Orchestration Load Indicators and Patterns: In-the-wild Studies Using Mobile Eye-tracking

Luis P. Prieto, *Member, IEEE,* Kshitij Sharma, Łukasz Kidzinski, and Pierre Dillenbourg

Abstract—Orchestration load is the effort a teacher spends in coordinating multiple activities and learning processes. It has been proposed as a construct to evaluate the usability of learning technologies at the classroom level, in the same way that cognitive load is used as a measure of usability at the individual level. However, so far this notion has remained abstract. In order to ground orchestration load in empirical evidence and study it in a more systematic and detailed manner, we propose a method to quantify it, based on physiological data (concretely, mobile eye-tracking measures), along with human-coded behavioral data. This paper presents the results of applying this method to four exploratory case studies, where four teachers orchestrated technology-enhanced face-to-face lessons with primary, secondary school and university students. The data from these studies provide a first validation of this method in different conditions, and illustrate how it can be used to understand the effect of different classroom factors on orchestration load. From these studies we also extract empirical insights about classroom orchestration using technology.

Index Terms—Orchestration, Orchestration load, Eye-tracking, Cognitive load, Classroom studies.

1 Introduction

Far from initial visions of a 'teacherless' classroom, teacher facilitation has been shown to be a crucial factor for the effectiveness of technology-enhanced learning (TEL), especially in authentic, face-to-face settings [1], [2]. Supporting this facilitation, however, has also been declared one of the foremost challenges in the area of learning technologies, under the label 'orchestrating learning' [3].

Orchestration has been defined as "the process of productively coordinating supportive interventions across multiple learning activities occurring at multiple social levels" [4]. Although different learning technology researchers use the term with a variety of meanings [5], there is a certain consensus that orchestration specifically addresses the challenges of TEL practice under the multiple constraints of *authentic educational settings* [6].

From a learning technology designer perspective, orchestration-related research denotes a focus on "usability at the *classroom* level" [7]. In this sense, researchers sometimes use the term 'orchestration load', in analogy to cognitive load [8], [9]: a useful construct related to *individual* usability, which has been studied thoroughly in cognitive science, educational psychology and human-computer interaction [10], [11].

However, in contrast with cognitive load, which is often studied in controlled laboratory conditions, face-to-face classroom orchestration load is hard to simulate accurately in a lab. This has led most researchers to study it by observing authentic classroom conditions (e.g., a real course with dozens of students and a teacher). This difficulty, alas, has also led to the term being used in a rather high-level and abstract manner [8], [9]. Indeed, the few attempts made at quantifying orchestration load rely on ad-hoc proxy metrics such as classroom workflow efficiency [12].

In our previous work, we have explored the feasibility of mixing physiological, behavioral and subjective measures to study orchestration load in authentic settings, in a more concrete and detailed manner [13]. However, we still lack models of the factors that affect orchestration load, and reliable measures to help us compare different classroom situations in terms of orchestration load. This paper presents such a method for quantifying orchestration load. Our proposed method (see figure 2) combines the observation of teachers' behavior with statistical modelling, and attempts to disentangle the influence that different factors like the concrete teaching activity at hand, or the social plane of teacher-student interactions, have on such load.

As an initial validation of the proposed model and measurement method, we have applied them in several case studies conducted in a variety of technology-enhanced classrooms, including both commonplace technologies like laptops, and more novel ones such as augmented paper tabletops. These case studies comprised a total of 14 sessions led by four teachers with different levels of experience, and students ranging from primary school children to university students. In each of these cases, we used reasonable postulates (e.g.: in the same situation, an experienced teacher will be less loaded than a novice one) as the 'ground truth' for validating the models.

In the following section, we describe the main related work on the measurement of cognitive and orchestration load. Section 3 proposes a model of factors that affect orchestration load, and section 4 describes a method for quantifying orchestration load built upon such model. Section 5 describes the four case studies of application of the proposed methods. The following sections discuss the implications of these studies, and our current and future work on this line of research.

• *All co-authors are with the Computer Human Interaction for Learning and Instruction (CHILI) Lab, École Polytechnique Fédérale de Lausanne, Switzerland. Additionally, Kshitij Sharma is with the Department of Operations, Faculty of Business and Economics, University of Lausanne, Switzerland.*

Manuscript received November XX, XXXX; revised December XX, XXXX.

2 Classroom Orchestration and Cognitive Load Measures

Classroom management by teachers not only is an important determinant of effective learning outcomes [1], [2]; it is also a highly demanding activity. It involves multiple activities performed in a public space with its own charge of history, and high levels of immediacy and unpredictability [14]. The introduction of digital classroom technologies has added another layer to this already complex space, and may help explain the widely-reported reluctance of certain practitioners about adopting novel technologies. The acknowledgement of this increased complexity also has prompted learning technology researchers to study classroom orchestration [4], [5], often in the sense of designing technologies for "classroom usability" [7].

By testing learning technologies in authentic classroom conditions and observing what technologies work in real lessons, researchers are starting to come up with design guidelines for technology that is "orchestrable": empowering teachers to take control of the technology [9], designing technologies for classroom visibility [8], supporting student accountability [15], supporting synchronous switch between group-level and class activities [16], etc. This new kind of usability research, however, is still in its infancy. Indeed, such heuristic approaches to classroom technology design, while certainly useful, can be compared with the initial approaches to usability studies, based on checklists [17] and heuristics [18]. In those early days, the main focus was on "a few users finishing the task" [19] – bearing resemblance to the "teacher heroes" [20] that very often agree to participate in our classroom technology studies.

In order to advance classroom usability research beyond this heuristic stage, we can take the advice from early usability researchers stating that usable computer systems require 1) a focus on users; 2) iterative design and testing; and 3) empirical measurements of usage [21]. The first two are already part of most learning technologies research, e.g., in the wide usage of iterative, user-centred methodologies like design-based research [22]. The third one, however, is made difficult in orchestration-related research, by the simultaneity and immediacy of authentic face-to-face classroom activities. We still lack models and methods for the *empirical* measurement of orchestration processes, beyond teachers' subjective reporting after the fact, or the ad-hoc definition of proxy measures of performance (such as the students' idle time in a problem recitation session [12]).

One potential path towards such empirical measurement of orchestration has been opened by another analogy between classroom orchestration and individual usability research: the notion of 'orchestration load' [8]. Orchestration load has been defined as "the effort necessary for the teacher – and other actors – to conduct learning activities" [9]. This is a direct application of the concept of cognitive load, which is the mental effort needed by a human to perform a certain task [23]. Cognitive load has been extensively used in human-computer interaction [24], and psychologists and usability researchers have devised multiple direct and indirect methods to measure it [25].

Hence, we could adopt the wealth of methods developed in the fields of psychology and human-computer interaction to measure the cognitive load of teachers orchestrating a classroom. Yet, the fact that orchestration (and orchestration load) inherently refer to the constraints and limitations of managing a classroom in everyday, authentic conditions [6] makes such transposition of methods quite challenging.

For instance, cognitive load can be measured directly by 'dual task' experiments: analyzing the performance of individuals that do a secondary task like detecting that a text in the screen changes color, while performing our primary task of interest. This kind of method, however, poses problems when the activity of interest is already a complex multi-tasking one [26], like teaching is. Subjective measures of cognitive load (i.e., asking a person how much effort did a task take) are also possible, but have similar limitations to relatively short, atomic tasks, as they rely on the subject's memory of the event. Indeed, in our initial pilots in this line of research we had to discard repeated subjective measures of mental load throughout a lesson, as they disrupted too much the flow (and teachers' experience) of the lesson. Once we rule out other direct measures of cognitive load like brain imaging (for obvious logistic reasons), this only leaves us with indirect measures of cognitive load such as physiological ones [25], which track involuntary reactions of the body to such increased load, e.g., variations heart rate or pupillary dilation. Among these, mobile eye-tracking has the advantage of being applicable while performing everyday tasks, while providing several measures that have been related to cognitive load in the literature:

- *Eye movements* (e.g., saccadic speed) had been related to cognitive load in fast choice [27] or driving tasks [28], as well as in information systems usage studies [29]. However, this relationship is in some studies direct and in other studies inverse, and there are also studies that failed to find such differences [30].
- *Fixation patterns* (e.g., fixation duration, or number of long fixations) have also been related to workload in a variety of situations, e.g., [29], [31], [32], [33]. However, the evidence about these measures is not unanimous either: for instance, a recent study failed to find effects of website complexity on fixation durations [34].
- By far, the most common eye-tracking variable related to cognitive load is the use of *pupillary response* (i.e., measures related to pupil diameter). This kind of measures have been related to cognitive load in a variety of tasks and conditions, since the 1960s [35]. However, these are also the most controversial ones, as certain studies found no significant difference in these measures [36], and some authors point out that these measures are prone to contamination [33]. On the other hand, the literature also mentions the possibility of technology limitations of eye-trackers, especially in older studies, which could also be a problem for the saccadic and fixation measures [37]. However, newer primary and replication studies using pupillary response have found overwhelmingly confirmations of this relationship (e.g., [29], [38], [39], [40]), including both visual and non-visual tasks, and studies done in realistic settings using mobile eye-tracking technology [41].

As we can see, no single eye-tracking metric (or any other method, for that matter) has been shown to measure cognitive load reliably for every kind of task [27] and, in fact, most cognitive load studies have only dealt with relatively simple, well-defined tasks. Furthermore, most of these cognitive load measurement methods and metrics have been tested in controlled laboratory conditions, to ensure that extraneous factors do not confound the measurements. Classroom management, on the other hand, is a complex multi-tasking activity, in which perception processes, modelling of students' understanding, real-time decision making and dynamic adaptation of a lesson plan fluidly take place. These difficulties may explain why there is very little research around cognitive load in classroom management and, for the most part, dealing with the concept in a very general manner [42].

3 A First Model of Factors in Run-time Classroom Orchestration

Although researchers in TEL have sometimes used the term 'orchestration' to include also pre-lesson planning or post-lesson reflection [5], in this paper we will look specifically at the orchestration load during the *run-time* of a face-to-face lesson [8]. More concretely, we want to model the *moment-to-moment* orchestration load of a teacher, which is similar to the instantaneous cognitive load used in psychology [43].

From Dillenbourg, Järvelä and Fischer's definition of orchestration [4] (see section 1), we can deduce that, at any moment, there are several *process variables* that may affect the orchestration load:

- The teacher's current coordination *activity*, e.g., explaining a concept to the learners, versus for instance monitoring them while they work.
- The *social plane* at which this coordination is happening: e.g., monitoring the work of a single student probably does not imply the same orchestration load as doing it for a class of 50 learners.
- The classroom resource (including as resources both objects, technology and people) on which the teacher is currently *focusing*: e.g., writing on the blackboard may not represent the same load as fiddling with complex geometry software.

Aside from these process variables, and taking into account Feldon's work on the role of automaticity in the cognitive load of teaching [42], we can also deduce that the orchestration load can be further influenced by factors such as teacher *expertise*, and the teacher's *familiarity* with the current kind of classroom situation (e.g., if a new technology for which the teacher has not developed automatisms yet, is being introduced). Additionally, we can also think of other external factors that may influence orchestration load, such as the presence of *another teacher helping* with the orchestration, as long as this help is efficient.

Therefore, to have an estimate of the load experienced by a teacher and enable meaningful comparison between two different classroom situations, we need to take into account the influence of all these variables (see figure 1). For instance, did one lesson require more lecturing than the other, which portrayed more monitoring time? Even if

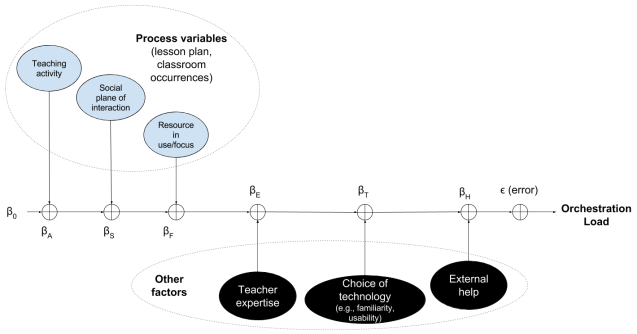

Fig. 1. A model of factors influencing orchestration load

the activity mix was similar, did the teacher spend more time fiddling with the computer in one case than in the other? Was the teacher equally familiar with both kinds of classroom situations? etcetera. It is interesting to note that, as educational technology designers, we may not be directly interested in the contribution of all these variables (since they often derive from the lesson plan or the teacher's peculiarities). We need, however, to understand the orchestration load that our novel technologies may indirectly cause by changing the mix of these factors.

In the absence of prior knowledge about the relationship between these variables and orchestration load, we propose to model the moment-to-moment orchestration load of a teacher as a *linear* combination of the variables, shown in formula (1). This choice of a linear model has the advantage of being relatively simple, and easier to interpret and generalize – both important traits given our goal of not only measuring orchestration load, but also deriving insights about the influence of these factors on classroom usability.

$$OL = \beta_0 + \beta_A \times A + \beta_S \times S + \beta_F \times F + \beta_E \times E + \beta_T \times T + \beta_H \times H + \epsilon \quad (1)$$

In this formula, A represents the current coordination activity, S is the current social plane of interaction, F is the current classroom resource on which the teacher is focusing, E represents the teacher expertise, T the familiarity of the teacher with the current technology or situation, and H is the presence of an additional helper in the orchestration. These variables themselves can be modelled in different ways, e.g., as a categorical set, as fuzzy variables, etc.

One interesting property of the model presented above is that the process variables (teaching activity, social plane of interaction, focus of the interaction) can be determined at any point in time by an external observer with relative ease. On the other hand, the last line of factors like expertise, familiarity and external help, can be assumed to be constant during a classroom lesson, and can also be easily determined, or can even be manipulated by researchers (something we do in the validation studies of section 5). Hence, the only piece of information missing is a measure (or an estimation) of the total instantaneous orchestration load a teacher is facing at a certain moment of a lesson (OL, in formula (1)).

4 Estimating Orchestration Load Using Physiological (Eye-tracking) and Behavioral Measures

As mentioned above, we need a way of tracking the evolution of instantaneous orchestration load. This estimation of the total instantaneous load can be used, along with the other observable variables of our model (like the current teaching activity or social plane of interaction), to build statistical models that estimate the influence of each factor on the orchestration load. Once we have an idea of these influences, we can make meaningful comparisons between situations, determine whether one them entailed distinctly more load than the other, and start disentangling why.

As we saw in section 2, there are multiple eye-tracking measures that have been related to cognitive load in a variety of visual and non-visual tasks, including eye saccade movements, fixation patterns and pupil size measures. However, there also exist studies that failed to find these relationships, in certain tasks and situations. This abundant but still conflicting evidence, along with the multi-task and uncontrolled nature of studying classroom orchestration 'in the wild', led us to think that no single eye-tracking (or physiological) measure would be sufficient to gather differences in orchestration load reliably, and that only by looking at *multiple* measures, and accumulating evidence over *longer periods of time*, we would have more chances at such discrimination. Weighing in the more recent advances in measurement technologies, among all the measures and studies mentioned above, we chose a set of measures that had worked well together as measurements of workload in more recent research, leading us to use those in [29]: mean pupil size, pupil size variation, saccade speed and number of long fixations ($> 500ms$).

These four eye-tracking metrics are combined into a single estimate for the instantaneous orchestration load at a certain point in time, by taking the first component of a principal components analysis (PCA) of the four eye-tracking metrics signals. This component was chosen as it is the linear combination of these methods that captures the most variance in the metrics' data, and is hence the most likely combination of variables to capture the different loads of contrasting classroom situations.

The fact that mobile eye-trackers provide by default a first-person audiovisual recording of the lesson from the point of view of the teacher also enables researchers to observe the process variables of our model. These variables can be extracted, for instance, through post-hoc manual coding of these videos by a researcher, indicating what activity the teacher is conducting, at what social plane, or what is the main focus of the gaze at any point in time. This manual coding of the first-person videos also may help to counter the potential contamination of the eye-tracking metrics taken in an uncontrolled classroom context [33], by *incorporating* these confounding factors into our models.

To summarize, we propose to use physiological data (from mobile eye-tracking technology), along with behavioral data (the manual coding of teachers' actions and social context from the video feed), to build models of the evolution of orchestration load. Our data analysis process thus follows five main steps, depicted in figure 2:

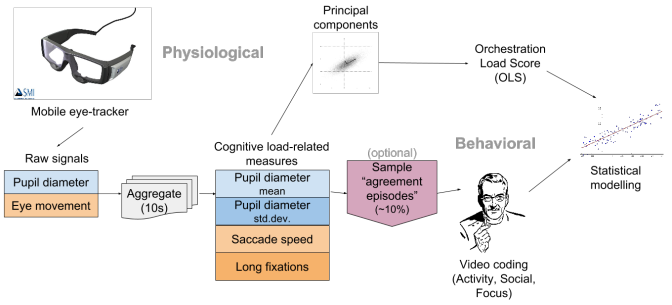

Fig. 2. Diagram representing our proposed method for the physiological-behavioral estimation of orchestration load

1) During the session, eye-tracking data and audiovisual feed of the lesson are recorded using a mobile eye-tracker worn by the teacher (see, e.g., figure 7).
2) From the raw eye-tracking data, we take multiple measures that have been related to cognitive load (pupil diameter mean, pupil diameter standard deviation, average saccade speed and number of fixations >500ms). We divide the lesson into windows or *episodes* of a certain length (e.g., sliding windows of 10 seconds, with 5-second overlaps), aggregating those eye-tracking measures for each episode.
3) After a principal component analysis (PCA) is run over our whole dataset of load-related eye-tracking metrics, the first component of the PCA is calculated for each episode, as our best estimation of the total instantaneous orchestration load (from now on, orchestration load score, or OLS).
4) By looking at the first-person audiovisual recording of the lesson, each episode is manually coded in terms of the process variables of the model (current teaching activity, social plane of interaction and main focus of the teacher's gaze). Given that there can be several hundred such episodes in a 40-minute lesson, a purposeful sampling of the episodes can be done, coding only those in which all four eye-tracking metrics are either above or below the median of the session (what we could call 'agreement episodes'). This sampling tries to capture critical moments in terms of orchestration load [44], and represents the episodes in which our four cognitive load-related metrics agree the load should be especially high, or especially low.
5) Using the orchestration load estimation derived from eye-tracking metrics (OLS, see step 3) and the video-codes extracted from the first-person video (step 4), we build linear statistical models of the orchestration load of the lesson.

This last step, the statistical modeling of orchestration load, deserves further discussion. Depending on the aim of our studies, this modelling can take two forms:

- To *validate* the OLS as a reliable measure of load, a logistic regression model can be built to predict a known binary 'ground truth' (the comparison of data from two classroom situations in which we know the orchestration load is noticeably different) as a function of the process variables and the OLS.

Checking the significance of the OLS as a predictor of such ground truth would provide evidence of its goodness as an estimator of orchestration load. This use is demonstrated in sections 5.2, 5.3 and 5.4.

- Assuming that the OLS is a good measure of instantaneous orchestration load, a linear regression model of the OLS as a function of the observed process variables can be built to derive *insights* about the influence of such variables on orchestration load, for a certain classroom situation. This kind of use is illustrated in section 5.5 and in section 6.

As noted in section 3, our choice of (generalized) linear models is mainly motivated by their advantages in terms of interpretability and generalizability. More concretely, the linear/logistic regression models used in this paper enabled us to express orchestration load with an easily understandable formula, to test the significance of predictors, and measure effect sizes and relative influence of the model's components. The process variables were represented as categorical variables, to match the categories the research team used when coding the first-person audiovisual feed. However, other kinds of modelling (e.g., non-linear) and representation of variables (e.g., fuzzy variables) could also be proposed, under the same basic methods proposed here.

5 Case Studies

5.1 Methodology

In order to provide evidence of the validity of the models and measures of orchestration load outlined above, we have applied them in several case studies performed in authentic classroom conditions. Table 1 shows how a total of 14 sessions were orchestrated by four different teachers, with students at different educational levels (from 11-year-old learners to university students). The research questions that these case studies pursued can be summarized as:

RQ1 Are the aforementioned orchestration load score (OLS) and the proposed linear statistical models able to discriminate the load of substantially different classroom situations?

RQ2 Can the application of these measures and models provide us with insights about the orchestration load of different classroom situations?

Among these four case studies, studies 1, 2 and 3 aimed at *validating* the aforementioned model and physiological-behavioral measures (RQ1). In each of these validation studies we compared two orchestration situations which were similar in most aspects, but varied significantly in one of the additional factors mentioned in section 3 (namely, teacher expertise, familiarity with the classroom technology, and having additional help from a fellow orchestrator). In these studies, we make the assumption that one of the situations imposed a substantially higher orchestration load on the teacher than the other (e.g., having a helper would entail *less* orchestration load than not having it). While it is debatable that the orchestration load was significantly higher due to these factors at all moments, it is reasonable to suppose so for the *aggregation* of the whole session. Furthermore, this kind of 'weak experimental control' is as much as we dared to manipulate otherwise realistic classroom situations, so as to maintain the authentic conditions that are the hallmark of orchestration-related research [6].

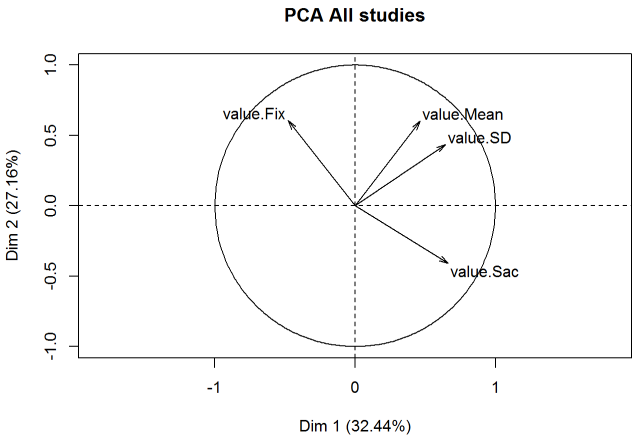

Fig. 3. Projection of the four eye-tracking measures recorded throughout the four case studies, on the space spanned by the first two PCA components. The Orchestration Load Score (OLS) is the linear combination of the four measures, with loadings represented by the projection of the four measures on the *horizontal* axis

On the other hand, study 4 was *not* manipulated, and is rather provided as an illustrative example of how the models and methods can be used to extract insights about the orchestration load of a certain classroom situation (RQ2), in a unmanipulated, non-comparative manner.

Initial explorations of the datasets from Studies 1 and 4 have been published before [45], as is the case for part of the eye-tracking data in Study 2 [13]. However, the application of the statistical models and methods depicted in sections 3 and 4, the main focus of this paper, is completely novel. Below, we describe the context and main results of the case studies. In all four studies, we followed the data gathering and analysis method described in section 4. Moreover, the anonymised datasets generated, full analytical code and detailed results are openly available online[1].

5.2 Study 1: Exploring Teacher Expertise in an University Course

One of the main factors that may affect the orchestration load of a classroom situation is the amount of teacher experience, as more expert teachers have internalized and automatized many of the small tasks involved in classroom management [42]. Hence, more expert teachers will manage a classroom with less orchestration load than novice ones. In this first study, we explore the following particularization of RQ1 above: *Is our modelling and OL score able to discriminate between similar learning situations, orchestrated by a novice and an experienced teacher?*

5.2.1 Context

The study took place during an authentic master-level course on 'Digital education and learning analytics', held at EPFL (Switzerland). As outlined in table 1, the lessons

1. https://github.com/chili-epfl/paper-IEEETLT-orchestrationload.

TABLE 1
Summary of main case study characteristics

Case study	Setting	Teachers	Sessions (length)	Number of students	Technological support	Main goal	Target variable
1	University	A (expert), B (novice)	2+1 (45-65′ each)	1 cohort, 10–12 each	Laptops, classroom projector	Validate OLS	Teacher expertise (novice vs. expert)
2	Primary school	C (expert)	2+2 (80′ each)	2 cohorts, 18–22 each	Laptops, classroom projector vs. Tabletops, classroom projector	Validate OLS	Familiarity with technology (usual vs. novel)
3	Open doors (primary)	D (novice, researcher)	4 (35-45′ each)	4 cohorts, 19–21 each	Tabletops, classroom projector	Validate OLS	External (human) help (without/with helper)
4	Open doors (primary)	D (novice, researcher)	3 (35-45′ each)	3 cohorts, 14–25 each	Tabletops	Illustrate insights	–

Fig. 4. Classroom setup during Case study 1 (master-level course)

took place face-to-face, with 10-12 students (depending on the session) using their own laptops, and the teacher using a laptop, projector and whiteboard. Figure 4 illustrates the physical setting of the classroom. The lesson plans combined fluidly lecturing, questioning of students and exercises. Two sessions were recorded with a teacher who had more than 10 years of teaching experience, and one session for a novice teacher[2] with only one year of teaching experience.

5.2.2 Results

To validate whether the model of factors and the proposed physiological measures (i.e., the orchestration load score – OLS) represent a good predictor for orchestration load in this context, we assumed as a 'ground truth' that the lessons orchestrated by the novice teacher represented a higher load (Load=1) than the ones orchestrated by the expert (Load=0). We trained a logistic regression model trying to predict this binary outcome (i.e., distinguish an episode orchestrated by the experienced teacher, vs. the novice one), using the three process variables coded by a human (teaching activity, social level of interaction, main focus of the gaze) and the OLS.

As we can see in table 2, once we take away the influence of the process variables, the OLS still remains a significant predictor of the assumed (binary) orchestration load ($p < 0.001$). Furthermore, this model including the OLS explains a considerable amount of the variance in the data (McFadden's pseudo-$R^2 = 0.88$). An analysis of variance of the different variables in the model confirms this, with the OLS explaining most of the deviance in the data ($p < 0.001$).

2. Actually, two sessions were recorded for the novice teacher, but technical problems during the recording of one of them forced us to discard the data, as the situation would have been no longer similar to the ones recorded with the expert teacher.

TABLE 2
Logistic regression model of assumed orchestration load (1–Novice, 0–Expert teacher) in case study 1

Coefficient	Estimate	Std. error	z	p-value
Intercept (Monitoring, Class-level, Focus on student faces)	-2.65	1.51	-1.75	0.08
Activity: Explanation	-0.28	1.46	-0.19	0.85
Activity: Questioning	-1.93	1.90	-1.02	0.31
Activity: Repairs	2.52	3.42	0.73	0.46
Activity: Task distribution	-16.49	2765	-0.01	1.00
Social: Individual	5.76	3.23	1.78	0.07
Focus: Projector	-3.49	2.90	-1.21	0.23
Focus: Student computer	1.00	1.72	0.58	0.56
Focus: Table	1.77	5.96	0.30	0.77
Focus: Teacher computer	-2.04	1.40	-1.45	0.15
Focus: Whiteboard	-2.91	2.78	-1.05	0.29
OLS	**8.42**	**2.30**	**3.67**	**<0.001*** **

It is important to note that the two teachers in our study may differ in many other aspects affecting orchestration load, aside from the amount of teaching experience. Hence, this study provides only preliminary evidence that the proposed modelling (and the OLS as an approximation to moment-to-moment orchestration load) can distinguish high- vs. low-load situations in this kind of classroom context. In the following study, we test them in a context in which such personal differences are controlled.

5.3 Study 2: Exploring Familiarity with Technology in a Primary School Classroom

As mentioned in section 3, another factor that is bound to affect the orchestration load experienced by a teacher in a lesson using technology, is the teacher's familiarity with the technology in use: a person trying out a tool for the first few times will not have yet developed automatisms to effortlessly use it in a classroom context, and hence will experience a higher orchestration load than during the usage of her usual set of tools. Hence, in this study we try to answer the question: *Are our linear model and OLS able to discriminate learning situations orchestrated by a same teacher, using the usual vs. a novel classroom technology?*

5.3.1 Context

In order to answer this question, we conducted another study in the context of face-to-face lessons in a Swiss in-

ternational secondary school (see table 1). We recorded two sets of two math lessons, in which an experienced teacher (who had more than 20 years of teaching experience) taught to two different cohorts of students about geometry. Each cohort had 18–22 students, depending on the session, all aged 11–12. In the first set of sessions, the teacher orchestrated students' work using a set of technologies she had used in previous courses (laptops with a geometry software[3], a projector and her usual classroom management software[4]) with each of the two cohorts. In the second set, the teacher orchestrated a collaborative groupwork game, with the same two cohorts, using a novel set of technologies: an augmented paper tabletop interfaces based on Tinkerlamp [46], and a projector used as public display. Both classroom setups are portrayed in figure 5.

5.3.2 Results

As we did in the first study, to validate whether the OLS is a good predictor for orchestration load in this context, we assumed as the ground truth that the lessons orchestrated using novel technologies (for the teacher) represented a higher load (Load=1) than those using the usual classroom technology setup (Load=0). After extracting the OLS from the recorded eye-tracking metrics of every 10-second episode, and manually video-coding the teacher activity, social plane and focus of the gaze of the 'agreement episodes', we trained a logistic regression model to predict the ground truth as a function of these variables.

As we can see in the coefficients of the logistic regression model of table 3, once we take away the influence of the process variables, the OLS still remains a significant predictor of the binary orchestration load ground truth ($p = 0.002$). Furthermore, this model including the OLS explains a considerable amount of the variance in the data (McFadden's pseudo-$R^2 = 0.75$). An analysis of variance of the different variables in the model confirms this, with the OLS explaining an appreciable part of the deviance in the data ($p < 0.001$).

To interpret these results, we can assume that the personal characteristics of the teacher did not vary across conditions, since this was a within-subject study. Nevertheless, as usual in this kind of authentic setting studies, other factors aside from familiarity with the technology may have played a role in the respective orchestration load of both sets of sessions. Most important among these is the fact that the teaching and learning activities performed were not the same in both sets of sessions: the novel technology ones implied a less habitual way of managing the classroom (alternance between small-group collaboration and class-wide competition), as compared with the usual-technology ones (which were based on individual/pair exercises following a worksheet). However, this would only exaggerate the differences in orchestration load between usual and novel sessions, thus making the proposed ground truth an ever more reasonable assumption. Therefore, this second study provides additional evidence that the proposed model and measures of orchestration load might be applicable to a variety of authentic classroom settings.

3. Geometer's Sketchpad, http://www.dynamicgeometry.com/.
4. NetSupport School, http://www.netsupportschool.com/

TABLE 3
Logistic regression model of assumed orchestration load (1–New classroom technology, 0–Usual technology) in case study 2

Coefficient	Estimate	Std. error	z	p-value
Intercept (Monitoring, Class-level, Focus on student faces)	1.71	0.54	3.17	0.002**
Activity: Explanation	-2.91	0.71	-4.08	< 0.001***
Activity: Questioning	-4.46	0.91	-4.88	< 0.001***
Activity: Repairs	-2.55	0.91	-2.81	0.005
Activity: Task distribution	0.31	0.87	0.36	0.72
Social: Small group	3.93	1.11	3.53	< 0.001***
Social: Individual	-0.25	0.91	-0.28	0.78
Focus: Student laptop	-20.45	1577	-0.01	0.99
Focus: Paper elements	-2.31	0.74	-3.14	0.002**
Focus: Projector	20.92	2921	0.01	0.99
Focus: Tabletop computer	17.39	1758	0.01	0.99
Focus: Teacher computer	-2.68	0.92	-2.93	0.003**
OLS	**0.76**	**0.25**	**3.04**	**0.002****

5.4 Study 3: Exploring External Help in an Open Doors Day

Following the non-exhaustive list of factors affecting orchestration load mentioned in section 3, we can also conceive a third way of manipulating the orchestration load of a classroom situation to validate our proposed methods. By providing the main orchestrator with an assistant, we postulate that her orchestration load during the lesson will be lowered. Hence, we performed a third study to validate the model and measures of orchestration load, looking at the following variant of our RQ1 question: *Are our linear model and OLS measures able to discriminate between similar learning situations, orchestrated by a same teacher, but with/without an assistant orchestrator?*

5.4.1 Context

To answer this question, we set up a semi-authentic study in the context of an open-doors day in our lab, in which cohorts of local school children aged 10–12, from nearby public schools, visit university labs. We transformed one of our rooms into a multi-tabletop classroom (figure 6), and one member of the research team acted as the main facilitator of four math lessons, with different cohorts of visiting students (19–21 students each, see table 1). The four sessions made use of augmented paper tabletops based on Tinkerlamp [46] and a public display/projector. The lesson plan of the sessions was identical, alternating mini-lectures, small-group and whole-class exercises as well as games about geometry and coordinate systems. To manipulate the orchestration load, in two of the sessions the main orchestrator was aided by an assistant. However, such aid was only provided in those moments where multiple groups were working independently, to ensure that the help was meaningful in terms of orchestration load – something hard to achieve in moments of whole-class lecturing, for example. The groups of students that each orchestrator would monitor/help were decided beforehand, to avoid further coordination load during the session.

Fig. 5. Classroom setup during two of the sessions in Case study 2: using usual laptop (left) and novel tabletop technology (right)

Fig. 6. Classroom setup during Case study 3 (open-doors day)

5.4.2 Results

To validate whether the OLS is a good predictor for orchestration load in this study, and taking into account our linear model of orchestration load, we assumed as a ground truth that the lessons orchestrated without an assistant represented a higher load (Load=1) than the ones orchestrated with the assistant (Load=0). This is assumed to be true at least for the parts of the lesson where the assistant had a meaningful role. Hence, after following the same steps as in our previous studies (calculating OLS from eye-tracking measures, and manual coding of agreement episodes), we selected *only* the episodes falling in parts of the lesson plan where students were working independently (i.e., in which the assistant orchestrator, if present, had a meaningful role). Using these selected data, we trained a similar logistic regression model to predict the aforementioned ground truth using the known process variables (activity, social plane, focus – from the video coding) and the OLS.

In the data of this study, once we take away the influence of the process variables, the OLS showed a positive correlation (i.e., coefficient) with the ground truth. However, when building a logistic regression model using the 'agreement episodes' as in previous studies, the OLS was no longer a significant predictor of the binary orchestration load we assumed as ground truth ($p = 0.13$). We conjectured that this lack of significance was tied to the low volume of data points used to build the model ($n = 163$ relevant episodes), compared to that of our previous validation studies.

In order to test this conjecture, we videocoded the full extent of the sessions in this study. As we can see in table 4, the linear regression model built with additional videocoded data (a total of $n = 610$ episodes) finds the OLS as a significant predictor of the assumed ground truth ($p = 0.016$). In both the initial and this second modelling, however, the logistic regression models explain relatively little variance in the data (McFadden's pseudo-$R^2 = 0.07$, in the model with more samples).

The above modelling results provide additional supporting evidence for the OLS as a discriminant measure of orchestration load, but also highlight some of the limitations of the proposed method in terms of the amount of data needed to build the statistical models of orchestration load. Especially in cases like this one, where we look at data representing only *a part* of a session rather than stretches of multiple sessions, a coding of the full extent of the session might be in order, so as to gather enough information to build the models.

5.5 Study 4: Modelling Orchestration Load in an Unmanipulated Setting

Once we have initial validation that the OLS is a good measure to discriminate situations entailing different orchestration loads, we can use it as a proxy for moment-to-moment orchestration load. Building linear models of OLS (like the ones presented in section 3) will help us understand how the different process variables affect orchestration load in a certain classroom situation, and will enable us to extract patterns of higher or lower load, and compare different kinds of classroom situations. Below, we illustrate the usage of this measure in an unmanipulated setting, to answer a version of RQ2: *What insights can we extract about the orchestration load in a multi-tabletop classroom, using a collaborative augmented paper game?*

TABLE 4
Logistic regression model to predict assumed orchestration load (1–Without assistant, 0–With assistant) in case study 3

Coefficient	Estimate	Std. error	z	p-value
Intercept (Monitoring, Class-level, Focus on student faces)	-0.64	0.35	-1.84	0.066
Activity: Questioning	1.14	0.36	3.15	0.002**
Activity: Repairs	0.87	0.22	4.04	< 0.001***
Activity: Task distribution	-0.04	0.31	-0.13	0.896
Social: Small group	0.16	0.26	0.61	0.543
Focus: Student backs	0.43	0.47	0.91	0.362
Focus: Projector	-0.05	0.38	-0.13	0.893
Focus: Tabletop computer	0.41	0.36	1.14	0.256
Focus: Teacher's computer	0.86	0.77	1.11	0.265
Focus: Other teacher	-16.13	637.80	-0.03	0.980
Focus: Teacher papers	0.84	0.64	1.32	0.187
OLS	**0.38**	**0.16**	**2.41**	**0.016***

TABLE 5
Linear regression model to predict the orchestration load score (OLS) in case study 4

Coefficient	Estimate	Std. error	t	p-value
Intercept (Monitoring, Class-level, Focus on student faces, Session 1)	-0.02	0.11	-0.22	0.82
Activity: Explanation	0.01	0.07	0.14	0.89
Activity: Repairs	-0.001	0.05	-0.02	0.98
Activity: Task distribution	-0.05	0.07	-0.80	0.43
Social: Small group	**-0.29**	0.12	-2.44	**0.02***
Focus: Student backs	**-0.38**	0.13	-2.90	**0.004****
Focus: Teacher desk	0.06	0.14	0.43	0.67
Focus: Tabletop computer	**-0.90**	0.11	-8.43	**<0.001*****
Session: 2	**0.30**	0.05	5.79	**<0.001*****
Session: 3	**0.29**	0.05	5.53	**<0.001*****

Fig. 7. Classroom setup during Case study 4 (unmanipulated open-doors day)

5.5.1 Context

Similarly to the previous study, this one took place in a semi-authentic classroom situation in the context of another open-doors day in our lab, with three cohorts of students aged 10–12, from local schools. As outlined in table 1, each cohort had between 14–25 students, depending on the session. Again, a member of the researcher team (a novice teacher) played the role of main orchestrator, with the usual teachers acting as passive observers (figure 7). The three recorded sessions had an identical lesson plan, in which the orchestrator combined whole-class mini-lectures with small-group collaborative exercises around augmented paper tabletops. In this kind of classroom context, we did not manipulate any variable of the orchestration load model.

5.5.2 Results

To get the main orchestration load trends, after calculating the OLS and video-coding the agreement episodes, we trained a linear regression model to try to predict the OLS using the known process variables observed from the subjective video (teaching activity, social plane, focus of the gaze), as well as the session identifier – i.e., we also wanted to know if one session had been noticeably more load than the others. The resulting model explains approximately 50% of the variance in OLS, and an analysis of the variance of the model shows that the social plane of interaction, the focus of the gaze and the session were all significant predictors ($F = 157$ with $p < 0.001$; $F = 46$ with $p < 0.001$; and $F = 22$ with $p < 0.001$, respectively), while the teaching activity was only marginally significant ($F = 2.59$ with $p = 0.05$).

As we can see in table 5, the main orchestration load trends, i.e., those coefficients that are significant predictors, are: a) that small-group interactions tended to be lower load than the reference level (class-level interactions); b) that looking at students' backs was lower load than looking at their faces (the reference level); or c) that looking at the tabletop itself tended to be lower load. Finally, we also see that d) the second and third sessions were on average higher load than the first session.

We can also speculate about the reasons behind these trends, not directly from the model, but from our contextual knowledge of the classroom situation and our observation of the videos. Class-level interactions may represent higher load than small-group ones due to the teacher needing to track and model the understanding of more students simultaneously (aside from keeping in mind the content of the lesson); the fact that student backs have no meaningful information to read about the progress of the students (as opposed to their faces, which may have) can be linked to the second trend. The other two trends initially seem counter-intuitive: however, the fact that the orchestrator was one of the designers of the tabletop application, and hence knew really well the user interface, might have led to lower loads when looking at the tabletop; the expectation that session 1 would be more load than the others due to the novelty of the situation, might have been offset by the fact that the three sessions occurred almost back-to-back and hence the fatigue might have had a greater role in the load than the relative novelty of the first session. However, these post-hoc rationalizations of the evidence should be taken rather cautiously, as usual in many studies on cognitive load [47].

6 Discussion

6.1 Eye-tracking Measures to Estimate Orchestration Load

Regarding our first research question (are we able to discriminate the load of substantially different classroom situations?), the results from studies 1, 2 and 3 support the notion that the Orchestration Load Score (OLS), obtained from multiple eye-tracking metrics as described in section 4, provides a discriminant measure of teachers' orchestration load, obtainable in authentic face-to-face classroom settings. These results, especially those of case 3, also show some of the limitations of the proposed method: for instance, the fact that a certain amount of data points is needed to discriminate the orchestration load of different classroom situations – i.e., taking only small parts of a session might not lead to clearly distinguishable OLS scores.

Our studies were set in a variety of classroom settings using different classroom technologies, which suggests that it may be applicable to a wide range of such situations. However, the evidence presented here is not definitive, and further comparative studies in authentic settings, with 'reasonable ground truths' like those described in sections 5.2–5.4, should be performed to continue the validation of this method.

In the same way, we can by no means claim that this OLS is generalizable to *any* classroom situation. The presented studies depict face-to-face classroom settings, with an emphasis on either teacher-led or collaborative groupwork activities. Hence, we cannot be certain of how much they can be applicable to other pedagogies like, e.g., project-based learning. Furthermore, our proposed models and measures in sections 3 and 4 are tailored for the case of face-to-face, real-time classroom orchestration. Other kinds of learning (e.g., distance learning) would require a radical re-thinking of the orchestration load factors and how to measure them: synchronous distance activities would translate more easily (as the number of factors involved in the orchestration may indeed decrease in that case), while asynchronous or blended learning activities may be hard to study using our proposed method, given their ubiquitous nature.

Another limitation of the method presented here is the fact that it is quite expensive, both in terms of equipment and human effort, since it requires manual researcher work to code the process variables from the eye-tracking videos. This behavioral video coding could be made more affordable through crowdsourcing, since the process variable categories chosen are relatively simple to assess, even for non-researchers. Alternatively, such video codes could be elicited automatically, as illustrated by our parallel work on applying machine learning techniques on eye-tracking and other wearable sensor data, to extract moment-to-moment teaching activities [48].

We should also note that our deliberate decision of proposing orchestration load as a mono-dimensional construct to be described as a linear combination of factors, is only an initial proposal and may be seen as an over-simplification of the complex reality of the classroom. While this simplicity has helped us propose and operationalize the validation studies presented here, it does not mean that such model is the closest one to reality. More complex, non-linear statistical models could represent more realistically the relationships between the factors affecting orchestration load. Another way in which this construct could be expanded is by making it multi-dimensional: while our application of PCA to the eye-tracking data would lend itself easily to this approach (by just using additional PCA components), the interpretation and validation of these additional orchestration load dimensions would require a more complex set of classroom studies.

The choice of eye-tracking metrics used to calculate the OLS scores in section 4 could also be contested, given the conflicting evidence from the literature. However, our proposed method could be easily expanded with additional eye-tracking metrics, or even with other physiological measures of load (e.g., heart rate or skin conductance). Indeed, even the combination of these metrics into a single score via PCA is a choice for which alternatives exist, from very simple averaging methods, to factor analyses, among others.

In a similar manner, we should note that the list of process variables in our orchestration load model (teaching activity, social plane of interaction, focus of the gaze/interaction) is not an exhaustive one. One could also look at the effect of each individual student, the effect of disciplinary remarks, or many other variables affecting class management. Similarly, the coding scheme chosen here for each of those variables is not exempt from improvement: for instance, a research effort focused on inquiry-based learning might want to categorize teaching activities in a more pedagogy-specific manner, to understand how different activities in the support of inquiries may affect orchestration load. Indeed, even the modelling of these variables as categorical sets is up for discussion: while we found it useful given our manual post-hoc video coding analyses, other alternatives such as the use of fuzzy variables (e.g., if the variables are output as probabilities by a machine learning process [48]), are also possible when modelling the orchestration load.

Other alternative models of orchestration load are also possible: one could argue that, given the complex nature of teaching, "the load of the teacher is always 100%", and what actually varies is the capacity and stress levels (and hence, the performance) of the orchestrator: some teachers may handle an individual model of the understanding of each student, while less capable ones may bunch all students' understanding into a single variable. Further research is needed to understand which of these ways of modelling orchestration load reflects more accurately a wide variety of classroom situations.

6.2 Orchestration Load Patterns and Learning Technology Insights

Our second research question inquired about the insights that could be derived from these kinds of studies about the orchestration load of different classroom situations. Study 4 above illustrates how the OLS and the manually-coded process variables can be used to build statistical models of orchestration load, and the use of such models to extract initial insights about the orchestration load trends and patterns of a classroom situation. By applying the same kind of methods to the data of all the studies presented so far,

we can get insights about the kind of influence that these different variables have on orchestration load, at different levels of specificity. Table 6 summarizes the main orchestration load trends of each study, and of different aggregations within our dataset. This table shows only the effect sizes in terms of Cohen's d (the difference in the means standardized by the variability of the data) of those predictors that are significant at the 0.05 level. For a more complete analysis, please see the paper's accompanying analytical code and open datasets[5].

At the single-study level, we can see for instance how in study 1 the episodes in the reference level (teacher monitoring at the whole-class level, looking at student faces) is the only significant predictor, although we can notice how the average OLS for these episodes was higher for the novice than for the expert teacher (which responds to our expectations). In case study 2, the teacher had significantly higher loads while explaining or questioning to students *only* when using the novel tabletop technology, even if looking at the tabletops themselves was appreciably less load than the reference level.

Joining the data from multiple case studies (second and third columns counting from the right in table 6) we can, for instance, make comparisons between the orchestration load patterns of laptop classrooms versus multi-tabletop classrooms – with the obvious caveats given the very small sample sizes, as we only have data from 2-3 different teachers in each setting. We can see how, in the tabletop classrooms both the reference level (monitoring at the class-level, looking at students), questioning episodes and tabletop-related technical problem-solving tended to be high load (a trend not so apparent in the more traditional laptop classrooms). In the laptop classrooms, we can see how task distribution episodes, and the focus on the whiteboard, seem to be lower load. We can also see several common orchestration load trends across both kinds of classroom technology setups.

Finally, joining the data from all the studies presented here, in the rightmost column of table 6, we can start extracting some general trends about how the different process variables seem to affect orchestration load. With the aforementioned caveats regarding sample size and generalizability, most of the variables seem to be significant predictors of orchestration load: the reference-level episodes (monitoring, class-level, looking at student faces) tend to be higher load, while the teacher focusing on tabletops, her own laptop, or the whiteboard, tend to be lower load. Regarding the social plane of interaction, both small-group and individual interactions seem to be noticeable lower load than class-level ones.

From the learning technology designer's point of view, these trends raise a number of interesting issues:

- The overall trend regarding the relative load influences of different social plane interactions (with class-wide ones involving significantly more load than individual and small-group ones) seems to support the claims that "usability at the classroom level" should be more thoroughly investigated [7], and that increased support is needed for teachers at this level of interaction. This seems to be especially true for monitoring activities, hence supporting several existing classroom technology design guidelines on the need to provide classroom-level workflow and awareness support [8], [15], [16]. In a sense, it also supports the recent rise of teaching and learning analytics as a field of study that tries to increase awareness of learning processes and outcomes, e.g., for orchestration purposes.
- Although this aspect is not thoroughly investigated in our studies, there is initial evidence that teachers with different levels of expertise may experience radically different load from factors such as class-level monitoring (see the trends of study 1 in table 6). Although this need of differentiated support by novice/expert users is consistent with many other areas of human-computer interaction [49], to the best of our knowledge it is a novel issue in classroom technology design, and should be further investigated in the future.
- Looking at the influence of technology on orchestration load, especially when introducing a new technology like our tabletops in study 2, we do not find indications such new technology as a direct source of higher load. Such direct relationship seems to be the underlying assumption of previous advice on designing "orchestrable technology" [9]). Our models, however, suggest that the relationship between orchestration load and technology might be rather an *indirect* one: larger loads are not necessarily tied to *looking* at the new technology's user interface, but rather to new teaching activities and social planes that are emphasized in the new mix of orchestration factors. This may imply that, at the classroom level, different classroom technologies vary the orchestration load by changing *what teachers do*, and how difficult this new mix of activities/planes is. For instance, the introduction of a new technology may increase the uncertainty of the learning situation, provoking that more effort is put into trying to read student faces, rather than looking more intently at the technology itself.
- This indirect relationship between orchestration load and classroom technology, in turn, supports the idea of studying orchestration in realistic classroom conditions, rather than lab settings where focus is often forced on direct technology usage and the measuring of such use. This is in line with Roschelle et al.'s synthesis of orchestration-related research [6].

In any case, we should not forget that these trends and implications are extracted from a limited dataset of classroom studies, and therefore are not guaranteed to be generalizable to every teacher and classroom setting. As more studies of this kind are performed, more generalizable and trustworthy insights will undoubtedly emerge. Furthermore, these trends and insights are tied to the modelling decisions made in terms of defining orchestration load as a mono-dimensional construct and as a linear combination of the observable variables. Hence, if alternative or more complex models of orchestration load were to be found more accurate in the future, more nuanced insights might

5. https://github.com/chili-epfl/paper-IEEETLT-orchestrationload.

also be derived from them (as long as these alternative models are easily interpretable).

7 Conclusion

In this paper we presented an overview of what orchestration load is, and a method to quantify it by combining physiological and behavioral measures (i.e., eye-tracking metrics and manual coding of first-person orchestration videos). The results of applying the proposed method to four studies encompassing 14 sessions led by four different teachers in a variety of settings, support the notion of an Orchestration Load Score (a linear combination of four eye-tracking metrics related to cognitive load) as a useful measure of orchestration load. Our studies show how this measure, along with orchestration process variables like teaching activity or social plane of interaction, can distinguish between situations in which it is reasonable to presume different levels of cognitive load (teachers with different levels of experience, novel versus usual classroom technologies, and having a fellow orchestration assistant). By building statistical models of orchestration load (which assume OLS as a reliable measure of orchestration load), we also obtained first glimpses into the respective influence of different factors that affect orchestration load, in different kinds of classroom situations.

These promising results, however, are not without limitations, mostly related to the small sample sizes in terms of number of teachers and classroom situations tested so far. Hence, future work along this line of research should be directed towards further validation studies of the proposed method (or variants of it) at a larger scale, in different classroom settings.

Within these limitations, our use of a deliberately simple construct definition (orchestration load as unidimensional) and a linear model of intervening factors, has already offered new evidence that partly confirms and partly complements existing guidelines for orchestrable classroom technologies. However, the modelling presented here is only an initial attempt, and could be expanded in future work in a number of ways: an expanded set of eye-tracking and physiological measures related to cognitive load could be triangulated in order to more reliably estimate the variations in instantaneous orchestration load; a larger set of factors affecting orchestration load could be considered when performing such modelling, and these variables themselves could be expressed in ways different from the categorical sets used in this paper; a more accurate modelling of the complex reality of the classroom can be attempted by validating different, non-linear statistical models of orchestration load, as well as multi-dimensional orchestration load constructs (e.g., using additional PCA components extracted from the physiological data); finally, the methods presented here could also be translated to other educational settings beyond the actual focus on co-located face-to-face classrooms. Well aware of the effort that exploring all these avenues of future research would entail, we have opened the anonymous parts of our datasets and the analytical code used in this initial exploration[6], so that the learning technologies community can jointly discuss and explore the benefits and limitations of these methods for the study of orchestration load and classroom technologies.

Once the methods and models proposed here (or their aforementioned expansions) have been validated more thoroughly in a variety of classroom settings, we will be able to add them to the other tools in our usability research toolbox as classroom technology designers. The proposed method could indeed be applied as well to more general educational research (e.g., understanding the orchestration load of different pedagogical methods, in a sort of 'multimodal teaching analytics' [48]). Furthermore, the method presented here could even be applicable beyond the realm of education, in the study of many other human activities that are highly social and involve heavy multi-tasking (e.g., healthcare, knowledge workers, etc.) – what we could call 'multimodal professional analytics'.

6. https://github.com/chili-epfl/paper-IEEETLT-orchestrationload.

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). Special thanks also go to Patrick Jermann, the International School of Lausanne and all the student and teacher participants in EPFL's 'Journée des classes' open-doors events. We also would like to thank the anonymous reviewers and editors that helped us improve the research presented here.

References

[1] F. Gómez, M. Nussbaum, J. F. Weitz, X. Lopez, J. Mena, and A. Torres, "Co-located single display collaborative learning for early childhood education," *International Journal of Computer-Supported Collaborative Learning*, vol. 8, no. 2, pp. 225–244, 2013.

[2] J. Onrubia and A. Engel, "The role of teacher assistance on the effects of a macro-script in collaborative writing tasks," *International Journal of Computer-Supported Collaborative Learning*, vol. 7, no. 1, pp. 161–186, 2012.

[3] R. Sutherland and M. Joubert, "D1.1: The STELLAR vision and strategy statement," STELLAR Consortium, Deliverable, 2009, available online at http://goo.gl/IevQs0 (Last visit: 28 Nov 2016).

[4] P. Dillenbourg, S. Järvelä, and F. Fischer, "The evolution of research on computer-supported collaborative learning," in *Technology-enhanced learning*. Springer, 2009, pp. 3–19.

[5] L. P. Prieto, M. H. Dlab, I. Gutiérrez, M. Abdulwahed, and W. Balid, "Orchestrating technology enhanced learning: a literature review and a conceptual framework," *International Journal of Technology Enhanced Learning*, vol. 3, no. 6, pp. 583–598, 2011.

[6] J. Roschelle, Y. Dimitriadis, and U. Hoppe, "Classroom orchestration: synthesis," *Computers & Education*, vol. 69, pp. 523–526, 2013.

[7] P. Dillenbourg, G. Zufferey, H. Alavi, P. Jermann, S. Do-Lenh, Q. Bonnard, S. Cuendet, and F. Kaplan, "Classroom orchestration: The third circle of usability," *CSCL2011 Proceedings*, vol. 1, pp. 510–517, 2011.

[8] P. Dillenbourg, "Design for classroom orchestration," *Computers & Education*, vol. 69, pp. 485–492, 2013.

[9] S. Cuendet, Q. Bonnard, S. Do-Lenh, and P. Dillenbourg, "Designing augmented reality for the classroom," *Computers & Education*, vol. 68, pp. 557–569, 2013.

[10] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and instruction*, vol. 4, no. 4, pp. 295–312, 1994.

[11] S. Oviatt, "Human-centered design meets cognitive load theory: designing interfaces that help people think," in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 871–880.

TABLE 6
Summary of the statistically-significant process variable factors and their effect sizes (in terms of Cohen's d [50])

Case	1	1	2	2	3	4	1+2	2+3+4	1+2+3+4
Teacher	A	B	C	C	D	D	A+B+C	C+D	A+B+C+D
Expertise	Expert	Novice	Expert	Expert	Novice	Novice	Varied	Varied	Varied
Technology	Laptops + Projector	Laptops + Projector	Laptops + Projector	Tabletops + Projector	Tabletops + Projector	Tabletops only	Laptops + Projector	Tabletops	Varied
Students	Young adults	Young adults	11-12yrs	11-12yrs	10-12yrs	10-12yrs	Varied	10-12yrs	Varied
No.sessions	2	1	2	2	4	3	5	9	14
No.episodes	142	49	236	166	667	332	427	1165	1592
Adjusted R^2 of model	0.11	0.30	0.39	0.29	0.48	0.44	0.16	0.45	0.35
Reference (Monitoring, Class-level, Focus on student faces)	-0.72	+1.15			+0.70			+0.66	+0.45
Activity: Explanation (lecturing)				+1.1					
Activity: Solve technical problems								+0.47	+0.48
Activity: Questioning				+0.97				+0.30	
Activity: Task distribution							-0.5		
Social: Small-group					-0.70			-0.63	-0.51
Social: Individual			-0.77				-0.50	-0.78	-0.69
Focus: Student backs					-0.51	-0.68		-0.68	-0.50
Focus: Table or desk									-0.77
Focus: Pieces of paper			-0.51	-0.98			-0.55	-0.58	-0.63
Focus: Projector				-1.13	-0.49			-0.54	-0.41
Focus: Student laptops									-0.3
Focus: Tabletop computers				-1.09	-0.92	-1.80		-1.07	-0.98
Focus: Teacher's computer			-0.88		-0.40		-0.51	-0.32	-0.61
Focus: Whiteboard							-0.61		-0.65

[12] H. S. Alavi and P. Dillenbourg, "An ambient awareness tool for supporting supervised collaborative problem solving," *Learning Technologies, IEEE Transactions on*, vol. 5, no. 3, pp. 264–274, 2012.

[13] L. P. Prieto, K. Sharma, and P. Dillenbourg, "Studying teacher orchestration load in technology-enhanced classrooms," in *Design for Teaching and Learning in a Networked World*, ser. Lecture Notes in Computer Science, G. Conole, T. Klobuar, C. Rensing, J. Konert, and E. Lavou, Eds. Springer International Publishing, 2015, vol. 9307, pp. 268–281.

[14] W. Doyle, "Ecological approaches to classroom management," *Handbook of classroom management: Research, practice, and contemporary issues*, pp. 97–125, 2006.

[15] A. Kharrufa, R. Martinez-Maldonado, J. Kay, and P. Olivier, "Extending tabletop application design to the classroom," in *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*. ACM, 2013, pp. 115–124.

[16] S. Kreitmayer, Y. Rogers, R. Laney, and S. Peake, "Unipad: orchestrating collaborative activities through shared tablets and an integrated wall display," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 801–810.

[17] S. Ravden and G. Johnson, *Evaluating usability of human-computer interfaces: a practical method*. Halsted Press, 1989.

[18] J. Nielsen, "Finding usability problems through heuristic evaluation," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1992, pp. 373–380.

[19] B. Bailey, "Usability testing: An early history," http://webusability.com/usability-testing-a-early-history/, (Last visit: 28 Nov 2016).

[20] P. Dillenbourg, "Exploring neglected planes: social signs and class orchestration," in *Proceedings of the International Conference of Computer-Supported Collaborative Learning (CSCL2009)*, 2009, pp. 6–7.

[21] J. D. Gould and C. Lewis, "Designing for usability: key principles and what designers think," *Communications of the ACM*, vol. 28, no. 3, pp. 300–311, 1985.

[22] F. Wang and M. J. Hannafin, "Design-based research and technology-enhanced learning environments," *Educational technology research and development*, vol. 53, no. 4, pp. 5–23, 2005.

[23] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture," *Instructional science*, vol. 32, no. 1, pp. 1–8, 2004.

[24] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally?: cognitive load and multimodal communication patterns," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 129–136.

[25] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educational Psychologist*, vol. 38, no. 1, pp. 53–61, 2003.

[26] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[27] W. Boucsein and R. W. Backs, "Engineering psychophysiology as a discipline: Historical and theoretical aspects," in *Engineering psychophysiology. Issues and applications*. Lawrence Erlbaum Associates, Inc. Mahwah, NJ, 2000, pp. 3–30.

[28] M. L. Reyes and J. D. Lee, "Effects of cognitive load presence and duration on driver eye movements and event detection performance," *Transportation research part F: traffic psychology and behaviour*, vol. 11, no. 6, pp. 391–402, 2008.

[29] R. Buettner, "Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology," in *KI 2013: Advances in Artificial Intelligence*. Springer, 2013, pp. 37–48.

[30] T. de Greef, H. Lafeber, H. van Oostendorp, and J. Lindenberg, "Eye movement as indicators of mental workload to trigger adaptive automation," in *International Conference on Foundations of Augmented Cognition*. Springer, 2009, pp. 219–228.

[31] R. W. Backs and L. C. Walrath, "Eye movement and pupillary response indices of mental workload during visual search of symbolic displays," *Applied ergonomics*, vol. 23, no. 4, pp. 243–254, 1992.

[32] J. H. Goldberg and A. M. Wichansky, "Eye tracking in usability evaluation: A practitioners guide," *To appear in: Hyönä*, 2002.
[33] A. Poole and L. J. Ball, "Eye tracking in hci and usability research," *Encyclopedia of human computer interaction*, vol. 1, pp. 211–219, 2006.
[34] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decision support systems*, vol. 62, pp. 1–10, 2014.
[35] D. Kahneman, J. Beatty, and I. Pollack, "Perceptual deficit during a mental task," *Science*, vol. 157, no. 3785, pp. 218–219, 1967.
[36] H. Schultheis and A. Jameson, "Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods," in *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, 2004, pp. 225–234.
[37] R. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," *Mind*, vol. 2, no. 3, p. 4, 2003.
[38] J. Klingner, "Measuring cognitive load during visual tasks by combining pupillometry and eye tracking," Ph.D. dissertation, Stanford University, 2010.
[39] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM, 2010, pp. 141–144.
[40] B. G. Stuijfzand, M. F. Van Der Schaaf, F. C. Kirschner, C. J. Ravesloot, A. Van Der Gijp, and K. L. Vincken, "Medical students' cognitive load in volumetric image interpretation: Insights from human-computer interaction and eye movements," *Computers in Human Behavior*, vol. 62, pp. 394–403, 2016.
[41] A. Szulewski, A. Gegenfurtner, D. W. Howes, M. L. Sivilotti, and J. J. van Merriënboer, "Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool," *Advances in Health Sciences Education*, pp. 1–18, 2016.
[42] D. F. Feldon, "Cognitive load and classroom teaching: The double-edged sword of automaticity," *Educational Psychologist*, vol. 42, no. 3, pp. 123–137, 2007.
[43] B. Xie and G. Salvendy, "Prediction of mental workload in single and multiple tasks environments," *International journal of cognitive ergonomics*, vol. 4, no. 3, pp. 213–242, 2000.
[44] L. P. Prieto, Y. Wen, D. Caballero, K. Sharma, and P. Dillenbourg, "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*. ACM, 2014, pp. 339–344.
[45] L. P. Prieto, K. Sharma, Y. Wen, and P. Dillenbourg, "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*, 2015, pp. 212–219.
[46] S. Do-Lenh, P. Jermann, A. Legge, G. Zufferey, and P. Dillenbourg, "Tinkerlamp 2.0: designing and evaluating orchestration technologies for the classroom," in *European Conference on Technology Enhanced Learning*. Springer, 2012, pp. 65–78.
[47] T. De Jong, "Cognitive load theory, educational research, and instructional design: some food for thought," *Instructional Science*, vol. 38, no. 2, pp. 105–134, 2010.
[48] L. P. Prieto, K. Sharma, P. Dillenbourg, and M. J. Rodríguez-Triana, "Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 2016, pp. 148–157.
[49] M. Ziefle, "The influence of user expertise and phone complexity on performance, ease of use and learnability of different mobile phones," *Behaviour & Information Technology*, vol. 21, no. 5, pp. 303–311, 2002.
[50] J. Cohen, *Statistical power analysis for the behavioral sciences (revised ed.)*. New York: Academic Press, 1977.

Luis P. Prieto received his Ph.D. in Information and Communication Technologies from the University of Valladolid (Spain). He is a Marie Curie Fellow at the CHILI Lab in the École Polytechnique Fédérale de Lausanne (EPFL). His research interests include classroom orchestration and the use of wearable technologies to capture and understand complex practice with ICT and the use of such data for reflection. He has authored more than fifty academic publications on these topics.

Kshitij Sharma received his Ph.D. in Computer Science from the cole Polytechnique Fdrale de Lausanne (EPFL, Switzerland). He is a Postdoctoral researcher at the CHILI lab in EPFL and at the Faculty of Business and Economics (HEC) in the University of Lausanne, Switzerland. His research interests include eye-tracking, MOOCs, collaborative learning, big data analysis, and statistics. He has authored more than twenty five academic publications on these topics.

Łukasz Kidziński is a researcher in the CHILI Lab at the EPFL. His research involves designing adaptive learning systems, analysing datasets from massive open on-line courses and managing other data-heavy projects in the laboratory. In 2014, he received a Ph.D. in mathematical statistics from the Université Libre de Bruxelles, developing spectral methods for functional time series. He received two master degrees, in mathematics and in computer science, at the University of Warsaw.

Pierre Dillenbourg is a former school teacher and graduated in educational science (University of Mons, Belgium). He obtained a PhD in computer science from the University of Lancaster (UK), in the domain of AI for educational software. He is currently full professor in learning technologies in the School of Computer & Communication Sciences, where he is the head of the Computer-Human Interaction for Learning & Instruction Lab. He is also the academic director at the Center for Digital Education.