SPACLE: Investigating learning across virtual and physical spaces using spatial replays

Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{kjholste, bmclaren, aleven}@cs.cmu.edu

ABSTRACT
Classroom experiments that evaluate the effectiveness of educational technologies do not typically examine the effects of classroom contextual variables (e.g., out-of-software help-giving and external distractions). Yet these variables may influence students' instructional outcomes. In this paper, we introduce the Spatial Classroom Log Explorer (SPACLE): a prototype tool that facilitates the rapid discovery of relationships between within-software and out-of-software events. Unlike previous tools for retrospective analysis, SPACLE replays moment-by-moment analytics about student and teacher behaviors in their original spatial context. We present a data analysis workflow using SPACLE and demonstrate how this workflow can support causal discovery. We share the results of our initial replay analyses using SPACLE, which highlight the importance of considering spatial factors in the classroom when analyzing ITS log data. We also present the results of an investigation into the effects of student-teacher interactions on student learning in K-12 blended classrooms, using our workflow, which combines replay analysis with SPACLE and causal modeling. Our findings suggest that students' awareness of being monitored by their teachers may promote learning, and that "gaming the system" behaviors may extend outside of educational software use.

CCS Concepts
• **Applied computing ~ Computer-assisted instruction**
• **Human-centered computing ~ Activity centered design**

Keywords
Intelligent tutoring systems, visualizations, causal modeling, blended learning, classroom, teachers

1. INTRODUCTION
In recent years, there has been a growing interest within the learning analytics and educational data mining communities in multi-modal learning analytics: the collection and integrated analysis of diverse data streams (e.g., computer log files, motion sensor logs, field observations, and audio recordings) to obtain a richer picture of student learning (e.g., [32, 33, 43]). Some of this work has focused on blended learning contexts – introducing methods for measuring and studying learning-related interactions that cross physical and virtual spaces. For example, Baker *et al.* studied relationships between students' behavior patterns within educational software and their interactions with peers and teachers in the physical classroom by analyzing computer log streams that were synchronized with quantitative field observations [13].

In parallel, there have been recent calls for added rigor in the design of learning analytics tools for teachers and students. If monitoring, awareness, and reflection tools for classrooms are to be effective, the design of these tools will likely benefit from a theoretically and empirically informed understanding of the causal mechanisms by which they could positively impact student learning [30, 33]. This may include, for example, understanding the dynamics of learning environments in which such tools will be used, and identifying any existing teacher or student practices with which they may conflict [32, 46]. In addition, a better understanding of the nature and effects of both teacher and student decision-making in such environments will likely be essential to the design of tools that can promote more effective decision-making [33, 37].

We are currently designing real-time learning analytics tools to help K-12 teachers more effectively guide their students as they work with adaptive educational technologies in the classroom. To inform our design process, we wish to examine the effects that teacher-student interactions have on student learning over relatively short time periods (seconds to hours, corresponding to Newell's 'cognitive' and 'rational' bands of action [44]).

To this end, we introduce the Spatial Classroom Log Explorer (SPACLE), a prototype tool that facilitates the discovery of relationships between teacher activity, classroom layout, and student learning and behavior in blended classrooms. Unlike existing tools for retrospective analysis of blended class sessions (e.g., [1, 15]), SPACLE visualizes user-selected, moment-by-moment analytics about student and teacher behaviors within their spatial context. Although students' out-of-software behaviors and spatial positions in the classroom are very rarely considered in analyses of log streams from educational software, there is reason to suspect that spatial factors may impact learning (e.g., [21]). Some existing tools allow researchers to quickly alternate between analyzing software log data and examining webcam or screen recordings from students' computers (e.g., [1]). While such tools can reveal rich features of individual and small-group learning sessions (including on-task conversations and student affect), they are not designed to reveal broader relationships between the spatiotemporal dynamics of the classroom and students' learning. SPACLE enables researchers to visualize moment-by-moment analytics about both out-of-software interactions and students' current learning and behavioral states (as computed from software logs) over a spatial map of the classroom in which the data was

ACM 978-1-4503-4870-6/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027450

collected. By visualizing a relatively small set of features within an interactive replay of a class session, SPACLE may enable faster detection of qualitative patterns *across* students than would be feasible from higher-dimensional stimuli such as live classroom observations or video recordings. And like other forms of log replay – including screen replays [28, 39] and low-fidelity text-based replays [40] of students' interactions within educational software – SPACLE allows researchers to examine different sets of analytics across multiple replay sessions, using the same data set.

In this paper, we illustrate how SPACLE supports analysis of classroom behaviors and provide initial findings from an investigation of student-teacher interactions in classrooms using intelligent tutoring systems. These findings suggest, for instance, that student "gaming the system" behaviors [13, 34] may extend to interactions occurring outside of the software. Finally, we discuss how SPACLE can inform the evidence-based design of real-time monitoring and awareness tools for teachers working in blended classrooms.

2. BACKGROUND

Intelligent tutoring systems (ITSs) are advanced learning technologies that allow students to work at their own pace: providing step-by-step guidance during complex problem-solving practice and continuously adapting instruction to students' current state (a set of measured variables, which may include estimates of student knowledge, affective states, metacognitive skills, and more) [10, 25]. Several meta-reviews have indicated that ITSs can enhance student learning in classroom settings, compared with traditional classroom instruction and other forms of educational technology [17, 19, 20]. Another key advantage of adaptive educational technologies such as ITSs may be that, when they are used in classrooms, they can free the teacher to move throughout the classroom and provide one-on-one support to students who may need it most [8, 23].

ITSs also generate a wealth of data from students' interactions within the software, which have enabled fine-grained process analyses of student learning and behavior. For example educational data mining methods have been used to study the effects of students' off-task and gaming-the-system behaviors (e.g., [6, 13]), cognitive and affective states such as frustration, concentration, and confusion (e.g., [12]), and various micro-level features of ITSs' instructional design (e.g., [7]) on student learning with these systems.

Although ITSs are often designed for use in K-12 schools, the ITS literature has rarely studied the effects of elements of classroom context on students' learning with ITSs [1, 24]. For example: recent field studies suggest that a large proportion of K-12 students' help-seeking behavior, when using ITSs in classrooms, may occur entirely outside of the software; but the ITS literature tends to study the effects of within-software aspects of students' help-related behaviors rather than out-of-software behaviors such as asking a teacher for help [4, 6]. While "in-vivo" classroom studies aim to study the effectiveness of ITSs in the presence of contextual variables that are likely to be present in real-world classrooms (e.g., help from a teacher or peer, external distractions affecting individuals or groups of students, collaboration between students, etc.), they do not typically measure the effects of the contextual variables themselves – instead treating these as noise [26, 27, 37] (though see [1, 13]).

There is reason to expect, however, that some of these contextual variables may be important mediators of student learning. In particular, gaining a better understanding of the effects of teacher-student interactions in classrooms using ITSs may be crucial to understanding these systems' effectiveness in real-world contexts. For example, a large-scale, two-year evaluation study of Carnegie Learning's Algebra I tutor suggested that variability in the out-of-software support teachers provided to students may have been at least partly responsible for inconsistent results across evaluation years [29]. Similarly, recent work has found that the extent to which teachers override ITSs' built-in, mastery learning based problem selection may negatively impact student learning [31].

3. EXPLORING MULTIMODAL CLASSROOM DATASETS THROUGH LOG REPLAY

Even when an "in-vivo" classroom study is primarily designed to test preconceived hypotheses (e.g. the effectiveness of a particular educational technology design), researchers sometimes collect qualitative classroom observations during the course of the study. These observations can allow researchers to gain a richer picture of what went on during a given class session, which may in turn help explain and interpret study results. Often times, these qualitative observations lead to unexpected discoveries, which can later be investigated more thoroughly through targeted follow-up experiments and observation sessions or educational data mining. For example, classroom observations of students' misuse of ITSs inspired a line of experimental and data mining work dedicated to uncovering the underlying causes behind these behaviors, as well as design work dedicated to intervening on these underlying causes [34]. Similarly, our current work was originally inspired by informal classroom observations of teachers' interactions with their students during lab sessions, in the context of in-vivo experiments that were not intended to study (and did not explicitly consider the effects of) teacher-student interactions.

We have developed SPACLE[1] to extend this observation process, by enabling researchers to interactively re-examine classroom ITS-use sessions, within a virtual map of the classroom layout (c.f. [9]), while visualizing moment-by-moment analytics about individual students. SPACLE replays are multimodal in the sense that they combine multiple data streams – visualizing both analytics about students' out-of-software interactions (e.g., whether or not a student is raising her/his hand, talking to a peer, or talking to the teacher), and analytics generated from students' interactions within the software, such as whether students are inactive, abusing the tutor's help functions [6], making frequent careless errors [11], "stuck" on a current activity [5], confused, frustrated, or engaged in their current task [2].

In each replay session, SPACLE allows researchers to specify the analytics they would like to examine about the teacher, the students, and/or any summary information they would like to display at the class level (e.g. the percentage of the class that is "stuck" on their current task at a given time). These analytics can be implemented as custom plugin scripts, subject to minimal input and output constraints. Then, given a map of the classroom layout where observations took place, as well as a mapping from student identifiers to their seating positions within the classroom (both of which may be obtained, in approximate form, by asking a teacher to provide a printed or hand-drawn copy of the seating chart), SPACLE can generate visual replays that preserve potentially important spatial information.

[1] Available at: https://github.com/d19fe8/SPACLE

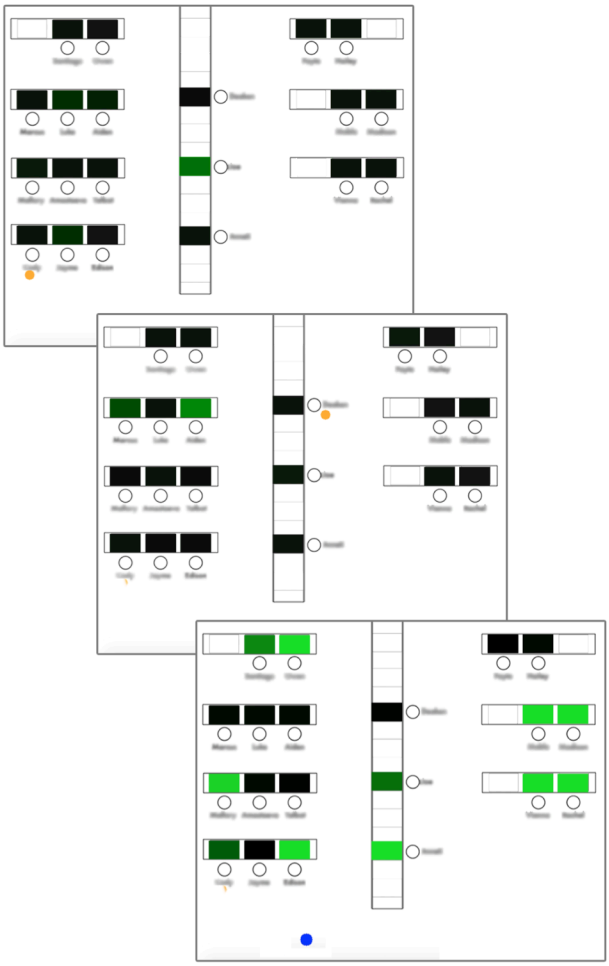

Figure 1. A sequence of screenshots from a replay of a class session generated by SPACLE. In the displayed classroom there is a long row of desks in the center of the room, oriented vertically, and several horizontal rows of desks on either side of it. Student names are obscured in this image. Students' inactive time, ranging from 30 seconds or less (black) to 90 seconds or more (bright green), is visualized on their "computer screens". The teacher circle takes on two colors (orange: on-task conversation with a student, blue: inactive/distracted or off-task conversation).

Specifically, researchers can import a class roster and an image (e.g., a scanned drawing) of a classroom layout into SPACLE, and then construct a virtual map of the classroom within the interface, by dragging, rotating, and resizing graphical representations of students (which are automatically generated, and pre-labeled, based on the class roster) into place, using the image as a guide. Each student is represented as a small circle with a rectangle directly above it (representing the student's computer screen) and a name or other identifier directly below it. Researchers can then choose to visualize moment-by-moment analytics about students by assigning certain analytics to appear either in students' circles (typically used to visualize out-of-software behaviors such as hand raising), or on their "computer screens" (typically used to visualize analytics about within-software interactions). In addition, if analytics about teacher behavior are present in a synchronized dataset, these can be visualized via a free-floating circle, which can change position on the map to represent the teacher's location in the classroom at a given time.

Aside from teacher position, all other analytics are visualized through color. For continuous-valued or ordinal analytics, colors can be assigned to two arbitrary end points within the range of values a given metric can assume, and these analytics will be visualized by interpolating between the two colors. For categorical analytics, colors can be assigned individually to different categories. Figure 1 shows a series of screenshots from a replay session (showing time slices several minutes apart). In this replay, the time elapsed since each student's last within-software interaction is displayed on their "computer screens", with end points of 30 seconds and 90 seconds. So, if a student has spent 30 or fewer seconds inactive, that student's screen will appear black, and if the student has spent 90 seconds or more inactive, the screen will appear bright green. In between 30 and 90 seconds of inactive time, a student's screen will appear to gradually transition from black to green. The teacher's position and current activities are also visualized in this replay, with "on-task conversation" indicated by an orange circle, and "inactive/distracted or off-task conversation" indicated by a blue circle. What is striking is the amount of inactivity in the third frame, during a period when the teacher is inactive, in the back of the classroom.

By examining a limited set of variables within a single replay session, researchers may be able to detect qualitative patterns across multiple students more rapidly than would be possible by watching video recordings or conducting live classroom observations [35]. And by visualizing different sets of analytics across multiple replay sessions, researchers can iteratively explore questions about potential mediators of student learning and behavior within the software. After formulating hypotheses based on replay analyses of a small number of classrooms, researchers can investigate further through quantitative modeling on larger samples. In addition to facilitating interactive replays, SPACLE can generate synchronized datasets that enable educational data mining techniques to be easily applied.

SPACLE is currently designed to work with ITS log data from DataShop, an open repository for data from educational technologies that currently houses over 700 data sets, many of which have been used in secondary analyses.[2] Prior to generating replays, SPACLE first synchronizes records of out-of-software events in the classroom (e.g. student and teacher behaviors, or class-level disruptions) with log data generated from students' interactions within the software. The records of out-of-software behaviors may be generated by hand (i.e., field observations conducted by human observers), or, in the future, via automated means such as sensors placed throughout the classroom (e.g. [21, 32]) or machine-learned detectors that attempt to infer out-of-software behaviors from ITS log data (e.g. [3]). The primary requirements SPACLE imposes on these out-of-software logs are that they either include continuous measurements (e.g. moment-by-moment recordings of a teacher's location and movements in the classroom) or discrete observations marked with approximate start and end times for a given behavior.

In our work thus far we have focused on using SPACLE to better understand and interpret the effects of classroom dynamics on student learning with ITSs – though we have also begun exploring broader uses of SPACLE as a design tool (see Discussion). In the next section, we illustrate how we've used SPACLE as a bridge

[2] https://pslcdatashop.web.cmu.edu

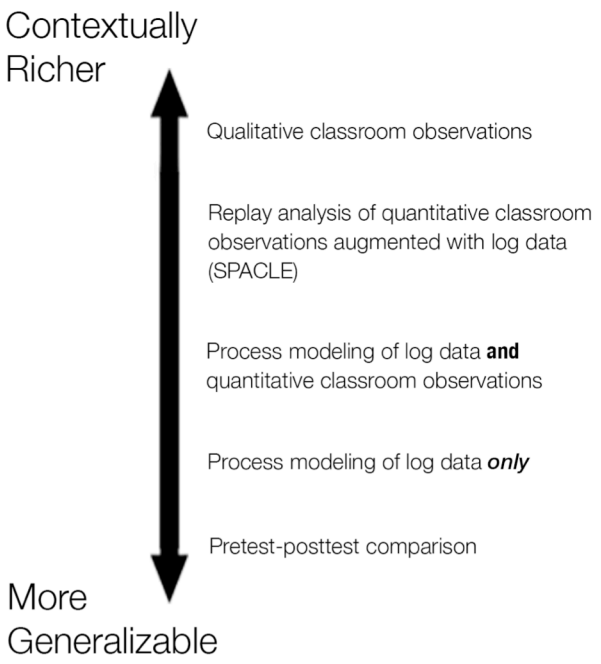

Figure 2. A spectrum of methods for understanding student learning in classrooms using educational technologies.

between qualitative analysis of classroom observation data and larger-scale data mining, in our own early investigations into potential effects of teacher behavior in ITS classrooms. After confirming that the ITS used in our study was effective overall (via analysis of students' pre- and posttest scores), we began exploring potential mediators of student learning (both within and outside of the software). Over the course of these explorations, we have gradually moved from contextually richer methods (classroom observations and replay analysis on small samples) to more generalizable methods (correlational analyses and causal modeling on larger samples), and then back again to aid in interpretation and additional exploration (see Figure 2).

4. CASE STUDY

4.1 Data Collection

The data we report in this study were originally collected during a classroom experiment aimed at evaluating how analytics generated from students' interactions with an ITS, presented on a prototype teacher dashboard, could help teachers plan more effective lectures for subsequent class sessions. However, the data analyzed in this paper are from a class period during which students worked with ITSs and teachers did not yet have access to a dashboard. Thus, these teachers often relied on direct observations of their students' computer screens, while walking around the classroom, in order to monitor their students' progress. This is a typical situation when teachers use ITSs in their classes.

In this study, 299 middle school students used *Lynnette*, an ITS for algebraic equation solving [14, 16], for 60 minutes, spread across up to two class sessions. Students' performance in equation solving was measured before and after using *Lynnette* via computer-based pre- and posttests, which were focused on measuring procedural skills. We used two test forms, which were identical up to the particular numbers used in equations. Test forms were presented in counterbalanced order across pre- and post-test. Test items were graded automatically, based on the correctness of students' final responses (i.e. without providing partial credit for intermediate steps in equation solving).

We collected live classroom observations from a sample of 9 out of 17 classrooms taught by 4 teachers, with a total of 151 students. Students who were absent during any of the pretest, ITS-use sessions, or posttest were excluded from subsequent analyses, leaving 135 students in total. In the remainder of this paper, only data from these 135 students are considered. Due to privacy concerns, we were unable to video record class sessions. Instead, during each class session, a member of our research team sat in the back of the classroom (in order to minimize any disturbance caused by their presence) and recorded coarse-grained field observations of teacher and student behavior. We recorded observations using *LookWhosTalking*[3], a tool for coding live classroom observations, developed at our institution, which was customized with a coding scheme we designed to facilitate both coding and eventual analyses. This coding scheme was adapted from the Baker-Rodrigo observation method protocol (BROMP) [42], and the TA observation protocol developed by Stang *et al.* [18]. We extend the TA observation protocol by distinguishing between different types of teacher interactions with students – namely, distinguishing whether a teacher is monitoring/observing a student or holding a conversation with that student, and further distinguishing between *on-task* and *off-task* teacher-student conversations (c.f. [13, 42]).

Following BROMP, up-to-date seating charts were elicited from a teacher prior to each class session, both to enable coding of student-teacher interactions, and for use as classroom maps during replay analysis [42]. Field observers recorded instances in which students raised or lowered their hands, and coded teacher behavior with reference to 6 broad categories:

1. **On-task conversation:** The teacher is engaged in a discussion with a student about the activity she/he is currently working on
2. **Off-task conversation:** The teacher is engaged in an unrelated discussion with the student.
3. **Talking to class:** The teacher is addressing the entire class (e.g., giving a "mini-lecture" based on observations made during a lab session)
4. **Monitoring:** The teacher is watching the class from a fixed location (e.g., the teacher's desk), or standing behind a student and scanning that student's computer screen over her/his shoulder (disambiguated by the teacher's current location, as described below)
5. **Outside the room:** the teacher is not in the classroom
6. **Inactive:** the teacher is in the classroom, but engaged in an activity other than one of the above (e.g., grading papers or checking email)

Within each of the broad behavior categories above, the position of the teacher in the classroom was recorded if the behavior persisted for at least two seconds. The teacher's position was coded either as the name of a student the teacher was standing behind (if the teacher was directly monitoring that student, or engaged in an on-task conversation), or a description of another location in the classroom, such as the teacher's desk. These field observations were then synchronized offline, using SPACLE, with the DataShop log data generated by Lynnette.

[3] Available at: https://bitbucket.org/dadamson/lookwhostalking

4.2 Analyses and Results

4.2.1 Pre-post analysis

A student's prior knowledge of equation solving (as measured by the pretest) was a strong predictor of their posttest score ($r = 0.79$, $p < .001$). Students went from an average of 43% on the pretest to 52% on the posttest – a significant improvement ($F(1, 133) = 17.66$, $p < .001$).

4.2.2 Replay Analysis

On average, teachers spent roughly 47% of their time either inactive or outside of the room. The proportions of time teachers were observed engaging in each of the other coded activities, within the remainder of the time, are reported in Table 1.

Table 1. Frequency of coded teacher and student behaviors during teachers' active time. Top row: average percentage of teachers' active time that was spent engaged in each of the coded behavior categories. Bottom row: average percentage of students for which a category was observed at least once.

	Teacher-student: On-task conversation	Teacher-student: Off-task conversation	Teacher: Talking to class	Teacher: Monitoring	Student: Hand-raising
Teacher time	33%	19%	4%	44%	n/a
% of students	28%	7%	n/a	34%	26%

In examining replays of a small number of class sessions, we observed a number of unexpected patterns – often re-running the replay with different combinations of analytics in order to explore particular questions more deeply. Almost immediately, we noticed that the teachers in our study tended to actively monitor their students in concentrated bursts, interleaved with (often lengthy) idle periods in which the teacher might either monitor the whole class from a fixed position in the room, or attend to an unrelated activity. During periods in which teachers were walking around the classroom, they occasionally provided students with apparently unsolicited feedback (i.e. feedback that was not preceded by the student raising her/his hand) based on their observations while watching a student's computer monitor over her/his shoulder.

In these replays, teachers appeared to selectively monitor certain students while consistently passing others by. In interviews with some of these teachers, they noted that they monitor their students strategically during computer lab sessions, relying on prior knowledge about their students' abilities and behavioral tendencies. In particular, two of the teachers we interviewed emphasized that they tend to focus on monitoring students who they expect are more likely to be off-task (e.g. browsing external websites instead of working with the software). However, replays displaying the amount of time each student spent inactive in the software suggested that teachers tended to neglect certain regions of the classroom, and overlooked students who truly tend towards greater time off-task.

For example, Figure 3 shows a group of students, on the right side of the classroom, who spent a large amount of time inactive over the course of a class session. Yet the teacher spent very little time

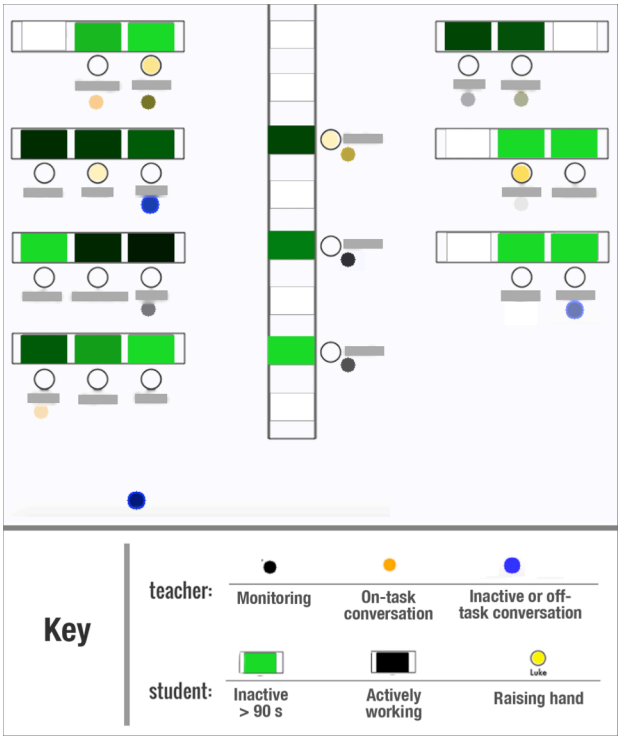

Figure 3. A time-lapse image of a SPACLE replay, summarizing a 60-minute class session. In this replay, the amount of time each student spent inactive during the entire session is displayed – ranging from black (less time) to bright green (more time). Student names are obscured in this image. Brighter yellow student circles indicate more frequent hand raising, and more faded colors of the teacher circle indicate less time spent with a particular student.

in this region of the room, and almost no time directly monitoring any of these students' screens. This may be viewed as evidence that teachers' intuitions are limited when it comes to judging which students are more likely to engage in off-task behavior. It is also possible that students sitting in regions of the room where a teacher is more active are more likely to remain on-task. Indeed, our replay analyses lend some support to this interpretation, as students frequently appeared to go off-task when the teacher moved to another region of the classroom, but then resumed working with the software once the teacher started moving in their general direction. And many students appeared to go off-task during periods in which the teacher was either inactive or outside of the room (see Figure 1).

A major takeaway from these replay analyses was that we might have previously underestimated the importance of spatial factors in the classroom when analyzing ITS log data. Although our original goal in collecting classroom observation data was to investigate the effects and predictors of teachers' helping behaviors in the classroom, replay analyses revealed that teachers' proximity seemed to have much more salient effects on student learning and behavior. A teacher's location in the classroom appeared to be related to whether or not particular students chose to be on-task, and the activity of students sitting next to one another often appeared to be temporally synchronized (similar to the "distraction ripples" observed by Raca *et al.* [21]). Furthermore, when the teacher was either distracted or outside of the classroom, many students appeared to stop working with the

software entirely. And students' willingness to raise their hands (as well as their likelihood of receiving help from the teacher as a result) appeared to increase during time intervals in which the teacher was nearby.

4.2.3 Relationships between student-teacher interactions and student learning outcomes

After using replay analysis to gain a richer qualitative picture of what went on in a small set of class sessions, we conducted quantitative analyses on the synchronized logs generated by SPACLE in order to investigate the robustness of some of the patterns we observed. Since we are ultimately interested in students' learning outcomes, we began by examining relationships between frequencies of various student-teacher interactions (evaluated per-student) and students' pre-post gains.

As shown in Table 2, neither a student's frequency of on-task conversations with the teacher nor their frequency of requesting help (via hand-raising) were significantly correlated with their performance at posttest, even when controlling for the student's pretest score. Interestingly, the frequency with which a teacher directly monitored a student was the only measured aspect of students' and teachers' interactions in the classroom that correlated significantly with posttest, and the relationship with direct monitoring remains significant even when controlling for pretest.

Table 2. Zero-order and partial correlations (controlling for pretest) between student-teacher interactions and posttest scores.

$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	On-task conversation	Off-task conversation	Direct monitoring	Hand raising
Zero-order correlation	0.00	0.13	0.39***	-0.02
Partial correlation	-0.08	-0.14	0.20*	-0.08

In order to better understand the mechanisms by which this apparent link might arise, we adopt a causal model search approach, using directed acyclic graphs (DAGs) to represent the qualitative causal structure among measured variables. We used the PC algorithm in the Tetrad V program[4] to search for an equivalence class of graphs that are consistent with a set of conditional independence constraints [22]. We included background knowledge about our experimental design as a search constraint: namely, that the pretest precedes all process variables, which in turn are all prior to the posttest. The PC algorithm is asymptotically reliable, and its primary limitations lie in its assumptions that the underlying causal dependencies between variables can be modeled with linear functions, and that there are no unmeasured common causes among variables.

To relax the second of these assumptions, we also used the FCI algorithm to learn an equivalence class of graphs, represented by partial ancestral graphs (PAGs). PAGs are representationally richer than DAGs, and may contain edges representing uncertainty over the nature of pairwise relationships between variables [22]:

- X → Y: X causes Y in every member of the equivalence class represented by this PAG.
- X ↔ Y: X and Y share a latent common cause in every member of the equivalence class represented by this PAG.
- X **o**→ Y: Either X causes Y, X and Y share a common cause, or both.
- X **o—o** Y: X is a cause of Y or Y is a cause of X. Alternatively, X and Y may share a latent common cause (either in the absence of a direct causal link between the two variables, or in addition to one).

Figure 4 shows the model found by PC, with path coefficient estimates included. The model fits the data well ($\chi^2 = 6.03$, df = 10, $p = .81$)[5], and contains a number of properties that are consistent with findings in prior literature on the effects of student help-seeking behaviors on learning gains with ITSs. For example, under this model increased use of the ITS's hint functionality appears to inhibit learning, overall [6]. Also, compatible with previous findings that on-task conversations with peers and teachers during ITS use may be negatively related to student learning overall, we find that on-task conversations with teachers appear to increase students' error rates within the software [13]. However, we did not replicate Baker *et al.*'s finding of a negative relationship between on-task conversations and learning gains, instead finding no relationship (perhaps owing, in part, to differences in the quality and effects of peer help and teacher help). Note that the observation of a negative relationship between on-task conversations and student error rates, and the absence of an observed relationship with learning gains may be, at least in part, due to a selection effect. Students who have more on-task conversations with the teacher are likely those who are having more difficulties in the software (for reasons that may not be captured by their performance on the pretest alone), and who are in turn likely to learn less [6, 13]. In addition, it is possible that a finer-grained coding of the nature or content of these on-task conversations would have revealed particular circumstances under which such conversations produce a measurable increase or decrease in student learning, as measured by posttest.

The observed positive relationship between the frequency of direct monitoring by the teacher and student posttest scores appears to have been mediated, in part, by students' hint-use behavior. One possibility this suggests – made more plausible by our observations during replay analyses – is that students who are more aware that the teacher is monitoring them are less likely to engage in maladaptive learning behaviors such as abusing software-provided hints, and are therefore more likely to learn the material. It is also possible, however, that the apparent link between teachers' direct monitoring and student learning gains

[4] Available at http://www.phil.cmu.edu/projects/tetrad/

[5] Note that in path analysis, the null hypothesis is that the estimated model is the true model, and the p-value represents the probability that a difference between the estimated and the observed covariance matrices at least as large as the realized difference would have been observed under the null hypothesis. As such, a p-value above a specified threshold (conventionally alpha = .05) implies that the model cannot be rejected.

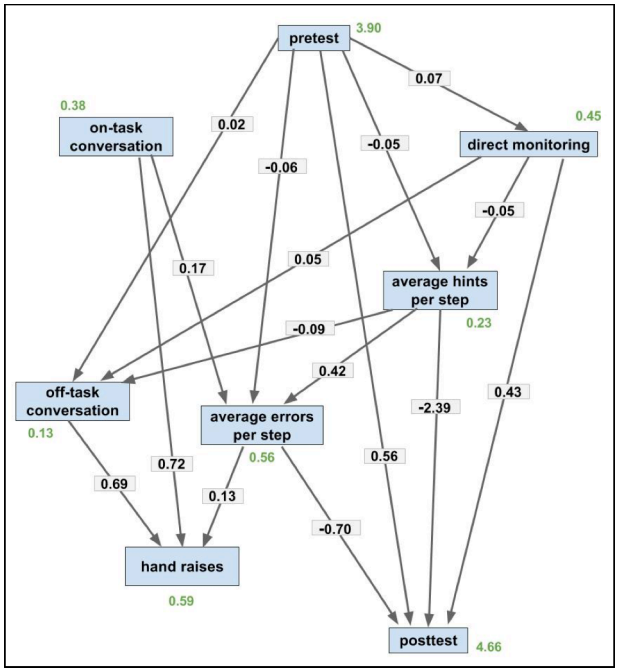

Figure 4. The model found by PC, with parameter estimates included. This model fits the data well: $\chi^2 = 11.31$, df = 12, p = .50.

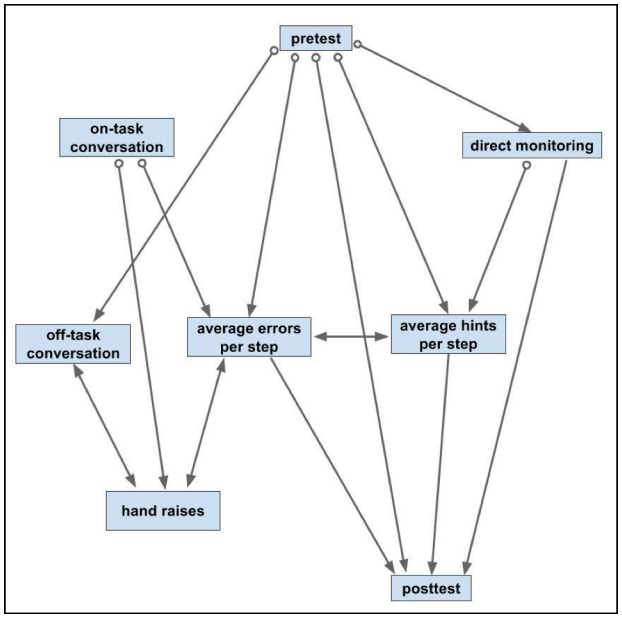

Figure 5. The PAG equivalence class found by FCI, which encodes the possibility of unmeasured common causes.

reflects a selection effect. For example, teachers may tend to more frequently monitor students who show signs of making progress in the software (or who the teacher believes are more likely to make progress). Interestingly, this model suggests that students with higher pretest scores may have been somewhat more likely to receive additional monitoring from the teacher. In a follow-up interview, one of the teachers in our study claimed to have intentionally placed a group of students in a relatively isolated and inaccessible area of the classroom, as these students were "a pain to deal with" – hinting at possible mechanisms by which this apparent bias could have arisen.

Stang *et al*. recently found similar results at the university level [18]. In their study of interactions between teaching assistants (TAs) and students in the 'hands-on' laboratory sections of large introductory physics courses, these authors found that the frequency of TA-student interactions was a strong and positive predictor of student engagement (defined as on-task behavior), which was in turn a stronger predictor of student learning gains than their pretest scores. Compatible with our findings, the authors found that this relationship held for interactions that were initiated by TAs, but not for those initiated by students. In addition, very brief visits by the TA appeared to be just as effective as lengthy interactions. The authors posited that this might be due either to a "policing" effect (i.e., frequent interactions motivate students to *not* stray off-task), or a "ventilation" effect (i.e., TA-initiated visits open the door for productive conversations with students). To gain a sense of the relative plausibility of these two explanations in our own dataset, we ran follow-up replay analyses with SPACLE, across two teachers and class sessions -- visualizing the rate of student hint requests on each student's "computer screen" by displaying a flash of color each time a student asked for a hint. These replays suggested that students might have been less likely to request hints when the teacher was nearby. In addition, students who were observed asking for multiple hints in rapid succession appeared to stop (or at least, pause) this behavior when the teacher was nearby or directly monitoring them – lending some support to Stang et al.'s "policing" hypothesis, while also remaining compatible with their "ventilation" hypothesis.

Given the potential for confounding factors, we used the FCI algorithm to learn a PAG causal model, relaxing the assumption of no unmeasured common causes (see Figure 5). The learned structure is largely the same, except that this model encodes the possibility that pretest may be related to direct monitoring, off-task conversation, hint use, and/or error rate by a common unmeasured cause, and that the same may be true for the relationships between direct monitoring and hint use, and on-task conversation and its children (hand raises and error rate). In addition, the learned structure suggests that students' frequency of hand-raising shares common unmeasured causes with their frequency of off-task conversation and their within-software error rates (which in turn may share a common cause with students' rate of hint-use) – perhaps indicating that these behaviors are symptoms of unmeasured cognitive, motivational, and affective states such as confusion and frustration [34]. However, the positive link between direct monitoring and student learning gains remains in every member of the equivalence class found by FCI.

5. DISCUSSION AND FUTURE WORK

We have introduced SPACLE, a prototype tool that facilitates exploratory, retrospective analyses of learning-related interactions that may cross over between virtual and physical spaces. In addition, we have demonstrated how SPACLE can support hypothesis generation, by using replay analysis to inform our own investigations into the effects of teacher-student interactions on K-12 students' learning with intelligent tutoring systems. We used SPACLE replays to inform quantitative log analyses in two ways: first as a means to quickly explore multimodal classroom datasets and identify important classroom behaviors that likely have an impact on learning, and second to continuously evaluate the relative plausibility of various, alternative hypotheses that were consistent with the results of our quantitative analyses.

Furthermore, through a combination of causal modeling and replay analysis with SPACLE, we have presented some convergent evidence for positive effects of teachers' monitoring behaviors on student learning in classrooms using ITSs. Specifically, our findings from causal modeling suggest that students who receive more frequent monitoring from teachers in ITS classrooms may learn more, and that this effect may be partially mediated by students' hint-use behavior within the software. Our use of SPACLE replays on a small subset of our data throughout the analysis process enabled us to evaluate the relative plausibility of various hypotheses that were compatible with these causal models. Students who are monitored by their teachers more frequently tend to engage less often in "gaming the system" behaviors such as hint abuse, and may also be less likely to go off-task.

These findings extend those of Stang *et al*. [13] by suggesting that more frequent visits from a teacher may promote engagement and learning not only at the university level, but also among considerably younger students (7^{th}-8^{th} grade). Our findings also lend support to the authors' prediction that their observed relationship between teacher visits and student engagement would generalize beyond their study's setting (inquiry-based laboratory sessions in an introductory physics course). In addition, our findings may help interpret Stang *et al.*'s observation that a teachers' frequency of interaction with a student predicts student engagement, independent of the length of these interactions. Our findings suggest that teachers' interactions may not need to have a verbal component in order to be effective – that is, K-12 students' mere awareness of being monitored may have a positive impact on their learning with self-paced systems such as ITSs.

Without using SPACLE for our initial explorations, we likely would not have turned our attention, in the first place, to studying potential *effects* of teachers' monitoring behaviors. Rather, we had originally collected classroom observations on teachers' monitoring behavior in order to study potential teacher blind spots during blended lab sessions (e.g., failing to notice important opportunities to help students learn the material, or exhibiting an unconscious bias towards helping and monitoring certain subsets of students). Informally, SPACLE replays suggested that teachers tended to spend a significant amount of time inactive during lab sessions and often overlooked students who tended to spend more time off-task. Another one of our initial goals for these analyses was to model and understand how teachers decide which students to help, in order to understand how their help giving might be better allocated. Contrary to our initial expectations, however, the frequency of teachers' verbal interactions with students was not a significant predictor of student learning, overall, even when examining on-task conversations only.

These results should not be interpreted as suggesting that on-task conversations with a teacher cannot be helpful. Indeed, we expect that there exist many scenarios in which help from a human teacher is likely to be *more* effective than the support ITSs can currently offer. As mentioned, a selection effect may be responsible, at least in part, for the absence of an observed relationship between on-task teacher-student conversations and student learning gains. However, this absence does suggest that any *overall* positive effect of such conversations is not strong enough to offset the selection effect. It may also be that, consistent with prior work on the effects of student hint-use within ITSs, on-the-spot support from a human teacher during blended lab sessions is helpful only under particular circumstances [6]. For example, it may be that current ITSs are generally more effective at teaching procedural skills, whereas human teachers can be more effective at teaching conceptual knowledge [36]. Under this interpretation, our pre- and posttests may not have been able to capture the effects of students' on-task conversations with their teachers, since these assessments were primarily designed to test students' procedural knowledge in equation solving. It may also be that the effectiveness of a particular on-task conversation with a teacher depends jointly upon student traits (e.g., the student's inclination to self-explain ideas presented by the teacher) and the type and quality of the help the teacher provides (e.g., completing a problem for the student as a worked example, versus prompting the student to work through the problem while verbalizing her/his thought process).

Although the coding scheme used in the current study was not fine-grained enough to capture such distinctions, we view the investigation of circumstances under which help from a human teacher is more beneficial than help from an ITS (or vice-versa) as a promising direction for future work. Such research could inform the design of more synergistic blended curricula – combining the complementary strengths of both human teachers and ITSs. It could also inform the design of learning analytics tools to help teachers more effectively support their students while they work with learning technologies such as ITSs in the classroom.

Several limitations of this work should be mentioned. The causal models shown in Figures 4 and 5 should not be viewed as the "true" models. First, although our data are from an experimental study, the data reported in this paper are from a portion of the study in which we did not directly intervene on any of the measured variables between pre- and posttest (except insofar as running an in-vivo classroom study can be considered an intervention in itself). As such, our data should be considered observational, and future experimental investigation is required to evaluate the causal nature of each link identified in our causal models. Second, by no means can we rely on the assumption made by our search algorithms, that the underlying relationships between our modeled variables are truly linear. Nonetheless, this model assumption is not unreasonable, as the relationships in the data we modeled appear approximately linear. Third, although our sample of 135 students is relatively large compared to many ITS studies, the reliability of our model search algorithms would be improved with access to larger samples, and it is generally impossible to compute confidence bounds when dealing with finite samples [45]. Fourth, although SPACLE allows us to quickly run exploratory replay analyses that capture spatiotemporal factors, time-series analyses could enable deeper analysis of individual links in our causal models by leveraging temporal precedence as a cue to causality (i.e., a scalable formalization of part of what human researchers do when observing SPACLE replays [35]). In our future work, we plan to apply algorithms for causal modeling from time-series data.

It is also worth noting that we observed teachers over a relatively brief period (60 minutes) in this study, as we were interested in investigating student-teacher interactions on a small scale. It would be interesting to observe ITS classrooms over longer time periods, in order to study how teacher practices (and perhaps also their effects on student learning) may evolve over time. And finally, in this study a single human observer manually collected classroom observations of teacher-student interactions. This required us to use a very coarse-grained coding scheme, and also limited the number of classrooms we could feasibly observe. In future work, we will automate parts of the classroom data collection process (building on recent work by Prieto *et al.* [32]), using a combination of low-cost sensors and manual observations. A semi-automated approach may enable more detailed coding

schemes by freeing human observers to focus on recording higher-level observations (e.g. semantic features of on-task conversations in the classroom).

In addition to using SPACLE for exploratory data analysis, we have also begun exploring the use of SPACLE as a design tool. First, we've begun using SPACLE in our own design work, to support the iterative design and prototyping of analytics for use in real-time teacher dashboards. SPACLE allows designers to experiment with alternative analytics (e.g. different alert thresholds for behavior detectors, or different measures of the same psychological construct) and examine the consequences of particular choices in a tangible way. Second, we have begun exploring the use of SPACLE as a means to investigate the distance between teachers' actual behavior in the classroom and their recollections of their activities. For example, after observing class sessions, we have asked teachers to walk us through their activities while student were working on computers, drawing their physical paths over top maps of their classrooms in the process. In comparing teachers' recollections with SPACLE replays, we've observed that the replay often reveals a much lower amount of teacher activity (often fairly concentrated in particular regions of the classroom) than teachers' recollections would suggest. In the future, we would like to explore the use of SPACLE-generated replays as "after action reviews" for teachers, to encourage them to reflect on their own activity patterns in the classroom. However, in order for such reflection tools to be used in teachers' daily practice, outside of exploratory design studies, the collection of classroom observations would need to be heavily automated.

In conclusion, our results have implications for the learning analytics, educational data mining, and intelligent tutoring systems communities. Using replay analysis with SPACLE, we generated a number of questions about the nature and effects of teachers' on-the-spot decision-making during blended lab sessions. Through both replay analyses and causal modeling, we observed rich relationships between students' out-of-software interactions in blended lab sessions using ITSs and their within-software learning and behavior. Some of the most salient observed effects involved no verbal interactions between students and their teachers, but rather appeared to be due to spatial factors (e.g. the teacher's position in the room, relative to a student) and perhaps classroom layout. We view these observations as suggestive that the influence of such out-of-software factors on student learning with ITSs and similar educational technologies has perhaps been under studied previously. Our finding that the frequency with which teachers monitor students is predictive of learning gains may indicate that one possible mechanism by which classroom monitoring tools such as real-time dashboards might be effective in promoting student learning is by simply making students aware that they are being monitored. In our future work, we plan to use SPACLE replays in conjunction with causal modeling, to construct models of teacher decision-making [32, 37] and identify additional links between teacher behavior and student learning. These models could in turn help inform the design of more effective learning analytics tools for teachers. Our findings also suggest that students may systematically take advantage of affordances offered by the physical classroom (e.g. teachers' limited attention and perceptual abilities) in order to decide whether and when to go off-task or abuse hints (consistent with previously reported informal classroom observations [34]). This hints at the usefulness of a broader notion of "gaming the system" than has been used previously – taking into account student behaviors that extend outside of the software.

6. ACKNOWLEDGMENTS

This work was supported by NSF Award #1530726, and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. In addition, we thank the DataShop and CTAT/TutorShop teams, Franceska Xhakaj, Jasper Tom, Ran Liu, Amos Glenn, Mary Beth Kery, and all participating teachers and students.

7. REFERENCES

[1] Liu, R., Davenport, J., & Stamper, J. (2016). Beyond log files: Using multi-modal data streams towards data-driven KC model improvement. In *EDM 2016*, 436-441.

[2] Baker, R.S.J.d, Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., & Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In *EDM 2012*, 126-133.

[3] Miller, W. L., Baker, R. S., Labrum, M. J., Petsche, K., Liu, Y. H., & Wagner, A. Z. (2016). Automated detection of proactive remediation by teachers in Reasoning Mind classrooms. In *LAK 2015*. ACM.

[4] 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 *CHI 2012*, 1381-1390. ACM.

[5] Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In *AIED 2013*, 431-440. Springer Berlin Heidelberg.

[6] 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. IJAIED, 26(1), 205-223.

[7] Rau, M., Scheines, R., Aleven, V., & Rummel, N. (2013). Does representational understanding enhance fluency – or vice versa? Searching for mediation models. In *EDM 2013*, 161-168.

[8] Schofield, J. W. (1995). *Computers and classroom culture*. Cambridge University Press.

[9] Mavrikis, M., Gutierrez-Santos, S., & Poulovassilis, A. (2016). Design and evaluation of teacher assistance tools for exploratory learning environments. In *LAK 2016*, 168-172. ACM.

[10] VanLehn, K. (2006). The behavior of tutoring systems. IJAIED, 16(3), 227-265.

[11] San Pedro, M. O. C. Z., Baker, R. S., & Rodrigo, M. M. T. (2011). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *AIED 2011*, 304-311. Springer Berlin Heidelberg.

[12] Fancsali, S. (2014). Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *EDM 2014*, 28-35.

[13] Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In *CHI 2004*, 383-390. ACM.

[14] Long, Y., & Aleven, V. (2013). Supporting students' self-regulated learning with an open learner model in a linear equation tutor. In *AIED 2013*, 219-228. Springer Berlin Heidelberg.

[15] Dyke, G., Kumar, R., Ai, H., & Rosé, C. P. (2012). Challenging assumptions: Using sliding window visualizations to reveal time-based irregularities in CSCL processes. In *ICLS 2012*, 1, 363-370.

[16] Long, Y., & Aleven, V. (2016). Supporting shared student/system control over problem selection with an open learner model in a linear equation tutor. In *ITS 2016*, 90-100. Springer International Publishing.

[17] Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, 105(4), 970-987.

[18] Stang, J. B., & Roll, I. (2014). Interactions between teaching assistants and students boost engagement in physics labs. *Physical Review Special Topics-Physics Education Research*, 10(2).

[19] Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901-918.

[20] Kulik, J. A., & Fletcher, J. D. (2015). Effectiveness of Intelligent Tutoring Systems A Meta-Analytic Review. *Review of Educational Research*, 42-78.

[21] Raca, M., & Dillenbourg, P. (2013). System for assessing classroom attention. In *LAK 2013*, 265-269. ACM.

[22] Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. MIT press.

[23] King, A. (1993). From sage on the stage to guide on the side. College Teaching, 41(1), 30-35.

[24] Roll, I., & Wylie, R. (2016). Evolution and revolution in Artificial Intelligence in Education. *IJAIED*, 26(2), 582-599.

[25] Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167-207.

[26] Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2), 147-179.

[27] Koedinger, K. R., Aleven, V., Roll, I., & Baker, R. (2009). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. *Handbook of Metacognition in Education*, 897-964.

[28] Aleven, V., McLaren, B.M., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. In *ITS 2004*, 227-239. Springer Berlin Heidelberg.

[29] Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 2013.

[30] Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64-71.

[31] Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016). How mastery learning works at scale. In *L@S 2016*, 71-79. ACM.

[32] Prieto, L. P., Sharma, K., Dillenbourg, P., & Rodriguez Triana, M. J. (2016). Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors. In *LAK 2016*, 148-157. ACM.

[33] Martinez-Maldonado, R., & Martinez, R. (2016). Seeing learning analytics tools as orchestration technologies: Towards supporting learning activities across physical and digital spaces. In *1st International Cross-LAK Workshop at LAK 16*, 70-73.

[34] Baker, R. S. (2011). Gaming the system: a retrospective look. *Philippine Computing Journal*, 6(2), 9-13.

[35] Baker, C., Saxe, R., & Tenenbaum, J. B. (2005). Bayesian models of human action understanding. In *NIPS 2005*, 99-106.

[36] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197-221.

[37] Borko, H., Roberts, S. A., & Shavelson, R. (2008). Teachers' decision making: From Alan J. Bishop to today. *In Critical Issues in Mathematics Education*, 37-67. Springer US.

[38] Xhakaj, F., Aleven, V., & McLaren, B. M. (2016). How Teachers Use Data to Help Students Learn: Contextual Inquiry for the Design of a Dashboard. In *EC-TEL 2016*, 340-354. Springer International Publishing.

[39] De Vicente, A., & Pain, H. (2002). Informing the detection of the students' motivational state: an empirical study. In *ITS 2002*, 933-943. Springer International Publishing.

[40] Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In Proceedings of the Educational Data Mining Workshop at *ITS 2006*, 29-36.

[41] Aleven, V., Xhakaj, F., Holstein, K., McLaren, B.M. (2016). Developing a teacher dashboard for use with intelligent tutoring systems. In *4th International Workshop on Teaching Analytics at EC-TEL 16*, 15-23.

[42] Ocumpaugh, J. (2015). Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

[43] Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2), 220-238.

[44] Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

[45] Robins, J. M., Scheines, R., Spirtes, P., & Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3), 491-515.

[46] Rodriguez Triana, M. J., Prieto, L. P., Vozniuk, A., Shirvani Boroujeni, M., Schwendimann, B. A., Holzer, A. C., & Gillet, D. (in press). Monitoring, Awareness and Reflection in Blended Technology Enhanced Learning: a Systematic Review. In *IJTEL*.