The Burden of Facilitating Collaboration: Towards Estimation of Teacher Orchestration Load using Eye-Tracking Measures

Luis P. Prieto, Kshitij Sharma, Yun Wen and Pierre Dillenbourg
luis.prieto@epfl.ch, kshitij.sharma@epfl.ch, yun.wen@epfl.ch, pierre.dillenbourg@epfl.ch
CHILI Lab, École Polytechnique Fédérale de Lausanne

Abstract: Teacher facilitation of CSCL activities is widely recognized as one of the main factors affecting student learning outcomes in formal face-to-face settings. However, the orchestration load that such facilitation represents for the teacher, within the constraints of an authentic classroom, remains under-researched. This paper presents a novel method to estimate the cognitive load of teachers during facilitation of CSCL sessions, using mobile eye-tracking techniques. Throughout three studies of increasing authenticity, we demonstrate the feasibility of this approach, and extract insights about classroom usability challenges in CSCL practice: the increased load of class-level facilitation, or the real-time monitoring of students' progress. This new instrument in the CSCL researcher's toolkit can help focus our attention in critical, fine-grained classroom usability episodes, to make more informed design decisions.

Keywords: orchestration, cognitive load, eye-tracking, teaching, facilitation

Introduction

Although CSCL often focuses on distance or informal learning, the physical, the face-to-face classroom in formal education is still the most common CSCL setting for many learners across the world. In the everyday life of these classrooms, teacher facilitation of collaborative learning is considered one of the main factors affecting learning outcomes (e.g., Onrubia & Engel, 2012; Song & Looi, 2012; Gómez et al., 2013): although student interactions remain crucial in CSCL, we still expect teachers to prepare, manage, assess and even adapt CSCL activities in their everyday practice: what Dillenbourg, Jarvela & Fischer (2009) termed 'orchestration'.

CSCL research has proposed new technologies and innovations to support learning together, focusing greatly on their individual and small-group usability. However, the study of pedagogical usability of CSCL innovations at the classroom level (e.g., will this nicely-designed, computer-supported group interaction be manageable for the teacher when ten groups are working at the same time?) has not been in the center of attention until recently (Dillenbourg et al., 2011). First studies and guidelines on the 'orchestration load' of CSCL facilitation are starting to appear (e.g., Cuendet et al., 2013), although they mostly remain on a qualitative, high-level of abstraction. Having a wider palette of techniques to measure orchestration load, not only from a qualitative perspective but also in a more objective manner, can be essential to understand the limited uptake that many CSCL innovations have – a growing concern in our community (Chan, 2011).

In this paper, we explore a method to estimate teachers' load when facilitating CSCL activities in authentic face-to-face classroom settings, using mobile eye-tracking techniques. This exploration is composed of three studies conducted in different contexts: a) a first analytic study of the load when performing a simple task in a lab setting; b) an exploratory case study, in which a researcher facilitated three CSCL sessions with a total of 61 primary school children, in a multi-tabletop classroom; c) a case study in the context of a university masters course, in which an expert and a novice teacher facilitated three sessions with the same group of 12-14 students, combining lecturing and collaborative work.

The rest of the paper is structured as follows: after a section detailing the related background in the areas of orchestration, cognitive load measurement and eye-tracking, we describe the general methodology of our studies; then, the context, methods and result of the three studies are sequentially described; afterwards, we discuss the results of the studies and conclude with their implications for research on orchestration, and current and future work.

Related background

Orchestration and orchestration load

When CSCL is enacted in a formal educational setting, teacher facilitation can greatly affect the effectiveness of learning (Onrubia & Engel, 2012; Song & Looi, 2012; Gómez et al., 2013). In this direction, Dillenbourg, Jarvela & Fischer (2009) define orchestration as "the process of productively coordinating supportive interventions across multiple learning activities occurring at multiple social levels". In this paper we will focus mainly on orchestration as the *run-time coordination of CSCL activities*, even if that might not be the only aspect that CSCL orchestration entails (Prieto et al., 2011; Roschelle, Dimitriadis & Hoppe, 2013).

In the literature we can find several examples of research efforts trying to address this run-time coordination in different educative contexts: Pérez-Sanagustin et al. (2009) implement flexible group management mechanisms on blended CSCL (computationally-interpretable) scripts, to automate the flow of learning activities. Muñoz-Cristobal et al. (2013) study how sharing the load of creating and managing learning artifacts can ease the orchestration of CSCL activities across multiple spaces. Cuendet et al. (2013) gather lessons learnt from several studies and propose five general design guidelines to lessen orchestration load when designing augmented reality technologies for the classroom. However, these works generally focus on qualitative, high-level conclusions about orchestration load, or ad-hoc measurements of orchestration efficiency (e.g., waiting time of students), based on a specific activity flow.

Dillenbourg et al. (2011) posit orchestration as "usability in the third circle", referring to the classroom-level usability (as opposed to single user and group-level usability, the focus of many HCI and CSCW/CSCL research works, respectively). However, as we have seen, we currently lack a finer-grained, more objective way of measuring orchestration load of a CSCL innovation for the teacher/facilitator, which might complement current subjective assessments of orchestration load (e.g., gathering the teachers' feedback after using a technology). Thus, we ultimately aim at helping respond to questions like: Which moment of a classroom enactment poses more load? Which technology design alternative will make classroom facilitation easier?

Orchestration load could be said to have two main components: a) the physical/logistic load (e.g., moving around the classroom, writing on the whiteboard, distributing worksheets); and b) the cognitive load of assessing what is going on in the classroom, weighing different courses of action and taking decisions about how to best support the ongoing CSCL process. This latter kind of load is probably the most important of the two in many situations, but it is also quite difficult to observe directly. Fortunately, cognitive load has been extensively studied in other contexts by psychology and the field of human-computer interaction (HCI).

Measuring cognitive load

In cognitive psychology, cognitive load (CL) is related to the executive control of working memory, and the limited capacity of human cognitive processing capacity (Paas et al., 2004). Cognitive Load Theory (CLT), its most direct application to the learning domain, has been extensively used in CSCL and educational psychology in general, and has served to discover several effects that may hamper learning by producing unwanted extraneous CL: split-attention effects, modality effects, worked example effects, etc. (van Merrienboer & Sweller, 2005).

Multiple techniques have been used to measure cognitive load, which Brunken et al. (2003) classify along two dimensions: direct measures (targeting cognitive load itself) vs. indirect ones (targeting other constructs related to CL), and subjective measures (those relying on subject introspection and memory) vs. objective ones. For instance, researchers have tried to measure CL using self-report questionnaires (subjective, direct/indirect), brain imaging techniques (objective, direct), dual-task performance (objective, direct) or physiological measures such as heart rate monitoring or eye-tracking (objective, indirect). However, it is important to note that many of these techniques have been applied mostly to relatively simple tasks, under controlled lab conditions. The measurement of cognitive load in more complex everyday activities is still considered difficult and needs to be further explored (van Merrienboer & Sweller, 2005).

Eye-tracking and teacher cognitive load in authentic settings

As we have seen, cognitive load has been used extensively to study learning tasks – but could it be used to study teaching as well? Although cognitive processing is often mentioned in teacher education literature, papers that focus specifically on this topic are not common: as a rare exception, Feldon (2007) re-interprets findings from teacher education research in terms of cognitive psychology, and highlights the role of developing automaticity to drive teacher training effectiveness.

Although measuring CL during CSCL facilitation could play a crucial role in designing technology for classroom-level usability (in the same way that CLT helped design technologies more conducive to student learning), the activity of teaching presents a unique set of challenges to such measurement, especially taking into account that orchestration is, by definition, concerned with what happens in authentic, non-controlled classroom settings rather than laboratory tests (Roschelle et al., 2013).

Brain imaging techniques, while being a direct and objective measure of CL, has been found unreliable, even in lab settings, by Paas et al. (2003), and are difficult to implement in a classroom setting. The fact that teaching is a complex and demanding activity, leading to overload (especially in the case of novice teachers –

see Feldon, 2007) makes dual-task measurements of CL (i.e., posing a secondary task to perform simultaneously with the main one) too obtrusive (Paas et al., 2003). Subjective self-report questionnaires, when administered during a lesson, can be equally disruptive, or have to rely on the subject's general memory of a long stretch of time (e.g., 50+ minutes in a normal lesson time span). The fact that "performance" in teaching is very hard to measure and also rules out performance-based measurements of CL. This leaves us with physiological measures of cognitive load, among which eye-tracking is considered quite reliable (Paas et al., 2003).

Eyetracking techniques measure and analyze the eye movements (fixation into a point of attention, saccades to the next fixation point, pupillary dilation responses, all in the scale of milliseconds), as a window into human cognition (Holmqvist et al., 2011). Although traditionally performed in lab settings using fixed equipment, the appearance of mobile eye-trackers (e.g., in wearable goggles format) now makes it possible to study such physiological response in the context of authentic activities. The relationship between pupillary response and cognitive load (e.g., mean pupil dilation is correlated with CL) has a long history in ergonomics (Kahnemann & Beatty, 1966). However, this kind of measures are known to be sensitive to varying lighting conditions (which are bound to occur in an authentic classroom). There exist multiple eye-tracking measures that can be related to CL: Buettner (2013), for example, identifies four such measures (mean pupil dilation, pupil dilation standard deviation, number of long fixations, average saccade speed). Thus, the triangulation of multiple eye-tracking measures could provide a more reliable way to trace teacher while facilitating CSCL in authentic classroom settings.

Methodology

Against this background, our main research question is: *can we use eye-tracking techniques to follow cognitive load of teachers facilitating CSCL in authentic settings?* In order to explore this question, we set out to apply the measurement of the four metrics used by Buettner (2013) in three studies (with increasing degree of authenticity). The main goal and features of these studies are depicted in Table 1, and further details of the context and methodology of each study can be found in each of the following sections.

Table 1: Main characteristics of the studies

Study	Analytical	Semi-authentic	Authentic
Setting	Laboratory	Multi-tabletop classroom, lab 'open doors' day	Authentic course, classroom including projector and laptops
Goal	Test method in different task, see evolution of CL over time	Feasibility of eye-tracking within classroom constraints, insights about multi-tabletop classroom usability	Feasibility of eye-tracking in real course, individual differences of novice/expert teachers
Subjects	16 participants	1 facilitator-researcher, 61 primary school students	1 expert teacher, 1 novice teacher, 12-14 students
Task	Game-based	Facilitation of small group collaborative work	Mix of lecture and collaborative work
Study duration	128 games in total, 1.5-4 minutes each	3 sessions, 35-45 min each	3 sessions (2 expert, 1 novice), 45-65 min each

Analytical study: Cognitive load in a simple game-based task

As a first step in our exploration of eye-tracking to follow teacher cognitive load, and taking into account that eye responses are often task-dependent (Holmqvist et al., 2011), we devised a first *test for the validity of the measurements proposed in Buettner (2013)*. Using existing eye-tracking data from a previous experiment in which participants played a computer game (unrelated to the task in Buettner's study), we estimated the evolution of the CL of participants throughout their experience, to see whether the results were consistent with what we knew about the game task in question and its temporal evolution.

Context and method

The eye-tracking data for this study was gathered during the course of an experiment on reward systems, collaborative/competitive behavior and stress (Senn & Goette, in progress). In the experiment, 16 subjects had to play a game of Tetris, either in collaborative or in competitive mode, with another participant, for 8 games. The maximum duration (if there was no 'game over') for each game was 5 minutes. Eye-tracking measures (eye movements, pupil dilation) were recorded, along with different game variables (e.g., game points, height of the stack of blocks, etc.).

The four measures used by Buettner (mean pupil diameter, pupil diameter standard deviation, saccade speed and number of fixations longer than 500ms) were calculated for every participant and game using these data, over a sliding window of 10 seconds (with a 5s slide from one window to the next). Then, a median cut was performed (using the median for each game), and a “load index” was calculated by counting the number of measures that were *above* the game median (thus going from 0 to 4), as a rating of how likely it is that a certain 10-second window represented higher cognitive load than other windows in that session. Then, this load index was compared with the two main game variables (the height and the variance of the stack of blocks), along the duration of each game (1).

Results

Figure 1a shows the average cognitive load index (brown curve) of participants as time went on, as well as the temporal evolution of the stack height (green curve) and stack variance (blue curve). If we think in terms of the particular task (the Tetris game), we get interesting insights into how the cognitive load evolves over time: at the beginning (low mean stack heights) the cognitive load is high (as many alternative options for placing a new piece are open to decide amongst), and it generally goes down as the game goes towards the end (higher mean stack height), until we eventually disengage from the game when we give up. Similar but opposite is the effect of stack variance (higher variance implies more complex stack profiles, difficult to process and with more open alternatives of placement). Figures 1b and 1c show the main descriptive statistics of both game variables on 10-second windows, classified by their load index. Our results show that the load index is positively correlated with stack variance, and negatively correlated with the mean stack height, and that such an effect is more clearly apparent in the extreme values of the load index.

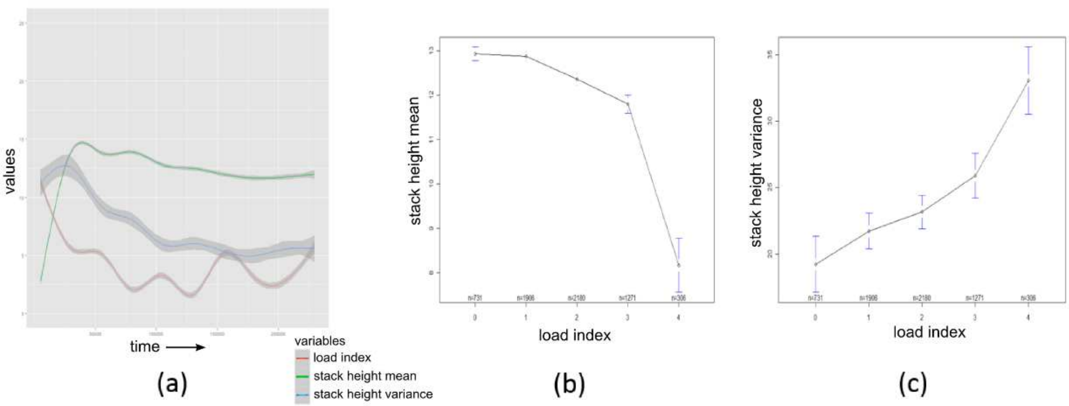

Figure 1. (a) Temporal evolution of averages of load index, stack height mean and variance; (b) and (c), stack height mean and variance in episodes classified by their load index

These results show that the load index, computed as described above, has potential for distinguishing different kinds of episodes occurring during the task (represented by moments with different mean stack heights and variances). We also see how CL may be related with the amount of *open alternatives* in each moment (in a sense, the uncertainty or the ‘entropy’ we perceive about the current game situation).

Semi-authentic study: Multi-tabletops at an open-doors day in the lab

Although playing Tetris is certainly far removed from the activity of teaching, the aforementioned results encouraged us to attempt the application of similar methods to the analysis of teacher facilitation of CSCL in a real-life setting. Thus, for the next study we aimed to explore the following research questions: *is it feasible to use a mobile eye-tracker to follow CL in a semi-authentic classroom setting? Does such analysis provide interesting insights about classroom usability of a novel CSCL technology? Can we detect specific classroom interaction episodes that imply high (or low) cognitive load?*

Context and method

The study was conducted in the context of an open doors day in our lab, in which whole classrooms of students from nearby primary schools had the chance to experience new learning technologies. In this occasion, a room was set up with five augmented paper tabletop devices running an augmented-paper tabletop collaborative learning software about mathematics (see Figure 2, and Caballero et al., 2014 for further details). A researcher (a novice in teaching to primary school children) played the role of the teacher/facilitator in this simulated math lesson about fractions, assisted by two other researchers and with the presence of two of the usual school teachers acting as observers, during approximately 40 minutes. In total, 61 primary school students (10-12 years old) attended these sessions.

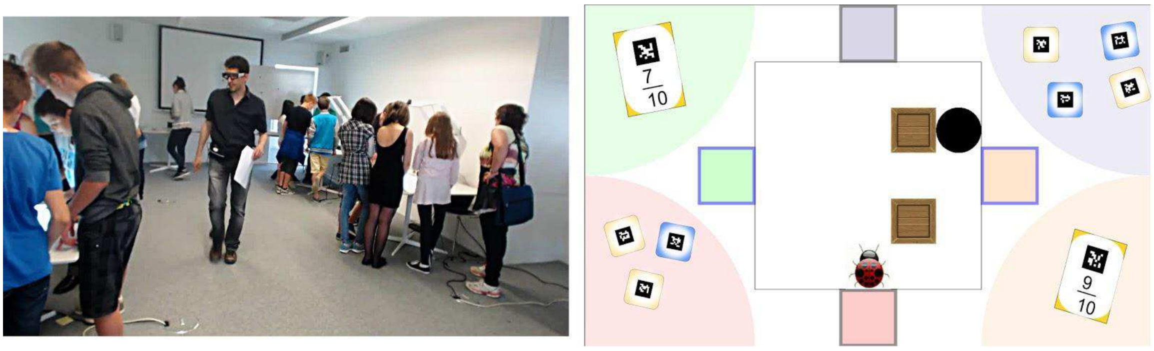

Figure 2. Setup of the multi-tabletop classroom and facilitator wearing mobile eye-tracker (left). User interface of the collaborative augmented paper game used by students during the session (right)

Again, eye-tracking measures of eye movement and pupil dilation were recorded, and a load index was calculated for (sliding) windows of 10 seconds (in a similar way as for the previous study). From this set of 10-second episodes we selected those that had minimum (0) or maximum (4) cognitive load index (as we saw in the previous study that the load index had more contrasting power in these extremes), and performed a qualitative video coding of them, to assess the main trends/patterns in orchestration properties of high/low load episodes. Taking into account the definition of orchestration by Dillenbourg, Jarvela & Fischer (2009, see 'Related work' section), each episode was coded along three dimensions: the *activity* or intervention the teacher was performing (explanation/lecture, monitoring, task distribution, repairs…), the *social plane* at which the activity was intended (individual, group or classroom level) and the *main focus* of the teacher's gaze during the episode (including students' faces or backs, the tabletop surfaces, the teacher desk, etc.). More details about the context and method used can be found in Prieto et al. (2014).

Results

A statistical analysis (Pearson's chi-squared test of independence) of the video coding for the episodes with minimum and maximum load index (n=315) revealed that high-load and low-load episodes had *statistically significant differentiated profiles* in all three coding dimensions: teacher activity ($\chi^2 = 15.3434$, df = 3, p = 0.001546), activity's social plane ($\chi^2 = 123.2922$, df = 1, p < 2.2e-16) and the main focus of teacher's gaze ($\chi^2 = 252.1052$, df = 7, p < 2.2e-16). By looking at the contributions of the different video codes to this statistical difference (the chi-squared test residuals for each code), we could identify the distinct profiles of high- and low-load episodes (i.e., what where the video codes appearing typically in one case or the other): *high-load episodes* had a higher chance of being transition/task distribution activities, and to occur at the classroom-level, featuring the students' faces (e.g., when explaining) or backs (e.g., when monitoring the progress of activities), or the teacher's own desk, which was cluttered with multiple paper elements to be distributed to the student groups as they progressed along the lesson activities. In contrast, *low-load episodes* occurred almost exclusively at the group social plane, while the teacher was focusing exclusively on one student tabletop.

The results of this study illustrate that it is feasible to use eye-tracking in a semi-authentic CSCL situation (e.g., the calibration procedure needed at the beginning of the eye-tracker recording could easily fit at the beginning of the lesson), and that the load index calculated using such techniques can be used to distinguish high/low load episodes. This enables focusing the researchers' attention in a smaller number of critical classroom usability episodes with distinct profiles (e.g., classroom-level activities have higher chance of high cognitive load). Furthermore, the study provided certain insights into classroom usability aspects of the specific CSCL technology used in this classroom: the need for classroom-level monitoring support in multi-tabletop classrooms and the dangers of a cluttered augmented paper user interface.

Authentic study: Master-level university course

After the promising results of the previous two studies, we set out to explore the following research questions: *is it feasible to use mobile eye-tracker to follow cognitive load in an authentic classroom/lesson?* But also, since eye movement patterns can vary greatly from person to person (Paas et al., 2003) and cognitive overload is known to be related to teaching expertise (Feldon, 2007), we aimed at exploring a second question: *do teachers with different teaching experience show different load episode patterns?*

Context and method

This study was performed in the context of a real master-level course at our university, on the topic of learning analytics and digital education. In this course several teachers and teaching assistants facilitate the different course sessions, which often combine lecturing and collaborative problem solving. We selected two of these teachers (one with more than ten years of teaching experience, the other a teaching assistant with two years of sporadic teaching assistantship), and recorded three course sessions of 45-65 minutes, with 12-14 students attending the class (two sessions for the expert teacher, one for the novice teacher). All sessions interspersed explanation/lecturing on the part of the teacher, with student individual and collaborative work (using laptops), and later debriefing of student outcomes. Thus, they represented the variety of situations and activities that often appear in authentic classroom settings.

The data gathering and analysis followed the same schema as for the previous study: recording of eye-tracking data using a mobile eye-tracker, calculation of the load index over 10-second episodes. From this load index, maximum and minimum load episodes were extracted and video-coded (along the same three dimensions of activity, social plane and main gaze focus). Then, statistical tests were run on the resulting video codes to check for the significance of the differences among video code distributions of high- and low-load episodes (both overall, and at the participant-teacher level).

Results

Again, a chi-squared test of independence of the *overall* video coding for the episodes with minimum and maximum load index (n=242) revealed that high-load and low-load episodes had statistically significant differentiated profiles in all three coding dimensions: teacher's activity (χ^2 = 9.904, df = 4, p = 0.04208), the activity's social plane (χ^2 = 14.8271, df = 1, p = 0.0001178) and the teacher's main focus during the episode (χ^2 = 45.2066, df = 7, p = 1.247e-07). In this case, looking at the most significant residuals of the chi-squared test, the profile of the *high-load episodes* again featured more often classroom-level activities, the monitoring of student work, and a focus on student faces, backs, or the whiteboard. *Low-load* episodes were most often repairs (e.g., solving a student doubt) at the individual or group level, when looking at the teachers' own computer/desk.

However, if we analyze the high- and low-load episodes at the level of *each participant teacher* (and the statistical test's residuals), we find interesting commonalities and differences: while for the novice teacher the high-load episodes gather more clearly under the classroom-level social plane (p = 6.992e-05), for the expert teacher such trend, while present, is *not* statistically significant (p = 0.2445). The same occurs for the teacher activity: while for the novice teacher the high-load episodes often feature monitoring, explanation or transitions, and the low-load ones are most often repairs (p = 0.000269), for the expert teacher such a trend is *not* significant (p = 0.4959). Regarding the main focus of teacher's gaze during the episodes, in both cases the differences are statistically significant (p = 0.00132 for the expert teacher, and p = 1.966e-06 for the novice one), with high-load episodes focusing on student faces or on the whiteboard, while low-load ones are most often focusing on the teacher computer or the projector (the residuals, in all cases, are smaller in the case of the expert teacher).

The results of this third study confirm some of the findings of the previous studies in this series (the increased load of classroom-level activities, or in trying to read students' faces, probably to assess their understanding). They also served to spot cognitive load patterns between an expert and a novice teacher (novice's high CL episodes being more concentrated in distinct kinds of activities/focus than the expert's). Finally, this study further confirms that this method for following CL of a teacher/facilitator is usable in authentic situations (although it requires the presence of a researcher/assistant for calibrating the device).

Discussion and future work

Overall, the three aforementioned studies show that the proposed combination of mobile eye-tracking measurements can be feasibly used in authentic classroom settings (2). Together with post-hoc qualitative video coding, the approach also showed potential for discriminating fine-grained critical episodes (either high- or low-load) during the CSCL enactment, without having to rely on the teacher's memory of the events. The distinct profiles of such episodes have helped us gain insights into the difficulties of orchestrating CSCL activities. Some of these results had already been anticipated by existing work on orchestration load: the need for awareness/monitoring support in multi-tabletop classrooms (Kharrufa et al., 2013), and the importance of classroom-level awareness in general (e.g., Cuendet et al., 2013), the challenge that clutter and a scattered user interface pose in augmented paper applications (Cuendet & Dillenbourg, 2013), etc. The results also show that the facilitation CL is highly dependent on the teacher's prior experience (also anticipated by Feldon, 2007).

Our results across the three studies also provide a new insight about orchestration load: the fact that it seems to be correlated to the amount of "open alternatives" (i.e., the perceived uncertainty/entropy of the classroom situation). More novice teachers seemed to be especially sensitive to these high-uncertainty situations (e.g., looking at students faces during an explanation, trying to assess their comprehension; looking at students' backs while working in the tabletops, trying to assess their progress). This need of novice teachers for classroom management support had already been mentioned in recent works like Raca & Dillenbourg (2013), and can be used as a starting point for new technologies that can ameliorate the challenge of this kind of episodes (e.g., a system that helps novice lecturers to assess the attention or comprehension of their students).

The present work, however, also presents several limitations, and the proposed method for estimating CL in facilitation of CSCL should be understood as only a first approximation. One limitation that this work shares with most research on orchestration is the limited number of participant teachers, which begs the question of the generalizability of these results: in contrast with learning in the classroom, teaching enactment is often a solitary endeavor. This problem, however, is difficult to solve, unless we resort to the sharing of datasets to help replicability and the accumulation of empirical evidence in this area. Another limitation is the cost of the mobile eye-tracking instruments themselves, and the fact that its initial calibration requires the presence of a researcher/assistant: therefore, it can be difficult to use on in-the-wild longitudinal studies. Also, we should note that the eye-tracking data analysis performed in this paper (featuring a session median cut), while more robust to classroom conditions that may vary from day to day (or hour to hour), will only capture relative load among different episodes *in the same session* (not absolute measures of load, or comparisons across sessions): a low-load episode in a difficult session might be actually more loading than a high-load one on an easy session.

This last weakness points us to the most immediate next steps in this research direction: the combination of these objective measures of CL with other complementary methods, e.g., to assess the overall load of the session using self-report questionnaires. Also, more nuanced analysis of CL could be performed by using more complex quantiles and/or different weights in the calculation of the load index. Another interesting path to follow is to attempt to separate the different components of CL, a problem that still proves difficult (Paas et al., 2003), and even more so in uncontrolled environments and complex tasks like teaching. The combination of different CL measurement techniques also has shown potential to help in this regard (DeLeeuw & Mayer, 2008). Last but not least, we aim to use this method (or its future enhancements) as part of our design-based research on the application of augmented paper in the orchestration of authentic primary school classrooms.

With this work, we add a new instrument to the CSCL researcher's toolkit to design and evaluate the material conditions of CSCL and novel proposed technologies, in authentic settings. However, this work could have implications for the design of pedagogical interventions as well: assuming equal potential for learning, are two classroom scripts equally easy to manage? In this area, considering orchestration cognitive load could help us make informed design decisions that could help drive teacher adoption of CSCL practices.

Endnotes

(1) The interested reader (and to facilitate study replicability) can find the pre-processing, analytic and visualization scripts, as well as further analysis results (and pointers to the available raw anonymous datasets) for this and the other studies described in this paper, at https://github.com/chili-epfl/cscl2015-eyetracking-orchestration.

(2) There might be concerns that the mere fact of using the eye-tracking device might have rendered the learning situations "un-authentic". However, the participant teachers reported "forgetting about them", and students themselves made few remarks about it in the first minutes of the lessons, and none more afterwards, until the lesson finished.

References

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53-61.

Buettner, R. (2013). Cognitive Workload of Humans using Artificial Intelligence systems: Towards Objective Measurement applying Eye-Tracking Technology. In *KI 2013: Advances in Artificial Intelligence* (pp. 37-48). Springer Berlin Heidelberg.

Caballero, D., Wen, Y., Prieto, L. P. & Dillenbourg, P. (2014). Single locus of control in a tangible paper-based tabletop application: an exploratory study. *Paper presented at the International Conference on Interactive Tabletops and Surfaces (ITS'14).*

Chan, C. K. (2011). Bridging research and practice: Implementing and sustaining knowledge building in Hong Kong classrooms. *International Journal of Computer-Supported Collaborative Learning, 6*(2), 147-186.

Cuendet, S., Bonnard, Q., Do-Lenh, S., & Dillenbourg, P. (2013). Designing augmented reality for the classroom. *Computers & Education, 68*, 557-569.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer-supported collaborative learning. In *Technology-Enhanced Learning* (pp. 3-19). Springer Netherlands.

Dillenbourg, P., Zufferey, G., Alavi, H., Jermann, P., Do-Lenh, S., Bonnard, Q., ... & Kaplan, F. (2011). Classroom orchestration: The third circle of usability. *CSCL2011 Proceedings – Vol. I*, 510-517.

Feldon, D. F. (2007). Cognitive load and classroom teaching: The double-edged sword of automaticity. *Educational Psychologist, 42*(3), 123-137.

Gómez, F., Nussbaum, M., Weitz, J. F., Lopez, X., Mena, J., & Torres, A. (2013). Co-located single display collaborative learning for early childhood education. *International Journal of Computer Supported Collaborative Learning, 8*(2), 225–244.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Kahneman, D., Beatty, J.: Pupil Diameter and Load on Memory. *Science, 154*(3756) (1966) 1583–1585.

Kharrufa, A., Martinez-Maldonado, R., Kay, J., & Olivier, P. (2013, October). Extending tabletop application design to the classroom. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces* (pp. 115-124). ACM.

Muñoz-Cristóbal, J. A., Prieto, L. P., Asensio-Pérez, J. I., Jorrín-Abellán, I. M., Martínez-Monés, A., & Dimitriadis, Y. (2013). Sharing the Burden: Introducing Student-Centered Orchestration in Across-Spaces Learning Situations. In *Scaling up Learning for Sustained Impact* (pp. 621-622). Springer Berlin Heidelberg.

Onrubia, J., & Engel, A. (2012). The role of teacher assistance on the effects of a macro-script in collaborative writing tasks. *International Journal of Computer Supported Collaborative Learning, 7*(1), 161–186.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Paas, F., Renkel, A., & Sweller, J. (2004). Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture. *Instructional Science, 32*, 1–8.

Perez-Sanagustin, M., Hernandez-Leo, D. & Blat, J. (2009) Towards supporting orchestrated computer supported collaborative learning scenarios. *IEEE Multidisciplinary Engineering Education, 4*(3):83–88.

Prieto, L. P., Dlab, M. H., Gutiérrez, I., Abdulwahed, M., & Balid, W. (2011). Orchestrating technology enhanced learning: a literature review and a conceptual framework. *International Journal of Technology Enhanced Learning, 3*(6), 583-598.

Prieto, L. P., Wen, Y., Caballero, D., Sharma, K. & Dillenbourg, P. (2014). Studying teacher cognitive load in multi-tabletop classrooms using mobile eye-tracking. *Paper presented at the International Conference on Interactive Tabletops and Surfaces (ITS'14)*.

Raca, M., & Dillenbourg, P. (2013). System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 265-269). ACM.

Roschelle, J., Dimitriadis, Y., & Hoppe, U. (2013). Classroom orchestration: synthesis. *Computers & Education, 69*, 523-526.

Senn, J, and Goette, L. (in progress). *Piece rate vs. team rewards in interdependent tasks: What can we learn from TETRIS?* Unpublished study.

Song, Y., & Looi, C.-K. (2012). Linking teacher beliefs, practices and student inquiry-based learning in a CSCL environment: A tale of two teachers. *International Journal of Computer Supported Collaborative Learning, 7*(1), 129–159.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. Educational psychology review, 17(2), 147-177.

Acknowledgments

This research was supported by a Marie Curie Fellowship within the 7th European Community Framework Programme (MIOCTI, FP7-PEOPLE-2012-IEF project no. 327384). We would like to thank all the teachers and students that agreed to participate in the study, and especially Patrick Jermann, for his feedback and insights.