key findings at each stage of the process. As part of this work, we introduce a novel prototyping method for data-driven algorithmic systems (Replay Enactments) and report on the use of Replay Enactments to anticipate the impacts of specific choices of analytic measures and visualizations on teachers' user experience (section 5.6) and behaviour (Section 5.7) prior to entering live classrooms. Finally, we present observations from Lumilo's use in live middle-school math classrooms (Section 6), and close with a discussion of our overall design process, providing recommendations for future LA co-design efforts and highlighting opportunities for future research (Section 7).

2. Background and Related Work

2.1. Co-Design of Learning Analytics Systems

Participatory or co-design approaches have a long history in the learning sciences, where researchers and designers have worked closely with teachers and/or students as active collaborators in the design of formative assessment tools, adaptive learning technologies, and educational simulations, among many other applications (e.g., Black & Harrison, 2001; DiSalvo & DiSalvo, 2014; Hoadley, 2017; Penuel, Roschelle, & Shechtman, 2007; Penuel & Yarnall, 2005; Shepard, 1997). In addition to helping designers gain deeper, more detailed understandings of stakeholder needs and values, co-design approaches can lead to unexpected technological innovations and also empower participants — giving them a voice in shaping technologies that impact their lives, and increasing the likelihood of technology acceptance and adoption (Bonsignore et al., 2017; Hoadley, 2017).

Despite these benefits, data-driven algorithmic systems such as LA tools introduce many new challenges for co-design. Helping non-technical stakeholders meaningfully participate in shaping *algorithmic elements* of such systems has been recognized as a central open challenge in the area of Human–Computer Interaction (HCI; Baumer, 2017; Dennerlein, Kowald,

[1] Some studies included in this article have previously appeared in conference publications (Holstein, Hong, Tegene, McLaren, & Aleven, 2018; Holstein, McLaren, & Aleven, 2017b). The current article expands substantially on these prior publications by contributing the following: 1) an expanded discussion of the design process and methods used; 2) a discussion of findings from *classroom deployment* of the resulting system, highlighting needs and nuances that emerged only once the tool was used "in the wild"; 3) methodological reflections and recommendations for future LA co-design efforts; and 4) an expanded discussion of the Replay Enactments (REs) prototyping method, which was briefly introduced in our prior publications.

Pammer-Schindler, Lex, & Ley, 2018; Lee et al., 2018; Prieto-Alvarez et al., 2018; Zhu, Yu, Halfaker, & Terveen, 2018). Although recently proposed design models in the field of Learning Analytics (LA), such as the LATUX workflow (Martinez-Maldonado et al., 2016), encourage the active involvement of stakeholders at every stage of the design process, detailed case studies of actual co-design processes in LA remain rare (but see Prieto-Alvarez et al., 2018). Furthermore, when stakeholders are involved in the LA design process, they tend to be brought in at relatively late stages — for example, to provide feedback on existing prototypes of LA tools. Yet by this point in the design process, many fundamental design decisions have already been made, such as the pedagogical goals that a design is intended to address (Rodriguez-Triana et al., 2017; Prieto-Alvarez et al., 2018).

Martinez-Maldonado et al. (2016) and Mavrikis, Gutierrez-Santos, and Poulovassilis (2013) provide case studies of LA design processes that involve frequent consultations and user testing with classroom teachers, providing rich insights into how teachers can be involved at multiple stages of the design process. In addition, recent work by Dollinger and Lodge (2018) and Prieto-Alvarez et al. (2018) has begun to synthesize high-level strategies and frameworks for involving students throughout the entirety of the design process for LA tools. However, LA design workflows and frameworks alone provide limited methodological guidance for researchers and practitioners who wish to employ co-design approaches in their particular contexts, and many open questions remain about how to involve non-technical stakeholders most effectively throughout the entire LA design process.

To the best of our knowledge, the present case study provides the first end-to-end demonstration in the literature of how teachers can meaningfully participate throughout the design process for a complex LA tool — beginning with broad needfinding and generative design phases, and concluding with piloting and evaluation in real-world learning contexts. It is our hope that, in combination with emerging frameworks for LA co-design, this detailed case study will help to inform future efforts to deeply involve stakeholders in the design of learning analytics tools and other data-driven algorithmic systems.

2.2. AI Tutors as Teachers' Aides

Intelligent tutoring systems (ITSs) are a class of AI 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; Desmarais & Baker, 2012). Several meta-reviews have indicated that ITSs can enhance student learning, compared with other learning technologies or traditional classroom instruction (e.g., Kulik & Fletcher, 2016). However, ethnographic studies have revealed that, in K–12 classroom settings, teachers and students often use ITSs in ways not originally anticipated by system designers (e.g., Holstein, McLaren, & Aleven, 2017a; Ogan et al., 2012; 2015; Schofield, Eurich-Fulcer, & Britt, 1994). For example, Schofield et al. (1994) 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 in the Schofield et al. study tended to perceive that teachers provide better one-on-one help than an ITS, they also preferred ITS class sessions to more traditional class sessions — in part because of this increase in one-on-one teacher–student interactions.

Despite these benefits, modern ITSs have also been shown to have various limitations (e.g., Beck & Gong, 2013; Kai, Almeda, Baker, Heffernan, & Heffernan, 2018; Käser, Klingler, & Gross, 2016; Ogan et al., 2012; 2015). A rich strand of literature in human factors engineering has studied problems of "function allocation" between humans and machines in contexts where automation is helpful yet imperfect (e.g., Wright, Dearden, & Fields, 2000; Sujan & Pasquini, 1998; Wickens, Gordon, Liu, & Lee, 1998). Yet the question of how best to combine strengths of human and automated instruction has received relatively little attention within the learning analytics and AI in education literatures thus far. Over a decade ago, Yacef (2002) proposed a reframing of intelligent tutoring systems as "intelligent teaching assistants" (ITAs): systems designed with the joint objectives of helping human teachers teach and helping students learn, rather than only the latter of these objectives as is typical of ITSs. In line with the literature on function allocation in human–machine systems, other researchers have since proposed optimizing student learning by leveraging complementary strengths of human and AI instruction (e.g., Baker, 2016; Ritter, Yudelson, Fancsali, & Berman, 2016). That is, ITSs might be more effective if they could adaptively enlist the help of human teachers (c.f. Kamar, 2016), in situations that teachers may be better suited to handle. While there has been some work on real-time teacher support tools for ITS classrooms since the vision of ITAs was introduced, little work has explored teachers' actual needs and desires for such support, or how human instruction might most effectively be combined with AI instruction.

3. Overview of Methods

Throughout our design process, we employed a broad range of design research methods. At a high-level, our design process followed the LATUX workflow for designing and deploying LA awareness tools (Martinez-Maldonado et al., 2016). As described in the following sections, choices of methods at each phase of our iterative design process were made adaptively, based on our research/design team's areas of greatest uncertainty at a given stage of the process. A brief overview of the overall design process, illustrating major design phases and examples of methods used at each phase, is presented below. Each phase is discussed in detail in Sections 4 through 6. To inform future LA co-design efforts, we then summarize "lessons learned" in the form of broad methodological recommendations in Section 7.

1. **Initial needs analysis and concept generation**, including design activities such as generative card sorting exercises (Cairns & Cox, 2008), directed storytelling (Evenson, 2006), semi-structured interviews with teachers, and field observations (Hanington & Martin, 2012).

2. **Initial concept validation** via speed dating sessions with teachers (Davidoff, Lee, Dey, & Zimmerman, 2007).

3. **Iterative lower-fidelity prototyping**, gradually increasing the fidelity of both prototypes and simulated use contexts, using methods such as experience prototyping (Buchenau & Suri, 2000), role-playing and bodystorming exercises (Oulasvirta, Kurvinen, & Kankainen, 2003), participatory sketching and comicboarding (Halloran et al., 2006; Moraveji, Li, Ding, O'Kelley, & Woolf, 2007; Tohidi, Buxton, Baecker, & Sellen, 2006), and behavioural mapping (Hanington & Martin, 2012; Veitch, Salmon, & Ball, 2007).

4. **Iterative higher-fidelity prototyping** with replay-based simulation exercises, using Replay Enactments (REs) (Holstein, Hong, Tegene, McLaren, & Aleven, 2018): a novel prototyping method for LA tools and other data-driven algorithmic systems.

5. **Iterative classroom piloting and experimental evaluation**, using field observations, pre-post assessments of student learning, semi-structured interviews with teachers and students, and behavioural mapping.

4. Investigating Teachers' Needs and Desires for Real-Time Analytics

To first gain a better sense of teachers' challenges and needs for support in AI-enhanced classrooms, we conducted a series of formative design studies with a total of 10 middle-school math teachers, across five schools.[2] All participating teachers had previously used some form of adaptive learning software in the classroom, and all but one had previously used an ITS as part of their classroom instruction.

4.1. "Teacher Superpowers" as a Probe to Investigate Perceived Challenges

To encourage teachers to talk freely about challenges they face in AI-enhanced classrooms, without feeling constrained to those for which they believed a technical solution was currently possible, we initially avoided asking direct questions about "learning analytics." Instead, we developed a new probe: in a series of one-on-one study sessions with five teachers (across schools C and D), we asked, "If you could have any superpowers you wanted, to help you do your job, what would they be?" We first posed this question in a very broad sense, but then asked specifically about superpowers that teachers would find useful during classes in which their students work with an ITS or other adaptive learning technology.

In each study session, we asked teachers to immediately write down their "superpower" ideas on index cards the moment they thought of them — pausing ongoing conversation, if need be — to reduce the chance that they would lose track of an idea. In addition to identifying design opportunities within the cards teachers generated, we wanted to get a better sense of teachers' relative priorities among superpowers, and the reasons underlying these priorities (e.g., the relative severity of the daily challenges underlying these "superpower requests"). To this end, once a teacher finished generating superpower ideas, they were asked to sort them by subjective priority, while thinking aloud about the reasoning behind their sorting (cf. Cairns & Cox, 2008; Hanington & Martin, 2012).

Teachers were encouraged to generate new cards while sorting, in case the card sorting process inspired new ideas. After a teacher had finished sorting their cards, they were presented with cards generated by all teachers who had participated before them, and were given the option to include any of these cards in their own hierarchy. If a teacher found an idea generated by a previous teacher undesirable, they were instructed to omit that card from their hierarchy. If a teacher felt that a superpower idea generated by a previous teacher was synonymous or redundant with one of their own ideas, they were encouraged to align these cards horizontally, to indicate a "tie." For example, Figure 1 shows an excerpt from one hierarchy that emerged from this

[2] Additional participant demographics for all studies reported in this article are available at http://tinyurl.com/JLACodesign

iterative card generation and sorting process. One of this teacher's desired superpowers was "Omniscience," which the teacher considered synonymous with "Being able to see students' thought processes" (a card that a previous teacher had generated).

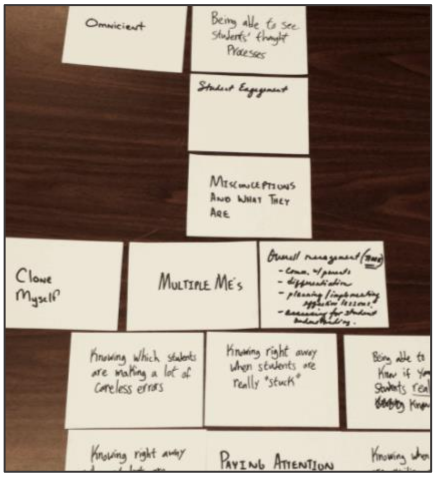

Omniscient
Being able to see Students' thought Processes
Student Engagement
Misconceptions And What They Are
Clone Myself
Multiple Me's
Overall management
- Comm. w/parents
- differentiation
- planning/implementing effective lessons
- assessing for student understanding.
Knowing which students are making a lot of Careless errors
Knowing right away when students are really "stuck"
Being able to Know if yo Students real Kno
Knowing right away
Paying Attention
Knowing whe

Figure 1. Excerpt of a hierarchy produced by one teacher's card sort. Superpower ideas the teacher considered more desirable are placed higher within the hierarchy (from Holstein et al., 2017b).

Overall, teachers tended to prefer "Seeing students thought processes" over most other superpowers, including "Seeing students' misconceptions." Some teachers elaborated that if they could really see and understand students' step-by-step reasoning, this would likely reveal students' misconceptions *and much more*. It is also worth noting that, although estimates of student knowledge (e.g., in the form of probabilities that a student has mastered particular skills) are one of the most central analytics presented by common reporting systems for ITSs (e.g., Heffernan & Heffernan, 2014; Khachatryan et al., 2014; Ritter, Carlson, Sandbothe, & Fancsali, 2015), the superpower "Knowing whether students really know something" ranked lower than most other common superpower ideas.

Across the card hierarchies teachers generated, some interesting regularities emerged. All five of the teachers we interviewed wanted the following abilities:

1. **See students' thought processes.** Teachers wanted to be able to see the chains of reasoning that led students from one mathematical expression to the next, without always having to ask students to "show their work," and without having to spend much time deciphering student work. Some teachers explicitly distinguished "seeing thought processes" from simply seeing percentage estimates of student's mastery over certain skills (which they were used to seeing in reports from adaptive learning software they had used previously), noting that such skill mastery estimates were less *actionable*. That is, if teachers could follow students' thought processes in real-time, this could provide opportunities for them to *"re-route"* students at the moment students *"take a wrong turn"* during a problem solving activity, rather than only providing delayed feedback once the student has moved past the relevant problem.

2. **Know which students are *truly* stuck.** Teachers noted that students often raise their hands during lab sessions when they don't actually need help. At the same time, teachers believed that many students who actually need help the most were the least likely to raise their hands. Being able to see which students actually need the teacher's help, at any given moment, would enable the teacher to better prioritize help across students and *"fight the biggest fires first.*"

3. **Know which students are "almost there" and just need a nudge to reach mastery.** Teachers noted that one of the most fulfilling parts of their jobs is *"seeing students to the finish line"*: working with students who are currently on the verge of understanding a new concept, and helping them reach that understanding more quickly. One teacher was initially conflicted over whether to include this superpower in his hierarchy, noting that students in this situation would likely reach mastery even without the teacher's help. But this teacher ultimately decided to keep this superpower in the hierarchy, acknowledging that, while he generally tries to spend most of his time working with struggling students, he would find it demotivating to spend *all* of his time doing so.

In addition, four out of five interviewed teachers wanted the following abilities:

4. **Temporarily clone myself (create "Multiple Me's").** Teachers wanted the ability to provide one-on-one support to multiple students simultaneously, rather than leaving real-time personalization entirely to the software. All of the teachers we interviewed reported that, while the level of personalization enabled by ITSs is one of its main attractions, such personalization also makes it more challenging for *teachers* to monitor their students' current activities and provide them with timely feedback.

5. **Have "eyes in the back of my head."** Teachers noted that some students take advantage of the challenges personalized learning software such as ITSs poses for classroom monitoring. They shared stories of catching middle-school students switching to non-academic websites when they thought the teacher was not watching, but then immediately switching back when they knew they were in visual range. Thus, much of these teachers' energy is spent "patrolling" the classroom and trying to make sure that everyone is on-task.

6. **Detect students' misconceptions.** Similar to teachers' desire to see students' thought processes, their desire to see student *misconceptions* was rooted in the actionability of this information. While teachers viewed "seeing students' thought processes" as enabling real-time correction of particular student errors, to help shape students' knowledge of procedures, they viewed "detecting students' misconceptions" as enabling the correction of persistent false beliefs that might hinder future learning.

7. **Know which students are making lots of careless errors.** Finally, teachers wanted to be able to more easily detect, in real-time, whether students are putting in the effort required to learn. Based on this information, they could decide on a case-by-case basis whether it would be most productive to spend their time providing additional *instruction*, or whether they should instead try to *motivate* the student.

4.2. Directed Storytelling

To more directly investigate teachers' needs for real-time support, we next conducted semi-structured interviews with 10 teachers across five schools. In these interviews, we asked teachers to walk us through specific, recent experiences using adaptive learning technologies in the classroom. When teachers brought up frustrations and challenges in the course of their storytelling, they were prompted to reflect on how they thought such systems might be better designed. Teachers were encouraged to imagine that there were no technical limitations, and in particular, no limits on what the system could measure about their students' learning.

Two researchers then worked though transcriptions of approximately five hours of video and audio recorded interviews, to synthesize design findings using two standard techniques from Contextual Design: interpretation sessions and affinity diagramming (Beyer & Holtzblatt, 1997; Hanington & Martin, 2012). 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. Affinity diagramming is a widely used, bottom-up synthesis method, aimed at summarizing qualitative patterns across study participants' responses, by iteratively clustering participant quotes into successively higher-level themes (Beyer & Holtzblatt, 1997; Hanington & Martin, 2012). Following several interpretation sessions, the resulting 301 extracted quotes were iteratively synthesized into 40 level-1 themes, 10 level-2 themes, and 4 level-3 themes.

The top-level (level-3) themes that emerged through Affinity Diagramming reflected strong desires to maintain control of the classroom, even when students are working with adaptive learning technologies, and to remain an effective force in the classroom, providing value over and above what these technologies can offer students. Quotes under these high-level themes were often accompanied by expressions of anxiety that educational technologists intend to *replace* their roles as teachers, instead of working to *support* these roles. In addition, the top-level themes reflected teachers' desires for analytics that could truly provide information they *did not already know* and teachers' concerns that real-time analytics in the classroom, if not designed carefully, could easily do more harm than good.

Within these top-level themes, teachers' design requirements and opportunities broke down into the following 10 mid-level themes:

1. **Help me to intervene *where, when,* and *with what* I am most needed.** Teachers wanted support in deciding how best to prioritize their time across multiple students who may compete for their attention at once, when to help (or refrain from helping) a given student, and how best to help. Given teachers' limited time during lab sessions, recommendations about how best to help students might come in the form of in-the-moment, personalized advice about effective instructional strategies to use, to address students' specific areas of struggle.

2. **Make sure the technology does not draw my attention away from my students!** Teachers worried that real-time analytics could easily draw their attention away from their students, thus defeating the purpose. Furthermore, teachers noted that some of the most useful real-time information comes from reading students' body language and other cues that likely would not be captured by a dashboard alone. As such, they emphasized that an effective classroom monitoring system would need to be designed to keep teachers' eyes and ears on the classroom to the greatest extent possible.

3. **How can I know whether what I'm doing is *actually working*?** Teachers noted that opportunities to receive immediate feedback on their own teaching are extremely rare. They often worry, especially after seeing student test scores, that much of what they have taught students over several weeks or months may have had no impact. Observing that intelligent tutoring systems can already track aspects of *student learning* in real-time, teachers wanted these systems to also provide *them* with timely feedback on the effectiveness of their own help-giving (e.g., one-on-one interactions with individual students or targeted mini-lectures provided to the whole class). Receiving such immediate feedback during a class session could allow them to adjust their instructional strategies on the fly.

4. **Help me understand the "why," not just the "what."** Given how busy teachers are working with students during ITS lab sessions, they wanted ITSs to provide them with summarized, directly actionable information whenever possible. A real-time support tool would need to provide concise diagnoses of issues the teacher could act upon. For example, rather than simply presenting teachers with the observation that a particular student is making frequent errors in the software, it would be valuable to also assist the teacher in determining whether this is due to carelessness or genuine struggle with the material (and if the latter, to help the teacher diagnose specific areas of struggle).

5. **I'm just one person: help ease my load.** Teachers emphasized the usefulness of group work and peer tutoring activities in reducing their orchestration load in the classroom. Some teachers suggested that one way an ITS could help them during a lab session would be to recommend groups of students who are likely to be able to help one another (perhaps adaptively matched by the ITS based on its knowledge of their current mastery of specific skills). This would lift some of the responsibility of helping students from the teacher's shoulders, and also enable the teacher to work with a larger number of students who may be struggling with similar issues, by meeting with groups rather than individuals.

6. **But how do I judge whether my students are *really* doing well?** Teachers wanted more support from the ITS in determining what constitutes "good" performance (e.g., is a 70% probability of mastery below or above "average" for a particular skill and amount of practice?).

7. **Help me monitor and manage student motivation.** Teachers noted that it would be useful to have real-time analytics about their students' motivation and affective states in the classroom, not just analytics about student learning and performance. Receiving real-time notifications about student frustration, for example, could allow teachers to intervene before students became too demotivated.

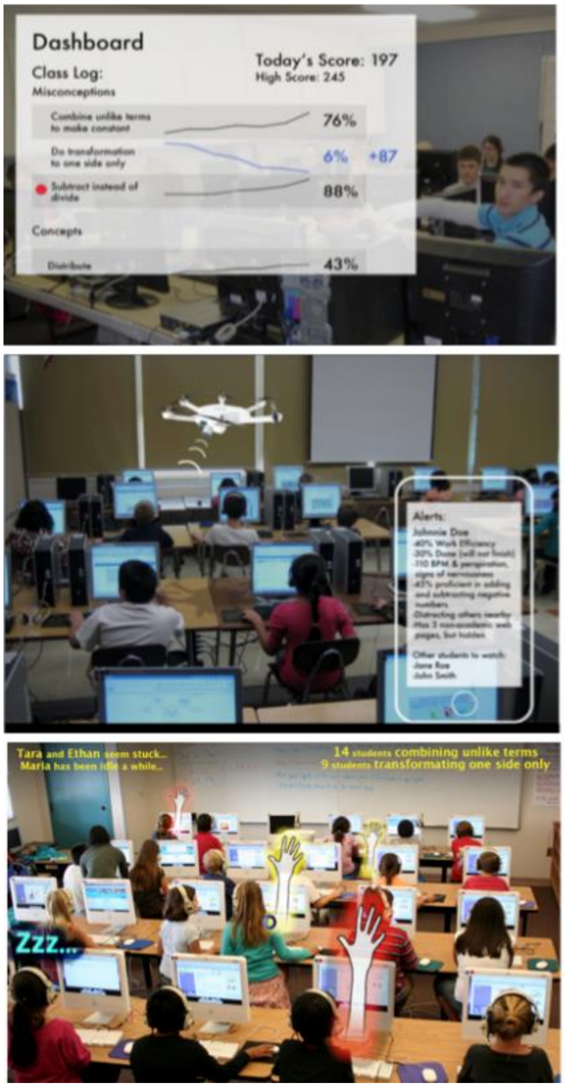
8. **What can you tell me about my students that I do not already know?** Teachers complained that reporting systems they had used in the past tended to provide them with a lot of unsurprising information about their students. Teachers wanted ITSs to take into account what they already knew about their students (e.g., *"[this student] is going to make slower progress, but that's only because she's so deliberate"*), and provide them with notifications only in cases that conflict with their expectations.

9. **Allow me to customize the technology to meet my needs.** Teachers emphasized that, in cases where an intelligent tutoring system's instructional design differs in some way from their own pedagogy (e.g., when the mathematical notation the teacher uses in their lectures differs from that the ITS will accept from students), teachers want to be able to quickly and easy adapt the software to meet their needs.

10. **Allow me to override the technology.** In addition to customization, teachers also wanted the ability to take control of the ITS on-demand. For some teachers, this simply meant being able to "freeze" all of their students' screens while giving an impromptu lecture in the midst of a lab session, to ensure they had students' attention. For others, this meant being able to load a "quiz problem" on all students' screens, to quickly assess the effects of a whole-class lecture on students' knowledge of particular skills.

4.3. Exploring Possible Futures Through Speed Dating

To further understand and validate needs teachers had revealed through the "superpowers" exercise and directed storytelling sessions, we adopted a "speed dating" approach, presenting teachers with futuristic classroom scenarios inspired by these needs. Speed dating is a design method for rapidly exploring possible futures with users, intended to help researchers/designers probe the boundaries of what users will find acceptable (which otherwise often remain undiscovered until after a prototype has been developed and deployed). In speed dating sessions, participants are presented with a number of imagined scenarios in quick succession (represented through sketches, storyboards, or role-playing exercises), while researchers observe and aim to understand their gut reactions to these scenarios (Davidoff et al., 2007). In addition to revealing ways technology concepts may cross boundaries of acceptability, this method can lead to the discovery of unexpected design opportunities when anticipated boundaries are found not to exist or when unanticipated needs are discovered. Importantly, speed dating can often reveal needs and opportunities that may not be observed through field observations or other design activities, such as those described in prior sections.


Dashboard
Today's Score: 197
High Score: 245
Class Log:
Misconceptions
Combine unlike terms to make constant
76%
Do transformation to one side only
6%
+87
Subtract instead of divide
88%
Concepts
Distribute
43%
Alerts:
Johnnie Doe
-40% Work Efficiency
-30% Done (will not finish)
-110 BPM & perspiration, signs of nervousness
-43% proficient in adding and subtracting negative numbers
Distracting others nearby
Has 3 non-academic web pages, but hidden
Other students to watch:
Jane Roe
John Smith
Tara and Ethan seem stuck...
Maria has been idle a while...
14 students combining unlike terms
9 students transformating one side only
Zzz...

"If I know that I lectured and after I lectured, now my misconceptions decrease significantly. Then I know, okay we're on the right track."
- Teacher 3

"Yeah, all I care about... it's all about the kids. I don't need rewards and stickers and points for what I am doing. Like, that's my job."
- Teacher 8

"...Are you kidding me? This is his heart beating and sweating. Oh my god... [But] part of me thinks, 'It would be nice to know how nervous they are,' because I was a very nervous math student back in the day."
- Teacher 9

"...Could I, like use [the drone] to zap my students if they're misbehaving? Nothing that would actually hurt them... a light jolt"
- Teacher 6

"I love this because at a snapshot I see who's struggling. I don't need to go from student to student to student..."
- Teacher 3

"There are always students who are shy and just don't raise their hands... [Some] raise their hands when they really don't need help."
- Teacher 6

Figure 2. Speed dating storyboards and reactions. Left: examples of panels from speed dating storyboards; Right: selected excerpts from teacher reactions to the illustrated scenarios.

We met with five teachers from our previous interviews, and presented them with futuristic classroom scenarios inspired by teachers' own expressed needs. Teachers were presented with eleven storyboards. Each storyboard presented a scenario intended to probe the boundaries of acceptability, generated based on teachers' most commonly requested "superpowers," themes from our directed storytelling interviews, and notes from field observations in teachers' classrooms (Holstein et al., 2017a; 2017b). Key findings from these speed dating sessions are summarized below.

Contrary to teachers' expressed desire for real-time support in prioritizing their time across multiple students during a lab session, teachers consistently rejected the idea of "time management" systems that explicitly nudge them to spend less of their time with certain students (e.g., those who seem to be doing well without the teacher's help) and more of their time with others (who may benefit from more assistance). For example, one teacher reacted strongly to this concept, stating, *"I don't need that… to remind me it's time to move on. I know that. As an educator, you know when you've got other kids to deal with."* Although recent work suggests that, contrary to this teacher's assertion, educators' intuitions about which students need the most help during personalized lab sessions can be limited (Holstein et al., 2017a), this teacher's comment reflects a core tension in the design of real-time, intelligent teacher supports. Teachers' comments in response to this storyboard suggested that such "time management" systems can be undesirable if they threaten teachers' autonomy in the classroom, and also if they remove teachers' ability to choose, on a case-by-case basis, between two conflicting desires during lab sessions (corresponding to two of the superpower ideas teachers generated): helping students who are "almost there" versus helping struggling students who may be most in need of help.

By contrast, teachers were highly receptive to technology designs that presented information to help them prioritize their time among students, without attempting to *directly recommend* specific actions. For example, the bottom panel in Figure 2 comes from one of the most positively received storyboards. This panel shows a heads up display, through which an ITS can inform the teacher that a given student may need their assistance (even if the student is not necessarily aware that they need help), without explicitly recommending that the teacher help a particular student at a particular moment. As shown to the right of this panel in Figure 2, some teachers noted that such a tool would be particularly helpful because students' hand-raising behaviour can be an unreliable indicator of their actual need for help.

A key reason teachers gravitated towards the concept of a heads up display was that, by displaying analytics directly overtop their view of the classroom, this technology would not draw their attention away from the classroom. Teachers' reactions to this and other storyboards, including a smart-watch-based classroom awareness tool, suggested that they might be quite open to and even *prefer* wearable interfaces for real-time use cases, as opposed to handheld displays such as tablets and mobile phones. In particular, teachers saw heads-up displays as an opportunity to have their own "private" smart classroom with analytics only they could see, preserving privacy between students.

5. Prototyping Real-Time Classroom Orchestration Tools

Our early design findings, discussed in section 4, laid the foundation for a broad research program. These design studies revealed strong needs for classroom AI systems that can effectively support *teachers* in addition to their students, enabling human teachers to remain in control of their classrooms, while freeing them up to do what they are uniquely good at. Building on these findings, we decided to narrow our scope, at least for an initial prototype, to the design of tools that specifically support real-time teacher awareness and decision-making in AI-enhanced classrooms, as opposed to system customization and control.

We next wanted to gain a more concrete sense of what real-time analytics would be most helpful to K–12 teachers during ITS class sessions, and how teachers would envision actually *using* such analytics during a class session, to inform their in-the-moment decision-making. In addition, we decided to further explore the idea of using heads up displays such as smart glasses, given teachers' desire to keep their heads up and their attention focused on the classroom, and given the enthusiasm around this idea that we had observed in our speed dating study. To these ends, we conducted a series of iterative design studies with a total of 16 middle-school math teachers across nine schools and six school districts. All participating teachers had previously used adaptive learning technologies in their classrooms, and 12 out of 16 had used an intelligent tutoring system as a regular component of their teaching.

5.1. Iterative Low-Fidelity Prototyping

To further understand teachers' needs and desires for real-time awareness support, before developing specific prototypes, we conducted a sequence of three lo-fi experience prototyping (Buchenau & Suri, 2000) and participatory comic-boarding (Moraveji et al., 2007) sessions with middle-school math teachers. For all studies, researchers travelled to schools to work with teachers in their own classrooms.

In each study session, a teacher viewed a computer screen showing a full-screen image of a classroom full of students working with adaptive 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 in front of the image (simulating the experience of using actual smart glasses). Floating text labels appeared over individual students' heads, alerting teachers to students' current detected knowledge or behavioural states, in accordance with common teacher "superpower" ideas in our earlier design work. For example, by looking around the classroom, teachers could instantly see that certain students were currently struggling in the software, might be off-task, or were frequently making careless errors. In addition, two class-level dashboards appeared against the front wall of the classroom, visible only through the "smart glasses," based on teachers' expressed desires for real-time information at the class-level. One of these dashboards showed a list of skills that had been practiced by multiple students in the class, but mastered by very few students, and the other dashboard showed a sorted list of common errors that multiple students in the class had recently exhibited.

The image showed a single instant during a class session, frozen in time, 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 glasses if this were an actual class session. Teachers were encouraged to remark on any information that was displayed to them, but which they did not find useful, as well as information that was not visible but which might inform their decision-making. For example, although one of teachers' "superpower" ideas was to be able to see when students are frequently making "careless errors," all teachers participating in this prototyping study expressed strong discomfort with the idea of a computer making judgments about students' motivation (e.g., "carelessness"). To facilitate brainstorming, teachers were also provided with a large, printed copy of the same classroom image shown on-screen, but with blank rectangles in place of the individual student labels and classroom analytics displays (see Figure 3).

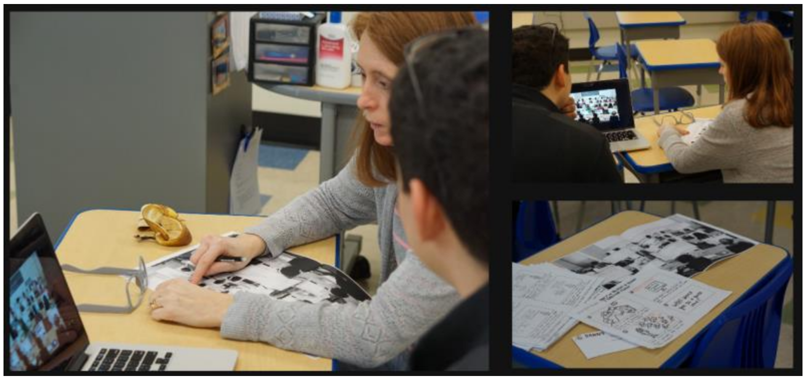


Figure 3. A researcher working with a K–12 teacher to design concepts and potential use scenarios during a low-fidelity prototyping session.

Throughout each session, teachers could use these blank spaces to sketch out new ideas for real-time information that might be displayed through the glasses. Each time a teacher generated an idea for new information, a researcher would press the teacher to describe how they envisioned using that information during a real class session. We found that the process of generating hypothetical use cases for particular analytics often led teachers to refine their ideas, as they realized that more, or different kinds of information might be needed to support particular decisions. As in the "superpowers" study, the ideas that a teacher generated during one study were ultimately incorporated into the version of the experience prototype (i.e., the image and overlaid analytics) that we would show to the following teacher.

At opportune moments throughout each study, researchers also probed teacher reactions to specific classroom scenarios involving the use of smart glasses, using storyboards prepared before the study. We took a participatory comicboarding approach (Moraveji et al., 2007), typically leaving the final panel or two of a comicboard blank. This allowed teacher to fill in the details of how *they* would imagine a classroom scenario progressing, or what decisions and actions they might take in that scenario, rather than relying entirely on a researcher-generated sequence of events.

During the first lo-fi prototyping session, we found that it was challenging for the teacher to imagine the actual experience of using mixed-reality smart glasses in the classroom. So, for the second and third sessions, we added an experience prototyping phase at the beginning of the study, using actual mixed-reality smart glasses (although with Wizard of Oz'd analytics, presented at a single instant in time). We used the Microsoft HoloLens,[3] which enabled us to place readily available, default HoloLens assets at fixed spatial positions throughout a teacher's classroom. When teachers then began the lo-fi prototyping and sketching exercises, they could refer back to this experience.

5.2. Iterative Mid-Fidelity Prototyping

We next moved to higher-fidelity prototyping sessions, given that we had observed strongly positive reactions to the concept of teacher smart glasses in early prototyping sessions and had also begun to gain a more detailed understanding of teachers' real-time information needs. We conducted an iterative sequence of prototyping sessions with five middle-school math teachers. As

[3] https://www.microsoft.com/en-us/hololens/hardware

in our earlier prototyping studies, researchers travelled to middle-school sites and worked with teachers in their own classrooms. Each study session lasted for 90 minutes. The teacher wore the HoloLens for the first hour and participated in experience prototyping activities (Buchenau & Suri, 2000), experimenting with different combinations and spatial configurations of analytics displays while generating ideas for potential use cases. Following this experience prototyping phase, teachers participated in a 30-minute semi-structured post-interview in which they had the opportunity to reflect on their experiences and provide more detailed design feedback. For these and subsequent prototyping studies, we narrowed our focus specifically to the context of middle-school math classrooms using ITSs for equation solving.

In order to present teachers with a range of design alternatives, we used a modified version of HoloSketch,[4] a HoloLens application for rapid prototyping of mixed-reality experiences. Using HoloSketch, we were able to position two-dimensional assets, including mock-ups of student- and class-level analytics displays created in Photoshop, throughout a teacher's physical classroom space. For example, when a teacher put the HoloLens on, they could see indicator symbols floating over empty student seats, and class-level analytics displays appearing as "wall decorations" that the teacher was able to reposition as they saw fit (see Figure 4, left). Throughout each prototyping session, the teacher had the opportunity to move about their classroom. Teachers were asked to think-aloud during these sessions, imagining what actions they might take in response to the displayed analytics if this were a real class session, and what other information might support them in making such decisions.

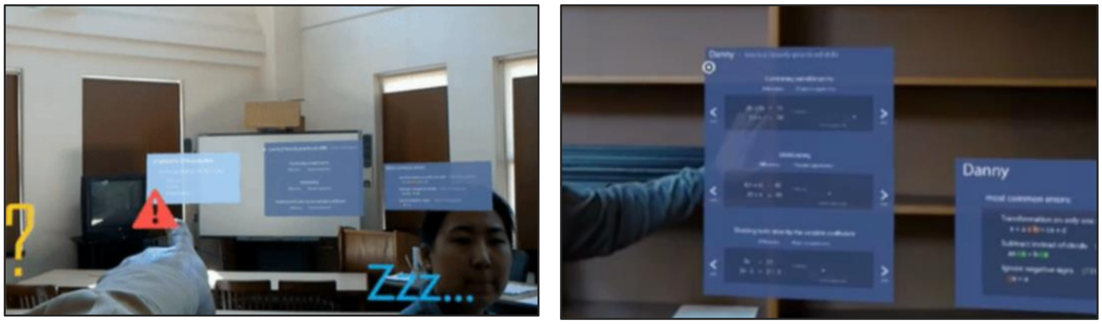


Figure 4. Screenshots from the teacher's point-of-view during a mid-fidelity prototyping session. Left: the teacher performs a think-aloud while positioning combinations of analytics displays throughout the classroom. Right: the teacher discusses design alternatives in a "gallery" at the back of the classroom.

In the first mid-fi experience prototyping session, we included all of the indicator symbols and analytics displays that teachers had consistently requested up until this point, in our lo-fi prototyping studies. Then, in-between sessions, we rapidly iterated on the design of individual student- and class-level displays, incorporating new ideas that teachers had generated during the previous session. Since our design mock-ups were synchronized with the HoloLens app as 2-D assets, we were also able to make modifications during a session based on teachers' live design feedback, by editing these assets on a laptop as a teacher viewed them (in an appropriate spatial context) through the HoloLens. Between prototyping sessions, we also reflected on our areas of greatest uncertainty. For each open question, we mocked up several design alternatives. Towards the end of each session, we brought the teacher to the back of their classroom, where (in mixed-reality) we had arranged an immersive "gallery" of these new design alternatives (see Figure 4, right). Teachers could reposition these information displays and experiment by decorating their classrooms with different combinations and arrangements of displays, all while thinking aloud and providing design feedback. Based on this feedback, we iterated on these designs prior to the next prototyping session, providing opportunities to validate previous teachers designs (and the needs underlying these designs) with new teachers in subsequent sessions.

5.3. Design Findings From Low- to Mid-Fidelity Prototyping

Our research team worked through transcriptions of approximately 12 hours of audio/video-recorded prototyping sessions, across eight teachers, to synthesize design findings through Interpretation Sessions and Affinity Diagramming (Beyer & Holtzblatt, 1997; Hanington & Martin, 2012). Following a series of Interpretation Sessions, the resulting 655 quotes were iteratively synthesized into 77 level-1 themes, 23 level-2 themes, 10 level-3 themes, and 7 level-4 themes. Key high-level findings (level-4 themes) are summarized below:

1. **Student-level indicators.** Five major categories of student learning and behavioural states emerged from these co-design sessions, shown in Figure 5. Teachers strongly preferred to keep these indicators visually simple, displaying a single graphical symbol above each student's head (as in Figure 6) to avoid information overload during a class session. However, it was also important to teachers that they could access brief elaborations on-demand (e.g., by

[4] https://github.com/Microsoft/MRDesignLabs/tree/master/ReleasedApps/HoloSketch

having such elaborations appear when looking directly at an indicator, as shown in Figure 6), which could help teachers better understand why an indicator was appearing for a student at a particular time. In line with our prior findings, all teachers expressed a desire to see positive information about individual students, not just negative information. In particular, teachers wanted to be able to see when students have been performing particularly well recently. Teachers found this valuable for several reasons, including motivating themselves (since seeing only negative alerts might be discouraging), motivating students (by identifying and praising students who have been doing well lately), and identifying students who may be under-challenged by the software.

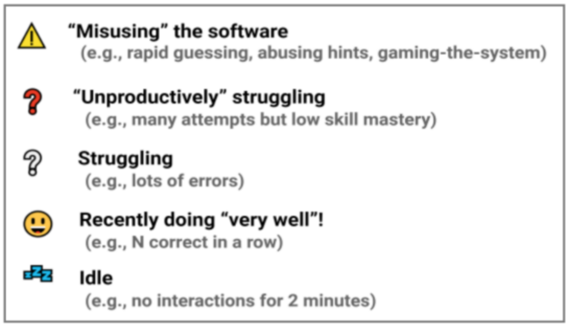

"Misusing" the software
(e.g., rapid guessing, abusing hints, gaming-the-system)
"Unproductively" struggling
(e.g., many attempts but low skill mastery)
Struggling
(e.g., lots of errors)
Recently doing "very well"!
(e.g., N correct in a row)
Idle
(e.g., no interactions for 2 minutes)

Figure 5. Consistently requested categories of real-time indicators (from Holstein, Hong et al., 2018).

2. **Student-level indicators.** Five major categories of student learning and behavioural states emerged from these co-design sessions, shown in Figure 5. Teachers strongly preferred to keep these indicators visually simple, displaying a single graphical symbol above each student's head (as in Figure 6) to avoid information overload during a class session. However, it was also important to teachers that they could access brief elaborations on-demand (e.g., by having such elaborations appear when looking directly at an indicator, as shown in Figure 6), which could help teachers better understand why an indicator was appearing for a student at a particular time. In line with our prior findings, all teachers expressed a desire to see positive information about individual students, not just negative information. In particular, teachers wanted to be able to see when students have been performing particularly well recently. Teachers found this valuable for several reasons, including motivating themselves (since seeing only negative alerts might be discouraging), motivating students (by identifying and praising students who have been doing well lately), and identifying students who may be under-challenged by the software.

3. **Sequences of student states can be information-rich.** In addition to seeing indicators that reflect students' current "states," teachers noted that it would be useful to see sequences of detected states that preceded a student's current state. For instance, if a student is currently "idle" or "misusing the software" in some way, it would be useful to know whether that student was *also* recently struggling. Teachers would then interpret the prior struggle as a possible cause of the student's current behaviour, and respond accordingly.

4. **The classroom as a dashboard.** During our experience prototyping sessions, teachers remarked that it felt natural to reference information displays that were distributed throughout their physical classroom spaces. In the absence of a real-time awareness tool, teachers were used to monitoring their students by scanning the physical classroom (e.g., reading students' faces and body language), and "patrolling" rows of student seats to catch quick glances of individual 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.

5. **Needs 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 unproductive social comparison and competition among their students (cf. Aguilar, 2018). 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 that they would want a mechanism to selectively share particular analytics during the course of a class session. Five out of eight teachers suggested it would be

useful to customize the shared visibility of particular analytics on a class-by-class basis. All of these teachers predicted an interaction effect in which real-time analytics might *motivate* higher-achieving classes by promoting healthy competition among students, while *demotivating* lower-achieving classes.

6. **Ground automated inferences in "raw" examples of student-generated artifacts.** Teachers emphasized the importance of having access to "raw" student-generated artifacts, in a familiar format, *in addition to* the system's automated inferences. For example, the mock-ups of "deep-dive" screens shown in Figure 6 display individual students' greatest "areas of struggle," including examples of actual student errors for each area of struggle that the student has recently exhibited. Showing these example errors is crucial not only in helping the teacher perform further diagnosis, but also in supporting teacher trust (cf. Bull & Kay, 2016; Lipton, 2016) or enabling the teacher to productively "second guess" system judgments.

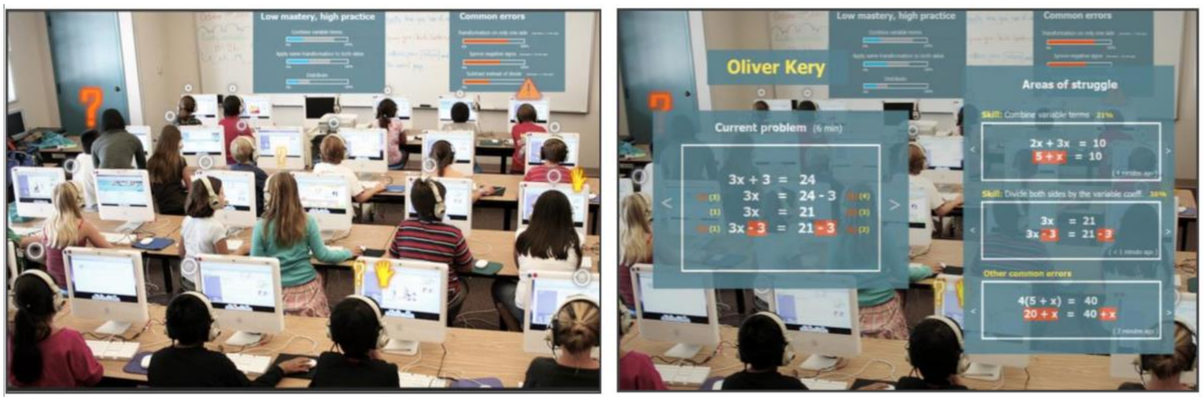


Figure 6. Design mock-ups based on findings from low- to mid-fidelity prototyping sessions (from Holstein, Hong et al., 2018). Left: Teacher's default view of the class. Each student has an indicator display floating above their head, and class-level analytics displays are positioned at the front of the class. Right: "Deep-dive" screens shown if a teacher "clicks" on an indicator.

7. **Enable teachers to "peek" at students' 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 current activities (potentially annotated, as in Figure 6). Although we had expected that teachers would prefer to simply walk over to a student and observe that student's screen directly, teachers noted that physically approaching students can cause them to alter their behaviour, potentially reducing the diagnostic usefulness of direct observations (cf. Holstein et al., 2017a).

8. **Support anonymous teacher–student communication ("Invisible hand raises").** Although most teacher design feedback focused on ways real-time analytics could help them regulate student learning, some teachers emphasized the importance of also creating opportunities to develop students' help-seeking skills (Aleven, Roll, McLaren, & Koedinger, 2016). Several teachers proposed the idea of giving students an "Ask the teacher" button within the tutoring software, which would trigger a "raised hand" symbol within the glasses. Teachers expected that, by providing students with a mechanism to request help that was not easily visible to other students, more students would feel comfortable requesting help (cf. Schofield et al., 1994). 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."

5.4. Development of a Higher-Fidelity Prototype

Up until this point, all prototyping sessions had relied upon Wizard of Oz'ing analytics, presented "frozen in time" at a single instant of a class session. However, the behaviour of a real-time learning analytics tool can be heavily dependent on the dynamics of specific data-generating contexts in combination with specific analytic methods/algorithms. We next wanted to begin prototyping the experience of using smart glasses to monitor a class session unfolding over time, using real student data and analytics. Based on our findings from low- to mid-fidelity prototyping, we developed a mixed-reality application called Lumilo (see Figure 7), using Unity3D[5] and the HoloToolkit[6] for the Microsoft HoloLens. Using a newly extended version of the CTAT/TutorShop architecture for ITS authoring and deployment (Holstein, Yu, et al., 2018), we developed an initial set of

[5] https://unity3d.com

[6] https://github.com/Microsoft/HoloToolkit-Unity

automated detectors of student learning and behaviour, making use of existing student modelling techniques (e.g., Aleven et al., 2016; Beck & Gong, 2013; Desmarais & Baker, 2012; Käser et al., 2016), to provide teachers with each of the key categories of indicators described in the previous section. When plugged into an ITS, the real-time analytics generated by these detectors would then be streamed to a learning management system, and finally to Lumilo, where they would update mixed-reality displays visible through the teacher's glasses. (Further details of Lumilo's technical implementation are provided in Holstein, Hong, et al., 2018; Holstein, Yu, et al., 2018).

To support future design explorations, we architected the initial prototype of Lumilo in a highly modular fashion, to enable rapid design iteration in-between future prototyping sessions, and even to make small adjustments within a single prototyping session, based on live teacher feedback. For example, alternative student modelling (detector) algorithms intended to measure the same teacher-identified construct (such as "unproductive struggle") could be interchanged for comparison during a prototyping session. All detectors included in our initial prototyping sessions were drawn from the LA, educational data mining, and AI in education literatures — where many automated detectors of student learning and behaviour have been introduced and validated, which rely only on students' interactions within the software (e.g., Aleven et al., 2016; Beck & Gong, 2013; Desmarais & Baker, 2012; Käser et al., 2016). For example, in order to drive a real-time indicator of "unproductive struggle," we explored the use of simpler methods such as Beck and Gong's detector of "wheel-spinning" (Beck & Gong, 2013), as well as more sophisticated methods such as Käser et al's (2016) "predictive stability." Each detector was implemented in a parameterized fashion, so that aspects of a detector's behaviour (e.g., tunable alert thresholds) could be adjust during and between sessions based on teacher feedback.

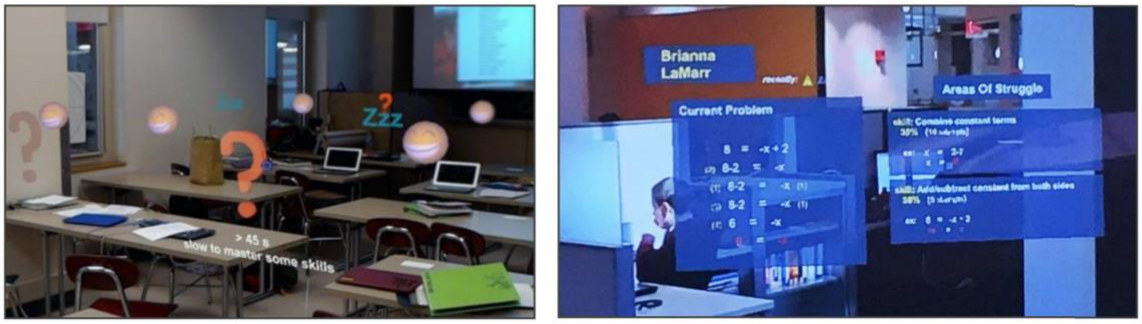


Figure 7. Point-of-view screenshots from teachers using Lumilo (from Holstein, Hong et al., 2018). Left: A teacher's view of student indicators, immediately following a pilot study in a live classroom. Right: A teacher's view of a classroom while wearing Lumilo (photos taken with no students present in the room to preserve student privacy).

We also developed a new logging library for Lumilo, which automatically logs teacher actions in a physical classroom space over the course of a class session. For example, Lumilo can record time-stamped logs of a teacher's physical proximity to a given student in the class moment-by-moment, as well the teacher's absolute location in the classroom, their proximity to landmarks (such as the teacher's desk or whiteboard), the target of a teacher's gaze, and all teacher interactions within the tool interface. These logs are recorded to DataShop, a major educational data repository (Koedinger et al., 2010).

5.5. Prototyping Real-Time Classroom Analytics Using Replay Enactments

Using this functional prototype, we next conducted an iterative sequence of higher-fidelity experience prototyping sessions (Buchenau & Suri, 2000), with a total of 10 math teachers across five schools. All participating teachers had previously used an adaptive learning technology in their classrooms, and seven out of 10 teachers had used an ITS as a regular component of their classroom instruction.

To rapidly prototype the experience of using Lumilo in a classroom, *prior* to running studies with the system in live classrooms with actual students (which can be costly in K–12 settings, and may be harmful to students if the prototype is still "rough"), we developed a new prototyping method: Replay Enactments (REs). Much like other recently proposed prototyping methods in the LA literature, such as the simulation methods presented in Martinez-Maldonado, Kay, Yacef, and Schwendimann (2012) and Mavrikis et al. (2016), REs involve replaying log data from students' interactions with educational technologies in order to prototype real-time analytics and visualizations with end-users (such as teachers or students). However, in the spirit of recent methods from the field of Human–Computer Interaction (HCI) for prototyping radically new experiences, such as User Enactments (Odom, Zimmerman, Davidoff, Forlizzi, Dey, & Lee, 2012), REs build on prior LA approaches by emphasizing *embodied role-playing* in physical classroom environments. In our initial work piloting this prototyping method with teachers, we found that pushing teachers to role-play while actually navigating throughout a physical classroom space seemed to contribute to an illusion of "actually being there." In addition, having the teacher move throughout the classroom provided early insight into potential effects of a classroom's layout and students' seating arrangement relative to this layout (cf. Holstein et al., 2017a).

In contrast to User Enactments, which typically involve Wizard-of-Oz'd scenarios, REs prototype an experience using authentic data and algorithms, evolving over time. Although this requires earlier investment in technical development, doing so can enable earlier, nuanced observations of the interplay between human and machine judgments, such as the ways in which a system's false positives and false negatives may impact the user experience (UX) of using a data-driven algorithmic system (cf. Dove, Halskov, Forlizzi, & Zimmerman, 2017). As such, REs responds to recent calls within the human–computer interaction and machine learning communities (e.g., within the nascent "UX of AI" community) for "new kinds" of prototyping methods that can address the unique challenges of prototyping data-driven algorithmic systems such as learning analytics tools (e.g, Doshi-Velez & Kim, 2017; Dove et al., 2017). The behaviour of such systems is highly dependent on interactions between particular data contexts (e.g., specific sociocultural contexts from which educational data was collected) and particular algorithms (e.g., specific machine learning models trained on specific datasets), which cannot easily be imagined ahead of time by system designers. For example, a recent study of industry product teams' challenges around algorithmic bias at major companies, including several companies working on learning analytics applications, found that a lack of suitable pre-deployment prototyping methods was a central pain point (Holstein, Wortman Vaughan, Daumé, Dudík, & Wallach, 2018).

Figure 8 shows a general, high-level description of an REs prototyping study. REs involve the simulation, at the highest level of fidelity feasible, of a relevant data-generating context (such as a classroom of students working with adaptive learning technologies). To generate this simulated context, authentic (rather than Wizard-of-Oz'd) data streams are replayed at the same speed at which the data were originally generated. Within this simulated context, participants are equipped to receive two key streams of sensory input: first, a simulated approximation of what the user would perceive in the target environment (e.g., in an actual classroom), even without access to real-time analytics or visualizations; and second, a particular form of cognitive augmentation (e.g., a specific tool design, including particular analytics and visualizations). Given a simulated context and a particular form of augmentation, the user is asked to complete an approximation of an authentic, complex task. Ideally, if the target task involves spatial navigation decisions (as in a vehicle driving task, a hands-on learning activity, or a simulation of active classroom teaching), the simulated task should reflect the physical nature of the target task.

These components are shown as nested boxes in Figure 8 to indicate that each can be swapped out for comparison purposes (e.g., the same form of cognitive augmentation, a particular learning analytics system design, might be tested across replays of datasets generated from multiple, diverse classroom contexts). In addition to generating qualitative insights, REs can be used to provide early insight into the effects of different tool designs (e.g., particular choices of analytics) might have on user behaviour. Since the use of data replays removes the possibility that user behaviour will influence the data streams being replayed (and thus removes the possibility of feedback loops), REs can also support early empirical evaluations of how effectively an early warning/alert system might steer users' attention.

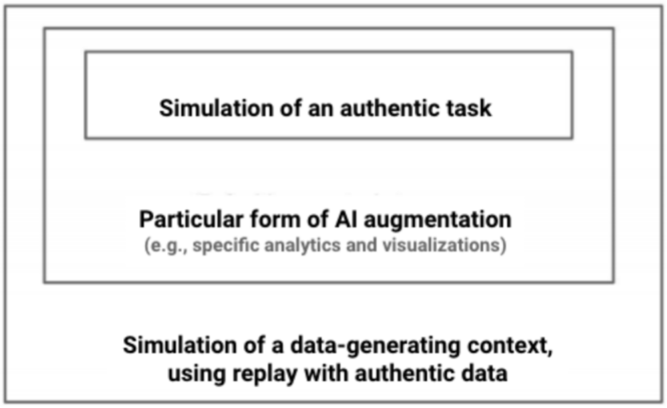

Simulation of an authentic task
Particular form of AI augmentation
(e.g., specific analytics and visualizations)
Simulation of a data-generating context,
using replay with authentic data

Figure 8. High-level diagram showing the modular, nested components of Replay Enactments sessions.

In each of an initial round of five REs study sessions, each held with a single teacher at a time, we brought teachers into a computer lab on our university campus. At each empty seat in the lab, we had placed a nametag with a fabricated student name before the study session began. On the corresponding computer screen, we had logged into the tutoring software, under that student's name. In addition, using Lumilo, we had positioned mixed-reality 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 class of students, using datasets previously collected from a multi-classroom study in which middle-school students used Lynnette, an ITS for linear 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, thinking aloud as they moved throughout the classroom space. If the teacher thought they might focus attention on a particular student at a particular time, they were asked to verbalize what they might say to that student in-the-moment, if the student were actually there. Teachers often became quite immersed in this task. As one teacher remarked, about halfway through the REs session, *"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."*

We ran separate REs sessions with a total of five teachers. Each of these sessions began with a 35-minute training and familiarization phase, during which the teacher could acclimate to using the system. This was then followed by a 40-minute simulation phase, during which the teacher was asked to think-aloud, and a 15-minute post-interview to elicit additional design feedback. To prototype the experience of using Lumilo under a diverse 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, where class tiers were based on those assigned by the schools from which these datasets were drawn. We then randomly assigned dataset to REs study sessions, so that the remedial and average classes were replayed for two teachers each, and the advanced class was replayed for the remaining teacher. To account for potential influences of a classroom's spatial layout, different computer labs, with a range of spatial layouts, were used across study sessions.

During REs sessions, our goal was to elicit teacher feedback not only on Lumilo's interface design and the visual presentation of analytics, but also on the specific choices of *learning analytics* used to drive Lumilo's real-time indicators and class-level dashboards. During each session's training and familiarization phase, teachers were provided with definitions for each indicator symbol. These included brief summaries of a detector's structure, the main features it relies upon, and the default settings of any free parameters (e.g., alert thresholds) used by an indicator or its corresponding detector.

Within the simulation phase of each session, teachers frequently monitored students' raw activity within the software (either by approaching a student's computer terminal and observing their screen, or by opening the student's deep-dive window through the glasses interface). In doing so, they often observed ways in which particular detectors might have been over- or under-sensitive, or may have been overlooking key features of student thinking and behaviour entirely. Such feedback provided opportunities for us to iterate on the selection and design of detectors and alert policies that drove Lumilo's real-time indicators in-between REs sessions (and sometimes even within a single REs session). For example, over several iterations, the definition of the "struggling" indicator evolved to include not only a threshold on a student's recent error rate, but also the automated detection of student *hint avoidance* (Aleven et al., 2016), as well as whether a student remained stuck despite having made good use of the software's built-in hint function, with the corresponding "question mark" symbol glowing gradually brighter, the longer the student remained stuck. By the final two REs sessions, teacher observations of under- or over-sensitivity, or mismatches between the analytics and a teacher's own judgments of a student's knowledge or behaviour, had become relatively rare.

Examples of other design features that entered the prototype during this iterative refinement process, included the ability to set visual "reminders" 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 few minutes. As one teacher put it, *"You want to stay with a kid until they have it mastered but... there's that advantage to saying 'Okay, try a few of these, I'll come back to you.' I've never found a good answer to that one."* In addition, we found that teachers saw great value in the ability to monitor individual students' activities from a distance, while walking around the classroom or while working face-to-face with a student seated across the room. As such, we enhanced Lumilo so that a teacher could have the deep-dive screen "tag along" with them as they walked (as opposed to hanging in space near the corresponding student, visible only when the teacher was looking in that student's direction). Finally, to give teachers "eyes in the back of their heads," a common need from our early "superpowers" design study, we added ambient, spatial sound notifications. For example, if a student was misusing the software, a teacher could privately perceive a soft auditory notification, as if it were emanating from that student's location in the classroom.

5.6. Design Findings from Replay Enactments

As before, we conducted Interpretation Sessions and Affinity Diagramming to synthesize design findings from transcriptions of approximately 18.5 hours of audio/video recorded think-aloud data and design feedback. The resulting 486 quotes were iteratively synthesized into 43 level-1 themes, 26 level-2 themes, 13 level-3 themes, and 5 level-4 themes. Key high-level findings from this synthesis (level-4 categories) are highlighted below.

We see these design findings as fruitful directions for future work.

1. **Value of continuous, real-time feedback on *instruction*.** Although Lumilo did not provide direct feedback to teachers about their own instruction, teachers frequently inferred potential *effects* of their instructional interventions (e.g., helping an individual student or providing a brief whole-class lecture) by monitoring changes in student and class state following an intervention. In fact, teachers were often tempted to infer causality even during REs sessions, in which no students were actually present. In line with findings from our earlier directed storytelling and speed dating studies, teachers emphasized that receiving more direct, in-the-moment feedback about the effects of their own teaching on student learning could help them adjust their instruction on-the-spot, and perhaps even *improve* their teaching over time (especially if this in-the-moment feedback was constructive).

2. **When many students need help on different topics at the same time, choice can be anxiety inducing.** During REs, teachers realized that when they were made more aware of student struggle during a class session, they also 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 they would otherwise be overloaded (e.g., when many students need their help simultaneously, or in more chaotic classes that require teachers to devote more attention to basic classroom management). 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."*

3. **Action recommendations *in addition to* awareness support.** As we progressed to higher-fidelity prototyping, teachers consistently noted that it would be helpful to have more explicit action recommendations from the system, to help them prioritize their attention across students and/or to decide how best to help particular students. For example, one teacher suggested that it would sometimes be helpful to receive recommendations for *"conversation starters,"* such as self-explanation prompts they could give a student, targeted to that student's current areas of struggle, to avoid providing "too much" scaffolding. It is clear from our early design explorations, however, that such a system would need to be designed with great care, to respect teacher autonomy and to ensure that system recommendations are aligned with teachers' goals.

4. **Automated support for dynamic, adaptive peer matching.** In line with findings from our early directed storytelling sessions, teachers noted that it would be useful to receive support in adaptively and dynamically assigning students to serve as peer tutors throughout a class session (cf. Diana et al., 2017; Olsen, 2017).

5. **Trade-offs between accuracy and interpretability.** Although teachers had expressed a preference for simpler, more interpretable analytics in lower-fidelity prototyping sessions, it became apparent during higher-fidelity prototyping sessions that the strength of this preference may depend heavily on the underlying construct that a real-time indicator was purporting to measure. For example, when it came to detection of "system misuse," it was important to teachers that they could easily understand (and thus justify to students) precisely the patterns of student actions that had led to this classification. By contrast, teachers appeared to be more open to the use of "black box" algorithms for detecting "unproductive persistence" if this meant alerting them to these students *earlier* (given that after this initial alert, teachers could apply their own discretion, using other information available to them).

5.7. In-lab Evaluation Using Replay Enactments

Prior to piloting Lumilo in live K–12 classrooms, we wanted to better understand its effects on teacher behaviour. We ran an in-lab evaluation consisting of an additional six REs, across which Lumilo's design was held constant, to investigate whether and how Lumilo might influence teachers' time allocation (cf. Martinez-Maldonado, Clayphan, Yacef, & Kay, 2015) across students of varying prior domain knowledge and learning rates compared with business-as-usual (i.e., without an orchestration tool).

Each session replayed data from a 40-minute class session, randomly selected from a pool of five "average" and "remedial" classes. An "average" class was replayed in four REs sessions, and a "remedial" class was replayed in the remaining two sessions. Advanced classes were omitted from the selection pool for this study, given that there was relatively little between-student variance in test scores in these classes. To minimize potential effects of student names or seating positions on teachers' behaviour, replayed students were randomly assigned names and seats in each session.

To measure how teachers allocated their time across students, we architected Lumilo so that the mixed-reality indicators positioned above students' heads doubled as proximity sensors within a physical classroom space. We measured each teacher's allocation of time to a given student as the cumulative time (in seconds) that they spent within a 4-foot radius of that student's indicator. If a teacher was within range of multiple students, time was accumulated only for the nearest of these students. We used hierarchical linear modelling (HLM) to predict teachers' time allocation across replayed "students" as a function of either students' prior domain knowledge (measured by a pre-test in the original class session being replayed) or students' learning (measured by a post-test, controlling for the student's pre-test score). As is the case in a typical classroom study, teachers did not have access to pre- or post-test data. Lumilo did not use any pre- or post-test data when presenting analytics to teachers. Using 2-level models with students nested within classrooms provided a better fit than 1-level or more complex models. Standardized coefficients for student-level variables are shown in row 2 of Table 1. As shown, teachers using Lumilo in REs spent significantly more of their time attending to replay "students" with relatively lower pre-test scores, or lower post-test scores (controlling for pre-test).

Table 1. Relationships between Teachers' Time Allocation across Replayed Students (in seconds) and Students' Prior Knowledge (pre-test score) and Learning (post-test score controlling for pre-test)
* p < 0.05, ** p < 0.01, *** p < 0.001

Class Type	Number of Teachers	Number of Classrooms	Average Class Size	Using Lumilo?	Pre-test	Post-test | Pre-test
Live	4	7	16	No (business-as-usual)	6.29	–5.49
REs	6	3	15	Yes	–4.66*	–21.19**

By contrast, row 1 of Table 1 shows results from a prior in-vivo classroom study with four teachers (across seven live middle-school classrooms), in which students worked with Lynnette while their teacher monitored and helped students (without access to an orchestration tool). Performing the same analysis as above, this time with data from the 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 1 (row 1). Although all participating teachers reported attempting to devote most of their time to students whom they expected would struggle with the material, we found no significant relationships between students' pre- or post-test scores and teachers' time allocation across students.

We took this contrast as preliminary evidence that Lumilo may aid teachers in focusing on and helping those students with lower prior knowledge. More importantly, we interpreted these results as suggestive that Lumilo may successfully aid teachers in identifying students who would have gone on to exhibit the lowest *learning* in an actual classroom session, at least without the teacher's help. Since the use of replay removes the possibility of a causal arrow from teacher behaviour to students' learning within the software, this method allows us to investigate counterfactuals such as the above for different forms of teacher augmentation. Conversely, classroom studies — although much costlier to run — allow investigation of the effects of a tool in the social context where it is ultimately intended to be used, in the presence of many competing influences on a teacher's attention and judgment.

6. Lumilo Goes to School in the Big City

Given promising findings in our in-lab evaluation study, we next brought Lumilo into real middle-school classrooms. Over the course of one school year, we iteratively piloted Lumilo in 18 middle-school math classrooms across five schools and tested the glasses in an additional 12 classrooms as treatment conditions in an in-vivo classroom experiment (Holstein, McLaren, & Aleven, 2018a; 2018b). This experiment investigated the effects of providing teachers with real-time analytics about student learning, metacognition, and behaviour on teacher and student behaviour, and students' out-of-software learning gains (Holstein et al., 2018b).

Among several other findings, the results indicated that students learned more when their teacher used Lumilo, compared with both business-as-usual and a condition in which teachers used a simpler form of classroom monitoring support (Holstein et al., 2018b). In particular, a teacher's use of real-time teacher analytics served as an equalizing force in the classroom. That is, whereas the use of AI tutors in K–12 has sometimes been shown in prior studies to *increase* achievement gaps — favouring students who are better prepared to learn from these systems (Rau, 2015; Reich & Ito, 2017) — teachers' use of real-time analytics in the classroom had the effect of *narrowing* the gap in learning outcomes across students of varying prior domain knowledge (Holstein et al., 2018b). This effect appeared to be mediated by a shift in the way teachers distributed their time across students during a class session. Without Lumilo, teachers distributed their time relatively evenly across students of varying prior domain knowledge. However, when using Lumilo, teachers allocated significantly more time towards students detected as struggling or exhibiting maladaptive learning behaviours — resulting in a dramatic shift of teachers' time and attention towards students with lower prior domain knowledge.

While our prior work reports on quantitative findings from these classroom pilots and experiments (Holstein et al., 2018a; 2018b), in the following we share observations from a year of deploying Lumilo "in the wild," with an emphasis on needs and nuances not captured our earlier design work with teachers. From the first classroom pilot onward, teacher and student responses to Lumilo were much more positive than we would have expected for an initial venture outside of the lab. By the time we entered the classroom, we had already encountered many "surprises" in our REs sessions by testing with diverse classroom datasets, and had iterated on Lumilo's design accordingly. In some cases, teachers found that particular design features that had emerged through iterative prototyping were *even more* useful in live classrooms than they had anticipated during REs. For example, one teacher who had participated in both an REs session and a set of pilot sessions with his own classes was frequently observed multi-tasking during a class session — using the glasses to peek at analytics for students across the room, and interleaving quick feedback between multiple students, even in the middle of working face-to-face with a particular student. This teacher reflected that he did not feel as strong a need to multi-task in the REs study, but in a live classroom where multiple students were constantly vying for the teacher's attention, *"[The ability to] take a student's screen with me, even if I'm over here working with another student is amazingly useful… that was well thought through."*

Teachers reported making frequent use of Lumilo's analytics to identify students who might need their help — in some cases, directly referencing these analytics in one-on-one conversation with a student: *"You have a smiley face [right now], but you're still having trouble with adding and subtracting variables from both sides ... That's what you need to watch, when you have variables on both sides, you need to subtract on both sides [not just one]."* In other cases, teachers used Lumilo's alerts in ways they had not anticipated during REs think-alouds. For example, in one class, a teacher noticed a "system misuse" alert over one student's head, with elaboration text indicating that this student seemed to be "Abusing hints?" However, the teacher believed that hint abuse was out of character, given what they knew about this student. When the teacher approached to find out how the student was doing, they learned that this student was actually colourblind and thus could not perceive the correctness feedback providing by the tutoring system (which was coded green for "correct" and red for "incorrect"). This had led the student to frustration and an apparent overreliance on the tutoring software's "hint" function. In a similar case, a teacher noticed that an otherwise diligent student had been idle in the software for over five minutes. The teacher approached this student and noticed that this student was playing online video games instead of doing the assigned work. The teacher asked how the student was *feeling* that day, which led the student to disclose that they had broken up with a significant other over the preceding weekend. In response, the teacher gave the student permission to take the day off from math, if needed.

Each class session began with the teacher introducing the glasses to the class, often accompanied by an invitation to laugh at the teacher's appearance while wearing the prototype (see Figure 9, left): *"Alright everyone get the giggles out now. You can laugh at me for the next five minutes, but after that it's time to work."* After each class, the teacher would invite their students to reflect on the experience, and provide design feedback from the student perspective. Overall, students also reacted very positively to teachers' use of Lumilo during ITS lab sessions. At the beginning of one class session, a student said, *"I'm a little afraid. [The researcher's] gonna let [the teacher] spy on us…"* During class, however, the student appeared to warm up to the idea. On multiple occasions, the teacher approached the student to provide (unsolicited) help based on what the teacher saw through the glasses. At the end of one of these student–teacher exchanges, after the student completed the majority of a problem without the teacher's assistance, the student and teacher high-fived. During the end-of-class reflection, the same student said, *"It was awesome how [the teacher] just knew when I needed help."* In another classroom, a student exclaimed, *"I want to be a teacher when I grow up!"*

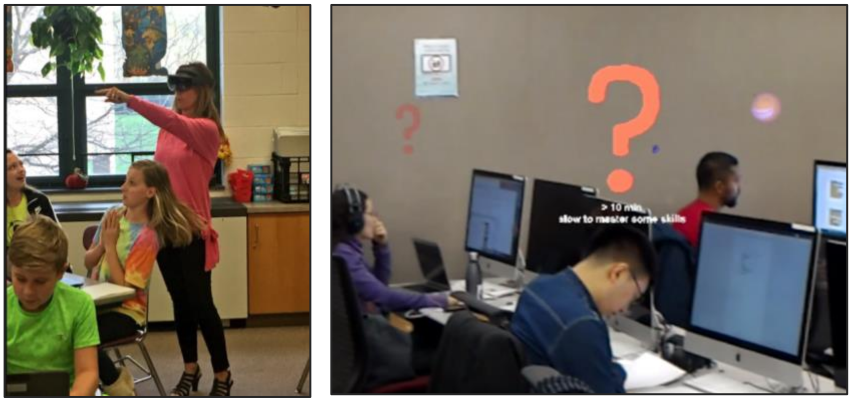


Figure 9. Left: A teacher using Lumilo while her students work with Lynnette, an ITS for equation solving, in class (from Holstein et al., 2018b); Right: A point-of-view screenshot through Lumilo.

Piloting Lumilo in 30 live middle-school classrooms also revealed several critical needs that Lumilo's design did not address. For example, both students and teachers across multiple classrooms emphasized needs for additional features to support "anonymous" non-face-threatening communication between students and teachers during a class session, beyond "invisible hand raises." In the absence of such features, students sometimes took matters into their own hands. For example, in one class session, a student used the equation-entry box in Lynnette's interface to write "secret" messages to their teacher, viewable through the teacher's glasses. In addition, while teachers reported that it rarely made sense to give a whole-class lecture given the non-synchronized nature of ITS class sessions, they realized that it would be useful to have more support in dynamically deciding between small group ("pull out") interventions versus interventions with individual students. While the Lumilo's design included analytics at the individual and whole-class levels, the design did not facilitate rapid filtering of students to identify relevant student subgroups. One teacher suggested it might have been helpful if *"the class [dashboard would let] you zoom in and see which students are struggling with a particular skill."*

Finally, after using Lumilo in the classroom, multiple teachers raised needs for greater transparency and control. For example, one teacher noted that although the analytics presented by Lumilo provide some insight into why the ITS might be making certain decisions (e.g., relating to adaptive problem selection), *"When a student asks me why they have to do twenty problems in level three [of the ITS], before [moving on], but another student only has to do two problems… I should be able to answer that."*

In addition, some teachers noted that the transparency provided by Lumilo helped make them more aware of some of the limitations of the intelligent tutoring system their class was using (and associated student modelling techniques). Given this enhanced awareness, teachers reported frustration over their relative lack of control (cf. Lee & Baykal, 2017). One teacher noted that it would be helpful if they could provide feedback when Lumilo's alerts miss the mark, to customize the alerts to their needs. Similarly, this teacher suggested that it would be nice if students had the option to see their own student model (including skill mastery estimates, metacognitive variables, and other information that the teacher can see through Lumilo) and contest what it says about their knowledge or behaviour (cf. Bull & Kay, 2016), perhaps allowing the teacher to review and approve these cases individually during class. Another teacher mentioned that when using Lumilo, they were seeing students struggle with the same issues over and over again, and the software's built-in hints did not seem to be helping in these cases. This teacher suggested that instead of "filling in" for the ITS by repeatedly giving different students the same feedback, *"It would be nice if [the ITS] could listen to what I tell [the student, and] just say that the next time a student gets stuck [with the same issue]."*

7. Discussion and Future Work

While recently proposed design models in LA encourage the active involvement of stakeholders at every stage of the design process (e.g., Martinez-Maldonado et al., 2016; Prieto-Alvarez et al., 2018), demonstrations of actual co-design processes in learning analytics remain rare. In this work, we have presented the first case study, to our knowledge, of how classroom teachers can be meaningfully involved throughout the entire design process of a complex LA tool. In addition to designing and deploying a specific LA tool, Lumilo, our co-design process has generated a rich *research agenda* for future "hybrid" systems that can adaptively leverage complementary strengths of human and AI instruction. In the process of designing Lumilo, we have also explored the development of a *new prototyping method*, REs, for LA tools and other data-driven algorithmic systems.

7.1. Methodological Reflections and Recommendations for Co-Designing LA Systems

Although recent work in the field of Learning Analytics encourages stakeholder involvement at every stage of the design/development of a LA tool — from early, generative design phases through piloting and evaluation in real-world educational contexts — demonstrations of successful end-to-end co-design processes for LA tools remain very rare in the literature. Furthermore, existing user-centred design workflows and frameworks (e.g., Dollinger & Lodge, 2018; Martinez-Maldonado et al., 2016) provide limited methodological guidance for effectively involving non-technical stakeholders at each phase of an LA design process. In the following, we present general recommendations for future LA co-design efforts, reflecting "lessons learned" and practices we have found valuable in the co-design process for Lumilo. We view our design process as broadly consistent with prior workflows in the LA literature, so here we focus on specific practices that go beyond those explicitly highlighted in prior work

1. **Begin with *stakeholder needs*, not analytics or visualizations.** In designing any tool, it is useful to begin with an understanding of the needs a tool might address. Yet design processes for LA tools often seem to begin by identifying technical solutions (e.g., particular data sources, analytic methods, and visualizations), and then searching for opportunities to apply these solutions (Rodriguez-Triana et al., 2017). In the early, generative phases of our design process, we explicitly avoided discussing particular solutions with teachers, to avoid limiting these conversations by teachers' conceptions of what is technologically possible. Instead, we found it much more

productive to explore teachers' current challenges and aspirations through probes like the "superpowers" exercise and through directed storytelling around teachers' lived experiences. Findings from such design exercises subsequently enabled us to "match make" between specific teacher information needs (e.g., real-time updates about specific student constructs) and current technical possibilities (e.g., existing analytic and student modelling methods intended to measure these constructs). In some cases, beginning from stakeholder needs led us away from the use of more abstract visualizations that are common in existing learning analytics dashboards — such as plots, graphs, and charts — towards the use of concrete, grounded representations of student data such as raw examples of student errors (cf. Bull & Kay, 2016).

2. **Regularly link analytics to action throughout the design process.** Although many learning analytics tools are designed to support awareness or reflection, the end goal of this enhanced awareness or support for reflection is commonly to support more informed *decision-making and action*. Throughout our design process, we found it highly useful to regularly link particular analytics to action. For example, in our early prototyping studies, we found that prompting teachers to reflect on what real-time *decisions* a particular information display might inform often led them to notice ways in which the display could be made more useful, usable, and/or trustworthy. In many cases, teachers would initially find particular visualizations interesting and appealing, but would change their minds when prompted to reflect on how they might actually use these visualizations to inform their classroom practice.

3. **Prototype specific user tasks and usage scenarios early and often.** In line with the previous recommendation, we have found it very useful to simulate specific user tasks and usage scenarios for an LA tool (e.g., by having stakeholders participate in role-playing and bodystorming exercises) as early and often as possible throughout the design and prototyping process. Such simulation exercises (combined with methods like think-alouds and cognitive task analyses) can help to surface information needs crucial for particular tasks or usage scenarios, but which users may not otherwise perceive or report "out of context." Since different usage scenarios for an LA tool can involve very different types of tasks and decisions (e.g., planning a lesson versus identifying students who need help right now), different constraints (e.g., time pressure), and different affordances for action, simulating specific usage scenarios can sometimes reveal that radically different LA display designs are needed for different scenarios.

4. **Prototype the behaviour of LA tools using multiple, diverse real-world datasets.** Finally, since the behaviour of LA systems can depend heavily on nuances of particular data-generating contexts, in combination with particular analytic methods/algorithms, we have found that it can be extremely informative to run prototyping sessions with diverse datasets. For example, by replaying data from real classrooms across a range of socioeconomic contexts and performance levels in REs sessions, we were able to anticipate several context-specific design challenges before entering live classrooms.

7.2. Designing New Prototyping Methods for Learning Analytics Systems

In the present work, we have introduced a new prototyping method: Replay Enactments (REs). We view REs as a step towards developing and formalizing a broader class of prototyping methods that can address the unique challenges of designing and prototyping complex, data-driven algorithmic systems. Moving forward, we expect that methods for prototyping with authentic algorithms and data from diverse data-generating contexts, will be invaluable in designing systems that are usable, useful, fair, and trustworthy.

A promising direction for future work is to explore how replay-based prototyping methods like REs might be refined to further structure participant feedback. While the current version of REs has participants engage in a relatively unstructured think-aloud while performing a task, future refinements might focus user attention on specific aspects of a data-driven algorithmic system's design; for example, to test the usefulness of particular forms of AI "explanations" in the context of specific user tasks (cf. Doshi-Velez & Kim, 2017; Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, & Wallach, 2018) or to support the discovery of spurious and undesirable biases in a system's behaviour (cf. Holstein, Wortman Vaughan et al., 2018). Similarly, instead of replaying a full class session, future work on replay-based prototyping methods like REs might explore methods to curate specific scenarios (i.e., data clips) with properties desirable for answering certain kinds of questions a research/design team may have. A potential tradeoff in developing more structured methods such as these is that this upfront structuring and curation may reduce opportunities for unexpected design findings to emerge (Odom et al., 2012). As such, a mix of less- and more-heavily structured studies may be desirable.

In order to test the interacting dynamics of specific data contexts, algorithms, visualizations, and human judgments in a simulated task context, REs differs from related prototyping methods like User Enactments and experience prototyping by requiring fairly heavy upfront investment in technical development. To a certain extent, higher upfront technical investment may be unavoidable when prototyping complex, data-driven algorithmic systems (Dove et al., 2017). However, promising direction for future work may be to explore lighter-weight prototyping methods that reap *some* of the benefits of REs earlier on in the design process. Finally, a fruitful direction for future work may be to explore new methods that can provide earlier insight into *social nuances* before deploying a system in the real world, while keeping development and recruitment costs low — perhaps by including multiple participants in role-playing exercises, or possibly a mix of multiple live and replayed participants.

8. Conclusions

The field of Learning Analytics (LA) is just beginning to explore what it might mean to co-design LA tools with stakeholders such as classroom teachers and learners (Dollinger & Lodge, 2018; Holstein et al., 2017b; Holstein, Hong et al., 2018; Prieto-Alvarez et al., 2018). In this article, we have presented an end-to-end case study and methodological guidance on how classroom teachers can be meaningfully involved throughout the entire design process of a complex LA tool. It is our hope that as the field of Learning Analytics matures, we will see more demonstrations of the benefits, opportunities, and challenges of co-designing LA tools with teachers, students, parents, and other critical stakeholders.

Declaration of Conflicting Interest

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

Funding

This work was supported by Grants R305A180301 and R305B150008 from the Institute of Education Sciences (IES) and Grant #1530726 from the National Science Foundation (NSF). Any opinions are those of the authors and do not represent the views of the NSF, IES, or the U.S. Department of Education.

Acknowledgments

We thank our anonymous reviewers and journal editors for their helpful feedback. Special thanks to all participating teachers, students, and schools, as well as Jodi Forlizzi, Gena Hong, Judy Kay, Mary Beth Kery, Roberto Martinez-Maldonado, Octav Popescu, Ido Roll, Peter Schaldenbrand, Jonathan Sewall, Mera Tegene, Cindy Tipper, Jasper Tom, Kalina Yacef, and Zac Yu.

References

Aguilar, S. J. (2018). Examining the relationship between comparative and self-focused academic data visualizations in at-risk college students' academic motivation. *Journal of Research on Technology in Education*, *50*(1), 84–103. http://dx.doi.org/10.1080/15391523.2017.1401498

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education, 26*(1), 205–223. http://dx.doi.org/10.1007/s40593-015-0089-1

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614. http://dx.doi.org/10.1007/s40593-016-0105-0

Baumer, E. P. (2017). Toward human-centered algorithm design. *Big Data & Society*, *4*(2). http://dx.doi.org/10.1177/2053951717718854

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA. (pp. 431–440). Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-39112-5_44

Beyer, H., & Holtzblatt, K. (1997). Contextual design: Defining customer-centered systems. Amsterdam, Netherlands: Elsevier.

Black, P., & Harrison, C. (2001). Feedback in questioning and marking: The science teacher's role in formative assessment. *School Science Review*, *82*(301), 55–61.

Bonsignore, E., DiSalvo, B., DiSalvo, C., & Yip, J. (2017). Introduction to participatory design in the learning sciences. In B. DiSalvo, J. Yip, E. Bonsignore, & C. DiSalvo (Eds.), *Participatory Design for Learning* (pp. 15–18). Abingdon-on-Thames, UK: Routledge.

Buchenau, M., & Suri, J. F. (2000). Experience prototyping. In *Proceedings of the 3rd Conference on Designing Interactive*

Systems: Processes, Practices, Methods, and Techniques (DIS '00), 17–19 August 2000, New York City, NY, USA (pp. 424–433). New York: ACM. http://dx.doi.org10.1145/347642.347802

Bull, S., & Kay, J. (2016). SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, *26*(1), 293–331. http://dx.doi.org/10.1007/s40593-015-0090-8

Cairns, P., & Cox, A. L. (Eds.). (2008). *Research methods for human–computer interaction* (Vol. 12). Cambridge, UK: Cambridge University Press.

Davidoff, S., Lee, M. K., Dey, A. K., & Zimmerman, J. (2007). Rapidly exploring application design through speed dating. In J. Krumm, G. D. Abowd, A. Seneviratne, & T. Strang (Eds.), *Proceedings of the International Conference on Ubiquitous Computing* (UbiComp 2007), *Lecture Notes in Computer Science*, vol. 4717 (pp. 429–446). Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-74853-3_25

Dennerlein, S., Kowald, D., Pammer-Schindler, V., Lex, E., & Ley, T. (2018). Simulation-based co-creation of algorithms. In *CEUR Workshop Proceedings* (Vol. 2190). RWTH Aachen University.

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1–2), 9–38. http://dx.doi.org/10.1007/s11257-011-9106-8

Diana, N., Eagle, M., Stamper, J. C., Grover, S., Bienkowski, M. A., & Basu, S. (2017). Automatic peer tutor matching: Data-driven methods to enable new opportunities for help. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 372–373). International Educational Data Mining Society.

DiSalvo, B., & DiSalvo, C. (2014). Designing for democracy in education: Participatory design and the learning sciences. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, & L. D'Amico (Eds.), *Learning and Becoming in Practice: Proceedings of the International Conference of the Learning Sciences* (ICLS '14), 23–27 June 2014, Boulder, CO, USA (Vol. 2, pp. 793–799). International Society of the Learning Sciences. https://dx.doi.org/10.22318/icls2014.793

Dollinger, M., & Lodge, J. M. (2018). Co-creation strategies for learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 97–101). New York: ACM. http://dx.doi.org/10.1145/3170358.3170372

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 6–11 May 2017, Denver, Colorado, USA (pp. 278–288). New York: ACM. http://dx.doi.org/10.1145/3025453.3025739

Evenson, S. (2006). Directed storytelling: Interpreting experience for design. In A. Bennett (Ed.), *Design Studies: Theory and research in graphic design* (pp. 231–240). Hudson, NY: Princeton Architectural Press.

Halloran, J., Hornecker, E., Fitzpatrick, G., Weal, M., Millard, D., Michaelides, D., Cruickshank, D., & De Roure, D. (2006). Unfolding understandings: Co-designing UbiComp in situ, over time. In *Proceedings of the 6th Conference on Designing Interactive Systems* (DIS '06), 26–28 June 2006, University Park, PA, USA (pp. 109–118). New York: ACM. http://dx.doi.org/ 10.1145/1142405.1142423

Hanington, B., & Martin, B. (2012). Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions. London: Rockport Publishers.

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. http://dx.doi.org/10.1007/s40593-014-0024-x

Hoadley, C. (2017). How participatory design has influenced the learning sciences. In B. DiSalvo, J. Yip, E. Bonsignore, & C. DiSalvo (Eds.), *Participatory design for learning* (pp. 34–39). Abingdon-on-Thames, UK: Routledge.

Holstein, K., Hong, G., Tegene, M., McLaren, B. M., & Aleven, V. (2018). The classroom as a dashboard: Co-designing wearable cognitive augmentation for K–12 teachers. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 79–88). New York: ACM. http://dx.doi.org/10.1145/3170358.3170377

Holstein, K., McLaren, B. M., & Aleven, V. (2017a). SPACLE: Investigating learning across virtual and physical spaces using spatial replays. In *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 358–367). New York: ACM.

http://dx.doi.org/10.1145/3027385.3027450

Holstein, K., McLaren, B. M., & Aleven, V. (2017b). Intelligent tutors as teachers' aides: Exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 257–266). New York: ACM. http://dx.doi.org/10.1145/3027385.3027451

Holstein, K., McLaren, B. M., & Aleven, V. (2018a). Informing the design of teacher awareness tools through causal alignment analysis. In J. Kay & R. Luckin (Eds.), Rethinking Learning in the Digital Age: Making the Learning Sciences Count. Proceedings of the 13th International Conference of the Learning Sciences (ICLS '18), 23–27 June 2018, London, UK (Vol. 1, pp. 104–111). International Society of the Learning Sciences.

Holstein, K., McLaren, B. M., & Aleven, V. (2018b). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In C. Penstein Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (AIED 2018), 27–30 June 2018, London, UK. (pp. 154–168). Springer, Cham. http://dx.doi.org/doi.org/10.1007/978-3-319-93843-1_12

Holstein, K., Wortman Vaughan, J., Daumé, H. III, Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 4–9 May 2019, Glasgow, Scotland, UK, Paper No. 600. New York: ACM. https://dx.doi.org/10.1145/3290605.3300830

Holstein, K., Yu, Z., Sewall, J., Popescu, O., McLaren, B. M., & Aleven, V. (2018). Opening up an intelligent tutoring system development environment for extensible student modeling. In C. Penstein Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (AIED 2018), 27–30 June 2018, London, UK (pp. 169–183). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-93843-1_13

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining*, *10*(1), 36–71. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/210

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In S. Kambhampati (Ed.), *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (IJCAI-16), 9–15 July 2016, New York, NY, USA (pp. 4070–4073). Palo Alto, CA: AAAI Press/International Joint Conferences on Artificial Intelligence.

Käser, T., Klingler, S., & Gross, M. (2016). When to stop? Towards universal instructional policies. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 289–298). New York: ACM. http://dx.doi.org/10.1145/2883851.2883961

Khachatryan, G. A., Romashov, A. V., Khachatryan, A. R., Gaudino, S. J., Khachatryan, J. M., Guarian, K. R., & Yufa, N. V. (2014). Reasoning mind genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education*, *24*(3), 333–382.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining* (pp. 43–56). Boca Raton, FL: Chapman & Hall/CRC.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. http://dx.doi.org/10.3102/0034654315581420

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW 2017) 25 February–1 March 2017, Portland, OR, USA (pp. 1035–1048). New York: ACM. http://dx.doi.org/10.1145/2998181.2998230.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., Noothigattu, R., See, D., Lee, S., Psomas, C. A., & Procaccia, A. (2018). WeBuildAI: Participatory framework for fair and efficient algorithmic governance. Pre-print.

Lipton, Z. C. (2016). The mythos of model interpretability. Paper presented at the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), 23 June 2016, New York, NY, USA. *arXiv preprint arXiv:1606.03490*.

Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2015). MTFeedback: Providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Transactions on Learning Technologies*, *8*(2), 187–200. http://dx.doi.org/10.1109/TLT.2014.2365027

Martinez-Maldonado, R. M., Kay, J., Yacef, K., & Schwendimann, B. (2012). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K.-K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (ITS 2012), 14–18 June 2012, Chania, Greece (pp. 482–492). Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-30950-2_62

Martinez-Maldonado, R., Pardo, A., Mirriahi, N., Yacef, K., Kay, J., & Clayphan, A. (2016). LATUX: An iterative workflow for designing, validating and deploying learning analytics visualisations. *Journal of Learning Analytics*, *2*(3), 9–39. http://dx.doi.org/10.1007/978-3-642-30950-2_62

Mavrikis, M., Gutierrez-Santos, S., Geraniou, E., Noss, R., & Poulovassilis, A. (2013). Iterative context engineering to inform the design of intelligent exploratory learning environments for the classroom. In R. Luckin, S. Puntambekar, P. Goodyear, B. Grabowski, J. Underowood, & N. Winters (Eds.), *Handbook of design in educational technology* (pp. 80–92). Abingdon-on-Thames, UK: Routledge.

Mavrikis, M., Gutierrez-Santos, S., & Poulovassilis, A. (2016). Design and evaluation of teacher assistance tools for exploratory learning environments. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 168–172). New York: ACM. http://dx.doi.org/10.1145/2883851.2883909

Moraveji, N., Li, J., Ding, J., O'Kelley, P., & Woolf, S. (2007). Comicboarding: Using comics as proxies for participatory design with children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07), 28 April–3 May 2007, San Jose, CA (pp. 1371–1374). New York: ACM.

Odom, W., Zimmerman, J., Davidoff, S., Forlizzi, J., Dey, A. K., & Lee, M. K. (2012). A fieldwork of the future with user enactments. In *Proceedings of the Designing Interactive Systems Conference* (DIS '12) 11–15 June 2012, Newcastle Upon Tyne, UK (pp. 338–347). New York: ACM. http://dx.doi.org/10.1145/2317956.2318008

Ogan, A., Walker, E., Baker, R. S., Rebolledo Mendez, G., Jimenez Castro, M., Laurentino, T., & De Carvalho, A. (2012). Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 5–10 May 2012, Austin, TX, USA (pp. 1381–1390). New York: ACM. http://dx.doi.org/10.1145/2207676.2208597

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*, *25*(2), 229–248. http://dx.doi.org/10.1007/s40593-014-0034-8

Olsen, J. K. (2017). *Orchestrating Combined Collaborative and Individual Learning in the Classroom.* Unpublished doctoral dissertation, Carnegie Mellon University.

Oulasvirta, A., Kurvinen, E., & Kankainen, T. (2003). Understanding contexts by being there: Case studies in bodystorming. *Personal and Ubiquitous Computing*, *7*(2), 125–134.

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. *Research and Practice in Technology Enhanced Learning*, *2*(1), 51–74. http://dx.doi.org/10.1142/S1793206807000300

Penuel, W. R., & Yarnall, L. (2005). Designing handheld software to support classroom assessment: Analysis of conditions for teacher adoption. *The Journal of Technology, Learning and Assessment*, *3*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1658

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Prieto-Alvarez, C. G., Martinez-Maldonado, R., & Anderson, T. (2018). Co-designing learning analytics tools with learners. *Learning analytics in the classroom: Translating learning analytics research for teachers.* Abingdon-on-Thames, UK: Routledge.

Rau, M. A. (2015). Why do the rich get richer? A structural equation model to test how spatial skills affect learning with representations. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 350–357). International Educational Data Mining Society.

Reich, J., & Ito, M. (2017). *From good intentions to real outcomes: Equity by design in learning technologies*. Irvine, CA: Digital Media and Learning Research Hub.

Ritter, S., Carlson, R., Sandbothe, M., & Fancsali, S. E. (2015). Carnegie Learning's adaptive learning products. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 633–634). International Educational Data Mining Society.

Ritter, S., Yudelson, M., Fancsali, S., & Berman, S. R. (2016). Towards integrating human and automated tutoring systems. In T. Barnes et al. (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (EDM2016), 29 June–2 July 2016, Raleigh, NC, USA (pp. 626–627). International Educational Data Mining Society.

Rodriguez-Triana, M. J., Prieto Santos, L. P., Vozniuk, A., Shirvani Boroujeni, M., Schwendimann, B. A., Holzer, A. C., & Gillet, D. (2017). Monitoring, awareness and reflection in blended technology enhanced learning: A systematic review. *International Journal of Technology Enhanced Learning, 9,* 126–150. http://dx.doi.org/10.1504/IJTEL.2017.084489

Schofield, J. W., Eurich-Fulcer, R., & Britt, C. L. (1994). Teachers, computer tutors, and teaching: The artificially intelligent tutor as an agent for classroom change. *American Educational Research Journal, 31(3),* 579–607. http://dx.doi.org/10.2307/1163227

Schuler, D., & Namioka, A. (Eds.). (1993). *Participatory design: Principles and practices.* Boca Raton, FL: CRC Press.

Shepard, L. A. (1997). *Insights gained from a classroom-based assessment project*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Sujan, M., & Pasquini, A. (1998). Allocating tasks between humans and machines in complex systems. In *Proceedings of the 4th International Conference on Achieving Quality in Software* (AQuIS '98), 30 March–2 April 1998, Venice, Italy (pp. 173–184).

Tohidi, M., Buxton, W., Baecker, R., & Sellen, A. (2006). User sketches: A quick, inexpensive, and effective way to elicit more reflective user feedback. In *Proceedings of the 4th Nordic Conference on Human–Computer Interaction: Changing Roles* (NordiCHI 2006), 14–18 October 2006, Oslo, Norway (pp. 105–114). New York: ACM. http://dx.doi.org/10.1145/1182475.1182487

Veitch, J., Salmon, J., & Ball, K. (2007). Children's active free play in local neighborhoods: A behavioral mapping study. *Health Education Research*, *23*(5), 870–879. http://dx.doi.org/10.1093/her/cym074

Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (1998). *An introduction to human factors engineering*. New York: Longman.

Wright, P., Dearden, A., & Fields, B. (2000). Function allocation: A perspective from studies of work practice. *International Journal of Human–Computer Studies*, *52*(2), 335–355. http://dx.doi.org/10.1006/ijhc.1999.0292

Yacef, K. (2002). Intelligent teaching assistant systems. In *Proceedings of the International Conference on Computers in Education* (ICCE 2002), 3–6 December 2002, Auckland, New Zealand (pp. 136–140). IEEE Computer Society. http://dx.doi.org/10.1109/CIE.2002.1185885

Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018, November). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human–Computer Interaction*, CSCW issue, Vol. 2, article #194. http://dx.doi.org/10.1145/3274463