Studying Teacher Orchestration Load in Technology-Enhanced Classrooms

A Mixed-Method Approach and Case Study

Luis P. Prieto(✉), Kshitij Sharma, and Pierre Dillenbourg

CHILI Lab, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{luis.prieto,kshitij.sharma,pierre.dillenbourg}@epfl.ch

Abstract. Teacher orchestration of technology-enhanced learning (TEL) processes plays a major role in students' outcomes, especially in face-to-face classrooms. However, few studies look into the fine-grained details of how such orchestration unfolds, the challenges and cognitive overload that using technologies at a classroom level pose for teachers. This paper proposes a mixed-method approach to the study of orchestration cognitive load, combining physio-behavioural (eye-tracking) and subjective measurements (questionnaires, stimulated recall interviews). We illustrate the approach by applying it to study the orchestration of two technology-enhanced geometry lessons, by a secondary school teacher. The results of our mixed-method analyses highlight the difficulty of classroom-level (as opposed to individual- or group-level) interactions, especially in modelling students' progress and understanding. Such insights can be useful in the design of new classroom technologies, and to focus researchers' attention on critical orchestration episodes during their evaluation.

Keywords: Orchestration load · Eye-tracking · Stimulated recall · Cognitive load · Classroom studies

1 Introduction

Teacher facilitation of the learning process plays a crucial role in students' learning outcomes, especially in co-located settings (see, e.g., [14,19]). TEL researchers have recognized this importance, and the unique challenges that practitioners face when applying technological innovations to their everyday practice, under the term 'orchestration' [22], often defined as "the process of productively coordinating supportive interventions across multiple learning activities occurring at multiple social levels" [11].

Although the term 'orchestration' has rapidly gained traction within the TEL research community, its current use is rather varied, even confusing (e.g., [22] identifies up to eight different aspects TEL researchers refer to when talking about orchestration). Taking the metaphor of orchestration as "classroom-level", or "third circle" usability (individual and small-group usability being

G. Conole et al. (Eds.): EC-TEL 2015, LNCS 9307, pp. 268–281, 2015.
DOI: 10.1007/978-3-319-24258-3_20

the first two [12]), we find that classroom-level technology usability studies are still in its infancy, undoubtedly due to the technical and logistic difficulties of doing research about such a complex activity, within the multiple constraints of authentic classroom conditions [25].

Current studies on the orchestration of novel educational technologies are normally based on ad-hoc quantitative measurements of efficiency (e.g., [1]), or on the combination of keen researcher observation and subjective questioning of the actors (e.g., [17]), majoritarily looking at whole sessions as the unit of analysis (i.e., "did the lesson work well?"). More recently, there have also been attempts to do finer-grained quantitative analyses of orchestration, using mobile eye-tracking [23,24]. However, we are still far from the current state on individual usability studies (e.g., website usability), in which the researcher can rely on an array of qualitative and quantitative methods to study the (admittedly, simpler) user tasks of interest, both from behavioral/physiological, and subjective perspectives. Paraphrasing Bob Bailey in his account of early usability studies in the 1970s, most classroom orchestration studies are "focused simply on determining if three or four participants were able to complete the tasks" [3].

This paper proposes *a novel approach to study instances of technology-enhanced, face-to-face orchestration, combining multiple methods for data gathering and analysis.* Our ultimate goal with this approach is to gain a deeper understanding of the orchestration process at different granularity levels, with the aim of obtaining insights to guide the design (and aid in the evaluation) of new classroom technologies that can overcome the unique challenges of orchestrating such classrooms.

After a section describing related work on orchestration studies, orchestration load and cognitive load as an important factor in it, we present our mixed-methods approach to the study of orchestration instances. The approach is illustrated through its application on a case study that recently took place in a school in Switzerland, in which a teacher conducted two 80 min lessons on geometry (with different groups of students) using laptops and other common classroom technologies. We conclude the paper with a discussion of the main implications of our results and future research lines derivating from this work.

2 Related Work

2.1 Orchestration and Usability at the Classroom Level

The complex challenges of classroom management are by no means a new topic in educational literature. The classroom has always been recognized as a highly-demanding public space with its own history, where multiple activities take place simultaneously, with high immediacy and unpredictability [13]. Once digital technologies started being integrated in the classroom, it became clear that technology not always provided solutions to these pressures, also adding a new layer of complexity to them. In TEL, this notion of managing multiple activities at different social levels using multiple technological tools, often in a formal education setting, has become known as'orchestration' [11,22,25].

Dillenbourg and colleagues, from a technology designer perspective, have put forward the notion of 'designing for orchestration', drawing the parallelism with the notions of usability for individuals (long studied by human-computer interaction) and for small groups (focus of computer-supported collaborative work/learning). Thus, classroom technologies should be designed to cope with the pressures and constraints of whole-classroom usage [12]. In this line, several authors have started putting forward guidelines for the design of classroom technologies, drawing from their experience testing technologies in authentic educational settings [8,10,17]: the establishment of mutual and class-wide awareness mechanisms, providing teachers with means of controlling the flow of activities and improvisedly adapting it, or dividing activities into sequences of phases, are some of the most commonly appearing ones.

These insights and guidelines are invariably derived from keen observation and reflection of the researcher team after trying out a certain technology or set of technologies for a few times in authentic classroom conditions. While this kind of expert advice is certainly very useful when designing new technologies for use in the classroom, as TEL researchers we still lack a systematic way of studying the process of orchestration in any technology-enhanced classroom, beyond reflecting (or asking the teacher) about 'how the session went'. As usability pioneers Gould and Lewis stated, usable computer systems require (1) an early and continual focus on users; (2) iterative design and testing; and (3) empirical measurement of usage [15]. The first two are already a part of how many TEL researchers approach technological innovation in the classroom. The third one is made very difficult by the simultaneity and immediacy of face-to-face classroom activities. If we want to produce technologies that are usable while performing such multidimensional task, gathering empirical evidence of how the process unfolds, not only at the temporal scale of a 60 min session, but also in more fine-grained critical episodes of minutes or even seconds will prove unvaluable (what Newell denominates the 'rational' and 'cognitive' bands of action [18]).

2.2 Orchestration Load, Cognitive Load and Its Measurement

One prominent notion appearing frequently in recent TEL literature is 'orchestration load', that is, "the effort necessary for the teacher – and other actors – to conduct learning activities" at the class-wide level [8]. Currently, orchestration load is still a fuzzy and abstract concept, but we can safely assert that it has a physical component (e.g., moving around the class, handing out worksheets to students) and a cognitive side as well (remembering the content to be explained, modelling students' learning progress, deciding how to adapt the lesson plan). In a TEL classroom, the cognitive part of orchestration is a more interesting focus of study. Fortunately, cognitive load (related to the executive control of working memory, and the limited capacity of human cognitive processing capacity [20]) has been widely studied in psychology and human-computer interaction.

Measuring Cognitive Load. We can find in the literature multiple methods to estimate cognitive load of a task, including subjective (e.g., questionnaires) and objective ones (e.g., dual-task performance), as well as direct (e.g., brain imaging) or indirect measurements (e.g., heart-rate monitoring) [5]. However, most of these measurements have been utilized in relatively simple, single-user tasks – not multi-dimensional and highly social ones, such as the orchestration of a classroom.

Indeed, the aforementioned characteristics of face-to-face classroom management tasks (simultaneity, immediacy, etc.) greatly restrict the variety of techniques that can be used to measure the orchestration cognitive load. For instance, subjective questionnaires can be used, but require continuous interruptions during the lesson, or relying on the subject's memory of a very long period (although this could be mitigated by using 'stimulated recall' interviews where the teacher is shown videos from the lesson). Reliable brain imaging is currently impractical given the mobility of classroom facilitation, and dual-task performance measures would be too obstrusive as they would add another task to an already multitask situation [21]. Physiological measures of cognitive load, such as pupillary response and other measurements detectable by eye-tracking are normally taken in carefully-controled lightning conditions (a luxury we often do not have in a classroom), and might be confounded by certain visual features of the setting (e.g., a wide spread of students in the visual field of the teacher might make average saccade speed higher). However, in previous works, we have shown the feasibility of using the triangulation across different measurements using mobile eye-tracking to record teacher orchestration *in situ*, approximating relative cognitive load within a session [23,24].

Multi-method Approaches. As we can see, there is no single method which can be reliably used to estimate the cognitive load of orchestration, at different temporal granularities, within the constraints of a classroom setting. An alternative is to take a pragmatic stance and use a mixed-methods approach [7], in which different qualitative and quantitative data gathering and analysis techniques (including both behavioral/physiological and subjective information) are triangulated to understand how the orchestration process (and its cognitive load) unfolds, from the second-to-second actions to the overall result of an hour-long session.

To the best of our knowledge, no other mixed-method approaches exist to systematically study the phenomenon of classroom orchestration. However, this sort of multi-method approaches are not unhead-of in areas related to cognitive load studies: [9] combined dual-task and subjective measurements of cognitive load to separate different cognitive load components in the context of multimedia learning tasks; [4] combined stimulated recall interviews and measurements of cognitive load to obtain a more comprehensive view of the effects of an ICT tool in learning. In the area of eye-tracking studies, [6] used four different eye-tracking measurements to distinguish cognitive load in a learning task supported by artifical intelligence.

3 Multi-method Approach to Study Teacher Orchestration

Against this backdrop, it can be seen that “empirically measuring” the multi-layered, time-constrained orchestration process is not an easy task, and probably cannot be accomplished using a single method or technique. Our aim thus is to propose a combination of methods that can be used consistently to track and better understand the unfolding and critical episodes of the orchestration of a lesson (typically spanning from 40 min to 2 h) and, if possible, to make meaningful comparisons among different sessions (as long as many of the sessions' elements, like the teacher, the kind of classroom, or the general activity structure, remain unchanged). Furthermore, our method intends to be feasible to use within relatively short research cycles and limited manpower (as iterativity is a common requirement of most current technology design/research approaches), by focusing the researchers' attention on the most critical orchestration episodes.

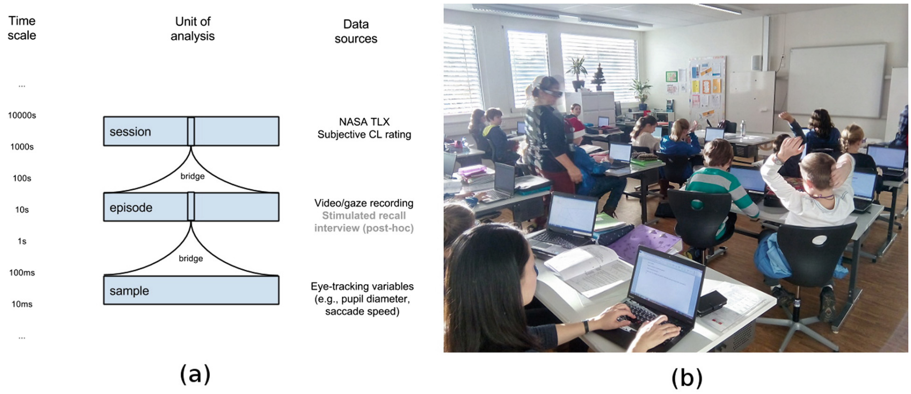


Fig. 1. (a) Units of analysis and data sources used in the proposed mixed-method approach. (b) Classroom disposition during one of the case study sessions.

Since we try to understand the orchestration process at different granularity levels, we have selected different data gathering techniques, best suited for each temporal scale and applicable within classroom constraints (see Fig. 1a): session-level subjective ratings of cognitive load (including both quantitative questionnaires such as NASA's TLX [16] and open-ended questions), first-person video/gaze recordings that can be analyzed by researchers, as well as used in stimulated recall interviews with the teacher (for the 10 s episode level). Finally, fine-grained physiological measures from the in-situ usage of a mobile eye-tracker provide information at the milliseconds scale. It is also important to note that, in order to understand how the teacher actions at these different scales relate to each other, analyzing the data at each scale has to be complemented with analysis processes that implement “bridges” among scales, aggregating smaller units of analysis to build meaning at higher-level ones [2]. The data analysis process is as follows:

1. *From Samples to Episodes (Eye-tracking).* Four eye-tracking measurements which have been related with cognitive load (pupil diameter mean, pupil diameter standard deviation, number of long fixations and average saccade speed, see [6]) are calculated throughout the session. Each of these four measurements is aggregated into 10 s episodes (thus bridging from sample- to episode-level), and the median value of these aggregated measurements for each session is calculated. By performing a median cut (i.e., for each 10 s episode, how many of the four measurements are over the session's median), a "load index" ranging from 0 to 4 is calculated, indicating the likelihood that a certain 10 s episode represented a higher load than the session average (see [23,24] for a more detailed explanation of this process).
2. *Subjective Episode Analysis (Stimulated Recall).* From this collection of 10 s episodes (and their associated load index based on eye-tracking measures), a subset is selected from consistently high- or low-load periods during the session, and 10 s video snippets of the subjective camera view of the teacher are generated for each selected episode. These videos are then used in a post-hoc stimulated recall interview with the teacher, in which the teacher is asked to rate subjectively each snippet using a standard 9-level mental effort scale, using a think-aloud protocol to express the rationale behind each rating. The numerical ratings are used to triangulate the load index obtained using eye-tracking, and the think-aloud output goes through qualitative analysis for triangulation and interpretation at larger granularity levels.
3. *Objective Episode Analysis (Video Coding).* All the extreme load index episodes (ELEs, i.e., those 10 s episodes with values 0 or 4) are then video coded by the researcher team[1], along three main dimensions characterizing orchestration (as per [11]'s definition of orchestration): the *activity* being performed by the teacher (e.g., explanation, monitoring), the *social plane* of the interaction (e.g., class-wide, with individual students) and the *main focus of the teacher's gaze* (a student laptop, students' faces, etc.). The video code counts are then aggregated for the whole session (thus bridging from the episode- to the session-level), and statistical tests (Pearson's chi-squared) are used to determine which coding dimensions and which video codes contribute most significantly to the differences between high- and low-load episodes, and using these codes to create distinct profiles for each.
4. *Session-level Analysis.* Finally, the session-level subjective ratings provided by the teacher (both quantitative scales and qualitative open responses) are triangulated with the qualitative data coming from the stimulated recall interview, to understand the overall perception of (cognitive) orchestration load at

[1] We restrict ourselves to the extreme load episodes as a way to keep the needed manpower under control (in a hour-long session the number of extreme load episodes to code tends to be around 100, while the total number of episodes can be close to 1000), and to focus the attention on the most critical orchestration episodes, contrasting them with the 'least critical' ones [23].

the session-level. These ratings can also be used to make comparisons among different sessions (e.g., which of two sessions was more difficult, and why). These inter-session comparisons can also be triangulated by comparing the medians used for the load index cut in step 1 above (after normalization by the first-episode values to cancel out physiological or ambiental differences among different sessions).

4 Case Study: Orchestrating a Technology-Enhanced Geometry Class

To illustrate our mixed-methods approach to study teacher orchestration, we provide an account of how it was applied in a small case study in secondary education. This study is part of a larger design-based effort to understand and enhance orchestration in classrooms using augmented paper interfaces. Within this larger project, however, the present study was intended as a test-drive of the approach, and to gather evidence about orchestration processes in the teacher's *usual* technology-enhanced practice. Hence, the goal of the study is not to prove or disprove that the technology worked better than an alternative – rather, it is to provide a deep comprehension of what was the teacher/classroom's baseline of technology usage. Further details, raw datasets and analytical code to generate the results of this study are publicly available online[2].

4.1 Context of the Study

The study comprised two secondary education geometry sessions that took place in an international school near Lausanne (Switzerland). The two sessions followed the same general structure, and were run by the same teacher (a practitioner with more than 15 years of teaching experience), with two different cohorts of students, of 22 and 23 students respectively (aged 11–12 years old). During the 80 min sessions, the teacher guided the students in individual and dyad work about geometrical figures and tessellations, using laptops and specialized geometry software (Geometer's Sketchpad[3]), interspersed with small periods of explanation/lecturing. Students thus were mostly working at their own laptops, following a paper worksheet that guided them through different exercises using the geometry software, attending to occasional explanations by the teacher, and asking questions when some exercise or concept was unclear. To support her orchestration of the lesson, the teacher was using a projector connected to her computer (see Fig. 1b), and the school's usual software for classroom management (NetSupport School[4]). During the two sessions, data was gathered from multiple data sources, and later analyzed as described in Sect. 3.

[2] https://github.com/chili-epfl/ectel2015-orchestration-school.

[3] http://www.keycurriculum.com/.

[4] http://www.netsupportschool.com/.

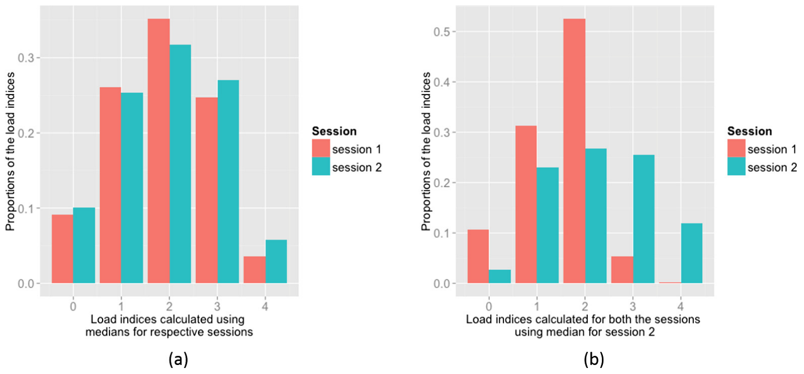


Fig. 2. Proportions of load indices calculated from eye-tracking measures. (a) Load indices calculated using each session's respective median. (b) Load indices calculated using session 2's median applied over both sessions; in this case load indices of session 1 are clearly skewed towards lower values, suggesting that session 2 was more difficult to orchestrate.

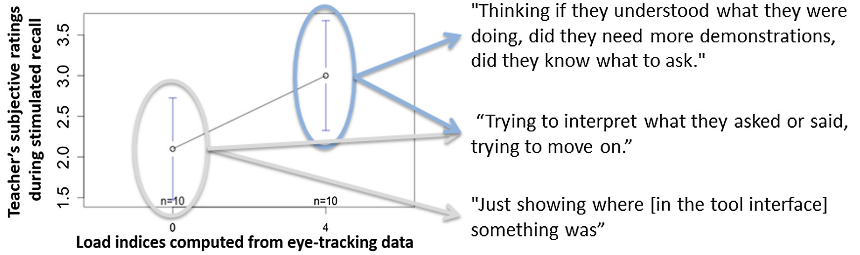
4.2 Results

From Samples to Episodes. Using the eye-tracking data from the sessions, we calculated the four eye-tracking metrics related to cognitive load (from [6]), and aggregated them (through averaging or counting) into 10 s episodes, using a rolling window with 5 s slide, thus obtaining the "load index" for each episode (i.e., likelihood that an episode is higher cognitive load than others in the session, from 0–4). As we can see in Fig. 2a, in both sessions the distribution of load indices is approximately gaussian (which is to be expected, given the way such index is calculated with respect to the session median for each eye-tracking metric).

Subjective Episode Analysis. From the 10 s episodes generated in the previous step, 20 were selected from sustained high-/low-load periods. The subjective video feed from the teacher's eyetracker for each of these episodes was extracted, and the teacher was asked to rate her mental effort (from 1–9) in each one and provide a rationale for each value, in a post-hoc stimulated recall interview.

We then triangulated the teacher's subjective ratings with the load indices computed from the eye-tracking data. We found that the teacher's ratings were significantly lower for the low load episodes than those for the high load episodes (F[1, 18] = 4.89, p = .04, see Fig. 3a), even if the correlation between such subjective ratings and the eye-tracking load indices was far from perfect.

A qualitative coding of the think-aloud protocol followed during the interview (see Fig. 3b) shows that in the higher-load episodes (both subjective and based on eye-tracking) teacher often mentions attempts to model students' understanding, or the need to take a decision about what would be the best immediate course of action. On the other hand, in lower-load episodes (from subjective rating and eye-tracking index), the individual explanations after a student error (what we could term 'repairs'), and the technical details or problems in using the technological



"Thinking if they understood what they were doing, did they need more demonstrations, did they know what to ask."

"Trying to interpret what they asked or said, trying to move on."

"Just showing where [in the tool interface] something was"

Fig. 3. Comparison of subjective load ratings and eye-tracking load index (means and confidence intervals), from a sample of 20 low (0) and high (4) episodes *(left)*. Examples of teacher quotes when explaining the rationale for the subjective ratings *(right)*.

tools appear more often. Interestingly, the need to make the lesson progress according to plan (i.e., be within the time/curriculum constraints) appears very often in high-load moments, but also appears in low-load ones (maybe indicating a background concern at all times, even in the lower-load moments)

Objective Episode Analysis. We manually coded the videos of all the ELEs (10 s episodes with an eye-tracking load index of 0 or 4), along the dimensions of teacher activity, social plane of the interaction and main focus of gaze. By aggregating the code counts for each of these dimensions in a session we can bridge from the episode-level analysis to have an idea of what kind of episodes are often high or low load (see Table 1). We can find several *common trends* in the orchestration load episode profiles of the two sessions:

- The high-load episodes are characterised by the teacher giving explanations/lecturing (EXP) and asking questions (QUE) (χ^2 [DoF = 5] = 131.14, $p < .001$); the social plane being whole-class (CLS) (χ^2 [DoF = 1] = 90.89, $p < .001$); and the main focus of the teacher being the faces of the students (FAC) (χ^2 [DoF = 3] = 161.95, $p < .001$).
- On the other hand, low-load episodes in both sessions are characterised by the teacher doing 'repairs' (REP – i.e., solving a question or misunderstanding of a student) more often (χ^2 [DoF = 5] = 131.14, $p < .001$); the social plane being individual/small-group (GRP) (χ^2 [DoF = 1] = 90.89, $p < .001$); and the main focus of the teacher being the activity paper sheets of the students (PAP) (χ^2 [DoF = 3] = 161.95, $p < .001$).

By looking at the video codes, however, we can also detect a few *differences between the two sessions* (see Table 1). Compared to the first session, in the second session we can see the appearance of a new category of episodes (disciplinary remarks, DISC), in the high-load episodes. Also, we can observe that the significance of differences in several of the video codes is increased for the second session: the predominance of repairs in low-load episodes, higher appearance of focus on paper worksheets in low-load episodes as well as the relative absence of focus on the students' laptops on high-load episodes.

Table 1. Video code counts of extreme (high and low) load episodes, for each dimension of coding. In parentheses, the standard Chi-squared residuals (absolute residual values larger than 1.96 are considered significant, marked with an '*'). Differences in significance between the two sessions are highlighted **in bold**

Session/	Activity						Social plane		Main focus of gaze			
Load	DISC	EXP	MON	QUE	REP	TDT	CLS	GRP	FAC	LAP	PAP	TCOMP
1 – Low	0	1 *	14	0 *	61	2	13 *	25 *	1 *	40	25	12
	(NA)	(-2.15)	(0.49)	(-2.94)	(1.81)	(-0.11)	(-2.75)	(2.03)	(-4.01)	(1.17)	(1.63)	(1.13)
1 – High	0	8 *	3	12 *	6 *	1	65 *	5 *	24 *	6	0 *	0
	(NA)	(3.47)	(-0.79)	(4.74)	(-2.92)	(0.18)	(4.44)	(-3.27)	(6.47)	(-1.89)	(-2.63)	(-1.82)
2 – Low	0	2 *	9	0 *	**66** *	10	15 *	36 *	0 *	48	**28** *	11
	(-1.41)	(-2.98)	(0.12)	(-2.57)	**(2.40)**	(0.71)	(-3.24)	(2.58)	(-4.60)	(1.77)	**(2.18)**	(0.32)
2 – High	**3** *	14 *	4	10 *	8 *	2	72 *	8 *	32 *	**8** *	0 *	4
	(1.98)	(4.20)	(-0.17)	(3.02)	(-3.38)	(-1.01)	(4.44)	(-3.64)	(6.48)	**(-2.49)**	(-3.06)	(-0.46)

Session-Level Analysis. Moving onto the overall session-level cognitive load analysis (and the eventual comparison between the two sessions), we can take a look at the subjective ratings provided by the teacher at the end of each session. In both cases the teacher assessed the overall load as involving 'some mental effort' (6 out of 9 in the standard subjective mental effort scale [5]). The teacher, however, considered the second session as slightly more difficult to manage (6 vs. 5 out of 9). This perception can be expanded by looking at the teacher's responses to the TLX questionnaire [16] about the sessions workload: In both cases responses were quite similar (overall workload scores of 53.3 and 56.3 out of 100 for sessions 1 and 2, respectively). However, by looking at the scores and weighting of the different workload components (mental, temporal or physical demands, performance, effort and frustration), we see that in session 2 the value and weighting of the 'frustration' component were much higher (value=5, weight=4 in session 2, vs. value=2.5, weight=1 in session 2), which also accounts for the larger part of the increase in overall perceived workload.

The qualitative (open) responses about self-perceived high-load episodes were also quite similar in both sessions, referring to worries about student progress ("when students could not follow the directions and were confused as what to do next"). However, we can find that only in the second session the teacher referred to the aforementioned disciplinary concerns ("when students try to log off while demonstrating"), which may help explain the higher frustration component of that session.

To triangulate this perception with more objective data, we can also look at the eye-tracking measures. If we apply the median values of the four eye-tracking measures of interest[5] of session 2 to the data in session 1, we can obtain a quantitative approximation to the relative difficulty between the two sessions. As we can see in Fig. 2b, in this case the distribution of load indices for session 1

[5] After normalization by the first 10 s episode of each session, to account for variability in the data due to the different time of the day or wakefulness of the teacher in the concerned days.

is skewed towards the lower end (thus indicating that fewer episodes in session 1 were as high a load as the high-load episodes in session 2, χ^2 [DoF = 4] = 344.29, $p < .001$). This again supports the idea that session 2 was (overall) higher load than session 1.

5 Discussion

The results of the different analysis methods outlined above provide us with a very detailed view of how the orchestration of these secondary school geometry lessons took place, from triangulated empirical evidence (both behavioral and subjective). The two sessions shared similar profiles in terms of high- and low-load episodes: high load episodes tended to be explanations or questioning (or the occasional disciplinary remark) in the class-wide plane, very often looking at the faces of students (in an attempt to assess their progress and understanging). On the other hand, low-load episodes tended to be individual or small-group repairs, with the teacher often focusing on the students paper worksheets or laptops. These similarities across sessions are to be expected given the fact that only the cohorts were different between the lessons, and the rest of the elements (teacher, classroom layout, technologies and subject matter) remained essentially the same.

Our triangulated data sources also helped us find orchestration load differences between these very similar sessions: both the subjective session-level data and eye-tracking metrics confirm a higher difficulty of session 2. The presence of more accentuated differences in the profiles of high- and low-load episodes seems to be a hallmark of sessions that are more difficult to manage. In this particular case, it looks like the presence of more disciplinary remarks and concerns by the teacher was one of the main drivers of this increase in orchestration load (especially, of its frustration component).

These results largely confirm several trends already observed in our previous work using mobile eye-trackers to follow orchestration load [23,24]: the higher load that class-level interactions pose for teachers, or the fact that assessing student progress across the classroom is also a high-load activity (often represented by looking at the students' faces, in the absence of other glanceable information), or the fact that more marked statistical differences in the high-/low-load episode profiles seem to be correlated with more difficult sessions (as in [23]'s comparison between a novice and an expert teacher). It is interesting to note that these common trends are preserved even across very different educational settings and technologies (small master-level classrooms with laptops, primary school tabletop classrooms, or mid-sized secondary school classrooms). Compared with this previous work to understand classroom orchestration in deeper detail, the method proposed here allows for better explanatory power by adding subjective data about the orchestration load, and also allows for meaningful comparisons among different sessions (at the expense of extra work by the research team).

An issue that might be surprising is the fact that most of the findings about the orchestration load do not seem very related to the technological support in the classroom. This can be partly explained by the fact that the teacher was using her usual set of technologies and setup (thus, in a sense, the technology was 'invisible' to her). However, this may not be the case if we were to compare technology-enhanced and traditional versions of same a lesson, or two different technologies applied to the same lesson. Also, we should consider that the technology might be playing a more subtle role in the orchestration load: the relatively low appearance of the teacher computer as the gaze of focus (where she could in theory access each student's screen through the classroom management software) seem to indicate that the teacher did not see great value in the way such monitoring was implemented[6], and preferred to roam physically around the classroom. Looking beyond what looks like a perfectly reasonable implementation of orchestration support features, into how it is used in actual classrooms, is one of our main aims when proposing the present mixed-method approach.

6 Implications and Future Work

After showing the current lack of methods to study in detail the cognitive load that facilitating a technology-enhanced classroom poses for teachers, we have proposed a mixed-method approach to such studies. The illustrative case study presented here shows how our approach enables the detection of common orchestration load trends across sessions and settings, but also how it detects and provides probable explanations for differences between two sessions' orchestration (which, let us remember, is very contextual and non-repeatable by nature).

The orchestration load trends detected by using our approach are largely compatible with much of the existing advice about designing technology for classroom orchestration. However, even if the insights derived from the application of this approach are not (yet) revolutionary, they provide empirical evidence to help technology designers *prioritize* which orchestration problems to work on first. In this concrete case, the results of our case study support the importance of research efforts in learning analytics and classroom-level visualizations, which are important not only in online environments, but also in face-to-face classrooms.

Naturally, the concrete results of this case study are hardly generalizable beyond the teacher, classroom technologies and activities described here. However, as more and more studies using these methods are performed, we expect to see several general trends about orchestration load emerge (indeed, some of these trends can already be seen throughout our previous and present work). In order to help other researchers to start performing such studies, and to enable the open discussion and improvement of these methods (e.g., more sophisticated methods of inferring cognitive load from eye-tracking data, the addition of other physiological measures), we have made the datasets and analytical code public.

[6] Teacher could see either a mosaic with the miniature version of all screens where little could be made out, or a single student's screen – which often provides an incomplete view unless the teacher actually walked there to talk with the student.

Also, we should note that these methods could be applied, not only to evaluate technological innovations in the classroom, but also pedagogical ones, from the point of view of the orchestration load they impose on the teacher (a factor often ignored by learning scientists).

Our own immediate future work includes the introduction of novel technologies in the classroom setting described here, and the observation of its effects on teacher orchestration load, as a way to help us in the design and evaluation of classroom technologies. Also, further validation and evolution of the mixed-method approach will be performed, by manipulating the conditions of similar classroom sessions (e.g., the number of facilitators) to provide some kind of"ground truth" to validate the selected physiological measurements (or the addition of new ones).

Acknowledgment. This research was supported by a Marie Curie Fellowship within the 7th European Community Framework Programme (MIOCTI, FP7–PEOPLE–2012–IEF project no. 327384). Special thanks to Wendy Farden and the International School of Lausanne for their collaboration.

References

1. Alavi, H.S., Dillenbourg, P.: An ambient awareness tool for supporting supervised collaborative problem solving. IEEE Trans. Learn. Technol. **5**(3), 264–274 (2012)
2. Anderson, J.R.: Spanning seven orders of magnitude: a challenge for cognitive modeling. Cogn. Sci. **26**(1), 85–112 (2002)
3. Bailey, B.: Usability testing: an early history. http://webusability.com/usability-testing-a-early-history/. Accessed on 17 March 2015
4. Beers, P.J., Boshuizen, H.P., Kirschner, P.A., Gijselaers, W., Westendorp, J.: Cognitive load measurements and stimulated recall interviews for studying the effects of information and communications technology. Edu. Technol. Res. Develop. **56**(3), 309–328 (2008)
5. Brunken, R., Plass, J.L., Leutner, D.: Direct measurement of cognitive load in multimedia learning. Edu. Psychol. **38**(1), 53–61 (2003)
6. Buettner, R.: Cognitive workload of humans using artificial intelligence systems: towards objective measurement applying eye-tracking technology. In: Timm, I.J., Thimm, M. (eds.) KI 2013. LNCS, vol. 8077, pp. 37–48. Springer, Heidelberg (2013)
7. Creswell, J.W.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage publications, Thousand Oaks (2013)
8. Cuendet, S., Bonnard, Q., Do-Lenh, S., Dillenbourg, P.: Designing augmented reality for the classroom. Comput. Edu. **68**, 557–569 (2013)
9. DeLeeuw, K.E., Mayer, R.E.: A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. J. Edu. Psychol. **100**(1), 223 (2008)
10. Dillenbourg, P.: Design for classroom orchestration. Comput. Edu. **69**, 485–492 (2013)
11. Dillenbourg, P., Järvelä, S., Fischer, F.: The evolution of research on computer-supported collaborative learning. In: Balacheff, N., Ludvigsen, S., de Jong, T., Lazonder, A., Barnes, S. (eds.) TEL, pp. 3–19. Springer, The Netherlands (2009)

12. Dillenbourg, P., Zufferey, G., Alavi, H., Jermann, P., Do-Lenh, S., Bonnard, Q., Cuendet, S., Kaplan, F.: Classroom orchestration: the third circle of usability. In: Proceedings of CSCL 2011, vol. 1, pp. 510–517 (2011)
13. Doyle, W.: Ecological approaches to classroom management. In: Evertson, C.M., Weinstein, C.S. (eds.) Handbook of Classroom Management: Research, Practice, and Contemporary Issues , pp. 97–125 (2006)
14. Gómez, F., Nussbaum, M., Weitz, J.F., Lopez, X., Mena, J., Torres, A.: Co-located single display collaborative learning for early childhood education. Int. J. Comput. Support. Collab. Learn. **8**(2), 225–244 (2013)
15. Gould, J.D., Lewis, C.: Designing for usability: key principles and what designers think. Commun. ACM **28**(3), 300–311 (1985)
16. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. **52**, 139–183 (1988)
17. Kharrufa, A., Martinez-Maldonado, R., Kay, J., Olivier, P.: Extending tabletop application design to the classroom. In: Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces, pp. 115–124. ACM (2013)
18. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1994)
19. Onrubia, J., Engel, A.: The role of teacher assistance on the effects of a macro-script in collaborative writing tasks. Int. J. Comput. Support. Collab. Learn. **7**(1), 161–186 (2012)
20. Paas, F., Renkl, A., Sweller, J.: Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. Instr. Sci. **32**(1), 1–8 (2004)
21. Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W.: Cognitive load measurement as a means to advance cognitive load theory. Edu. Psychol. **38**(1), 63–71 (2003)
22. Prieto, L.P., Dlab, M.H., Gutiérrez, I., Abdulwahed, M., Balid, W.: Orchestrating technology enhanced learning: a literature review and a conceptual framework. Int. J. Technol. Enhanced Learn. **3**(6), 583–598 (2011)
23. Prieto, L.P., Sharma, K., Wen, Y., Dillenbourg, P.: The burden of facilitating collaboration: towards estimation of teacher orchestration load using eye-tracking measures. In: Proceedings of the 11th International Conference on Computer-Supported Collaborative Learning, Vol. I, pp. 212–219 (2015)
24. Prieto, L.P., Wen, Y., Caballero, D., Sharma, K., Dillenbourg, P.: Studying teacher cognitive load in multi-tabletop classrooms using mobile eye-tracking. In: Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces, pp. 339–344. ACM (2014)
25. Roschelle, J., Dimitriadis, Y., Hoppe, U.: Classroom orchestration: synthesis. Comput. Edu. **69**, 523–526 (2013)