Dear Dr. Drachsler,

This document describes the changes made to the manuscript entitled: “Multimodal Teaching Analytics: Automated Extraction of Orchestration Graphs from Wearable Sensor Data” (Manuscript ID JCAL-17-149), addressing all the comments made by the two reviewers, as per your email on the 5 September 2017.

We would like to sincerely thank the reviewers for their time and their insightful comments. We hope to have addressed them satisfactorily.

Reviewer 1

Comment	Response	Changes made in the new version
The evaluation of the model shows that the author(s) want to have a somewhat comprehensive evaluation. There are Tables 1-8 which provide much details but take up much space, and so I wonder if they can be reduced.	We agree that the tables using different results of the four evaluation schemes we tried, for both personalized and generalized models, take up much space, and do not enable easy overall comparison between models. Following also other reviewers’ comments (see below), probably simplifying the evaluation schemes and substituting the tables with graphical summaries (i.e., plots of the results) would help in reducing the space results take and enabling easier comparison.	Tables 2, 4, 6-9 from the original version of the manuscript have now been substituted by more synthetic graphs summarizing the results of the models’ evaluations (Figures 3-4 in the new manuscript). The number of evaluation schemes reported has also been reduced (see comments below), leading to some reductions in the text (removal of original Figure 2, and modification of the text in p.11? describing the evaluation schemes, as well as the modification of text on the discussion of limitations of our evaluations in p. XXX). Also, we have simplified the number of results shown in terms of models built with different isolated data sources (e.g., video features only) that appeared in the (now removed) Table 2, only mentioning them briefly in the text for brevity’s sake (p. XXX).
In such kinds of work, what is an acceptable level of F1 scores? Figure 3 shows the actual and automatically-extracted orchestration graphs of a session, with F1 scores of 0.7-0.8 reported. Yet it seems that the two graphs have quite some	a) We agree that F1 scores (or any other metric) by themselves do not give a good idea of how much the results are within the current state-of-the-art, or are groundbreaking. The most closely-resembling work to ours is that of Donnelly et al. (2016a,	a) Further text about comparable work in the literature and their reported evaluation scores has been added to the discussion in page XXX. b) A brief note explaining the noticeable differences between real

differences, e.g. the method could not pick up the group work interaction, and in the automatically-extracted graph, the class level interaction alternates too much between the different types of activities. On this last point, we know that a human teacher cannot and does not flip-flop or switch too much types of teaching activities in counts of seconds --- so can the neural network learning pick this up, or one can program in such considerations?	2016b, 2017), which also tag classroom activities using multimodal data, report F1 scores of around 0.6-0.7. Other (reasonably) comparable works in the field of multimodal interaction (featured in the ICMI conference and similar venues) achieve top *accuracies* of 60-80% (and often, much lower), when using unstructured data like audio/video to extract activities (e.g., Morency et al., 2013; Chahuara et al., 2016; Dhall et al., 2017). b) Noticeable differences still visible between the real and automatically-extracted orchestration graphs, is, in our opinion, related to the fact that such graph represents not one, but *two* simultaneous prediction/extraction tasks. Hence, even with F1 scores around 0.7 there is a sizeable chance that *either* activity or social plane will have been mis-predicted. However, we thought that providing such graphical example (often absent from this kind of literature), would give an idea that our progress in this area, while remarkable, is still not yet ready for real-world/commercial use. c) Finally, we agree with the reviewer in the observation about rapid changes in the automatically-extracted activities. We were hoping that models like the LSTMs would pick on that (but probably insufficient data is available to train this kind of models yet, see comments below). Since the time of the original writing, we have used Markov Chains to understand the probability of transitions between states in the orchestration graph, and use that to enhance our predictions and provide a performance edge (now added to the new version of the manuscript).	and extracted orchestration graph in the graphical representation has also been added in page XXX. c) An additional kind of time-aware model (based on the aforementioned Markov Chain enhancement of a time-independent random forest model) has been tried on the data, and added to the results section (Figures 3-4, and text on pages XXXX).

Some of the spaces and decisions on what to pre-process the data are not explored. For example, the episodes are 10-seconds. What is the justification for 10 seconds as a unit of analysis? Does a longer episode improve the mean F1 scores?	We agree that this point is not adequately justified in the original manuscript. The choice of 10-second episodes was made on the basis of our initial work in orchestration graph extraction (ref. anonymized for review), where different episode lengths were tried: 10 seconds seemed to lead to better performance, and also was more adequate for the manual coding of episodes (as it is often hard to assign a code to a 1-second piece of action when transitioning between actions, or to assign a single label to a 30-second-long chunk of video action).	Clarifications have been added in p. XXX regarding the episode length choice.
Some theoretical justification (or literature review) of the types of teaching activity would strengthen the classification. Some examples would also help. Are they really mutually exclusive (e.g. questioning could be part of explanation or monitoring)?	We agree the classification scheme used for our experiments is not sufficiently justified in the original manuscript. The classification of the teaching activities is based on previous observational studies of classroom activities and routines (such as Fogarty et al., 1983; Prieto et al., 2011) as well as classroom observation schemas (concretely, the Flanders Interaction Analysis Categories (FIAC) -- see Flanders, 1970). These literature-driven categories were refined with the help of participant teachers (i.e., what kinds of activities were interesting for them) during a participatory preparation phase of the experiments. The activities are mutually exclusive, as in each of these kinds of activities the teacher's *immediate* intent is quite distinct (e.g., transmitting knowledge to students in explanation, solve a student problem in repairs, assess students' understanding in questioning, etc.). We are not considering or trying to infer the implicit intention of wider arcs of discourse in a lesson. Nevertheless, we agree that the speed and fluidity of classroom discourse can sometimes make it hard to tease these intentions	We have added a more extense justification of the categories chosen for teaching activities, including how we derived them and examples, in p. XXX.

	apart, and we definitely agree that more examples can enhance the reader's comprehension of the experiments we set up.	
I would like to read some explication of how the tagging of the episodes in the class going through the different planes of social interaction can help in teacher reflections, and some sense of what the educational research says about the spread of the 5 types of teaching activities in what types of classroom sessions. This can better motivate the work beyond the technical contributions.	We agree that the pedagogical/theoretical purpose of our multimodal analytics task was insufficiently explained. Our main overall goal to eventually provide tools for teachers' own reflection about their practice, which are currently insufficiently based on evidence about daily practice (Marcos et al., 2011). More concretely, the value of tagging social planes of interaction can be traced back to Vygotsky's socio-constructivism (1978), and is often included in classroom observation schemes (Richards & Farrell, 2011). The specification of social planes of interaction is a common practice in instructional design (especially, of collaborative learning, see Dillenbourg & Jermann, 2007), and hence it can be especially useful to track deviations between the intended instruction and the actual classroom enactment (see Lockyer, Heathcote & Dawson, 2013; and other work on the value of aligning learning design and learning analytics). An explanation of the value and origin of the teaching activities coded is provided in the answer to the previous comment. However, it should be noted that by providing such automatic coding we are not aiming at some kind of direct assessment tool on the basis of the observed/extracted codes (i.e., some distribution of activities and/or social planes being inherently better than others). Rather, we aim at providing "mirroring tools" for teachers' own personal reflection, as the teacher (and her contextual knowledge of the classroom) is	An extensive explanation has been added about the overall pedagogical aims of our research (p. XXXX), and the theoretical underpinnings and process to derive the sets of codes used for the automatic extraction of orchestration graphs (p. XXX).

	the best judge for what mix of activities is most desirable.	

Reviewer 2

Comment	Response	Changes made in the new version
One motivation for this study is the employment of deep machine learning techniques (recurrent neural networks) over more shallow ones. The examples given are in the field of image processing and speech recognition. The best choice of the model however depends by the nature of the data. Why did the author decide to use deep learning?	The use of deep learning techniques was not in itself a motivation for the study: rather, the main motivation being to understand what models could help us exploit the time structure of the classroom-recorded signals, to extract automatically useful pedagogically-meaningful constructs. This time structure had been insufficiently explored in our own previous work (ref. anonymized for review) and in much of multimodal learning analytics literature. We agree with the reviewer that the use of deep learning techniques might not be sufficiently motivated (the successes in speech/audio and video processing, and the importance of audiovisuals and speech in most of our dataset seemed enough to warrant some experimentation). Also, that the original text might mislead the reader about the relative importance of using deep learning techniques *per se*.	The overall motivation to exploit the time structure of the classroom-recorded dataset (p. XXXX) and the particular motivation to try out deep learning techniques (p. XXXX) have been further clarified, to address these concerns.
Another rationale for this study - claimed in page 4 - is the need of more automatic systems that avoid the error-prone manual processing. However, the labelling of the teaching episodes for this study is still done manually. How do you foresee this happening in a more automatic fashion?	It is indeed one of our ultimate goals (aside from supporting teacher daily practice and reflection, see the last answer to comments of Reviewer 1), to eventually develop tools that can relieve human researchers from the repetitive task of providing (simple) codes from audio/video. The possibility of this kind of approach, and first prototypes, are starting to appear in the literature (e.g., Fong et al., 2016). However, so far they	The rationale and future outlook of this kind of research work with regard to researcher tools (and the need for initial human coding) has been further explained in the revised version of the manuscript (p. XXXX).

	invariably require some amount of manual human coding (at the beginning at least) to provide a baseline on which machine learning models can act. We believe such next-generation researcher tools are indeed one of the major directions for the deep datasets that multimodal learning analytics research gathers (ref. anonymized for review).	
Please explain what is the the class distribution (frequency) of the 5 action codes and 4 planes of interactions.	We agree that showing the distribution for the different classes can be helpful to spot class imbalance problems (as there are, indeed). That is also the reason why we chose F1 scores as the main evaluation metric (as opposed to e.g., accuracy, which is more prone to misleading results due to class imbalance).	We have added a new figure (Figure 2 in the new manuscript) showing the code distributions for the two classification activities, as well as comments about the class imbalance and evaluation metrics, in p. XXXX.
The article lacks some examples. It would be especially beneficial to know, on what kind of activity and which level of detail predictions are expected from the model to understand the given data better. Is a predefined set of activities used? Are activities also learned from observation?	We understand this comment as similar to the other reviewer's on the lack of a justification about the coding scheme used for both classification tasks (basically, derived from literature and refined in a participatory co-design with the participant teachers), as well as examples that may help better understand both classification tasks. We agree, and have modified the manuscript accordingly (see next).	We have added text explaining in more detail how the coding schemes where derived, and examples of the different codes, in p. XXXX.
One of the research questions is not clear (page 13): "to what extent the qualitative differences in terms of the lesson's instructional designs affect this effectiveness?"	The fragment of research questions mentioned by the reviewer, was meant as a reference to the different evalutation schemes tried out in the original manuscript (at the session level, at the kind-of-situation level, or at the teacher level). Nevertheless, we agree that the wording was rather obscure (if not outright misleading). Since now these different evaluation schemes have been greatly simplified (see this same reviewer's comment below), that piece of research question is not needed anymore,	The research questions (in p. XXXX) have been simplified, taking away the controversial fragment. Similar fragments and related remarks have also been taken away throughout the paper (e.g., in the abstract).

	as it is not thoroughly investigated in the new version of the manuscript.	
Please explain why did you include the „rather obscure and much more numerous“ audio features in table 3, if they come without any explanation (p. 15)?	The inclusion of such a high number of features (especially, audio features) that are not necessarily interpretable stem from the nature of the machine learning task we were trying to tackle, and our ulterior motives to do so: ascertaining the immediate teaching activity and social plane of interaction is a rather easy task for even a non-expert human, and we are not trying to better understand the process of how this coding task is performed (rather, we only want to predict/replicate human results). This enables us to prioritize less the interpretability of results, and more the predictive power of our models and features (i.e., the use of such low-level audio features, and black-box machine learning models that find their own rules and sources of information) -- which is what we tried in the current study. In any case, we agree with the reviewer that this issue, and the reason to include such features, is worth expliciting in the manuscript.	Text explaining this aspect of our approach to the automated coding/extraction task (in p. XXX) has been added.
The authors did not specify how did they perform feature extraction. The audio data produces more than 7K features which is more than the size of the data points (5K). It is really strange that you manage to learn anything having more features than training samples. More convincing sounds the feature extraction approach, in which the 100 best features are selected. On this note, please discuss more how feature ranking was performed.	a) We agree with the reviewer that the limited dataset is one of the major weaknesses of our work, and the most probable reason for the LSTM models' low performance. However, the probabilistic/ensemble nature of some of our models (for instance, the high number of trees making up the random forest) enabled b) Regarding the feature selection/dimensionality reduction approaches used, in many cases this was done via Principal Component Analysis. The SVM using 100 best features, on the other hand, used a feature ranking performed by removing features that are highly	a) Explanatory text about this issue has been added in p. XXXX. b) The feature ranking used for the SVM models has been explained more extensively in p. XXX.

	correlated with each other, and using a measure of effect size: Cohen's *d* (Cohen, 1988), which gives an idea of how different the values of a feature are for different activity or social plane codes (an approximation to their predictive value for those codes). We agree this was not sufficiently described in the original manuscript.	
I have the feeling that the paper is aiming at too many analysis and is not going in depth to any of them. For example, with a dataset containing only 2 teachers it is difficult to train personalised generalisation models (i.e. models that adapt to each individual teacher). Similarly the role of time is neither explored in depth.	We agree with the reviewer that our original manuscript was maybe trying to "bite more than it could chew", given the limited dataset, especially in terms of teachers (only 2) and kinds of situations (only 5). Also, these different levels of evaluation made the description of results rather complex and the text, convoluted. We think that describing the results of the leave-one-session-out evaluations (LOSO, in the original manuscript), both for personalized and general models, simplify the paper, increase readability, but still provide an idea of the potential of our approach (acknowledging the limitations of the dataset and such evaluation). Prompted by this and other comment from the reviewers above, we have also added another kind of time-aware model to our study, so as increase the depth of our exploration of the time dimension (see the comments about Markov Chain-enhanced models above).	The methods section (p. XXXX) and the results sections (pp. XXXX) have been simplified to describe only the results of the leave-one-session-out evaluation schema. Anoter time-aware model (using Markov Chains) has been added to the results sections (pp. XXXX and XXXX), and further discussion about the limitations of the datasets and evaluatioin schemas has been added in p. XXXX.
The use of the look-back technique is indeed probably too naïve. Considering the features of the previous 9 sounds like what happen one feature episode t-9 is equally relevant as the same feature in t0. The RNN approach seems more convincing than the look-back. However it does not produce satisfactory results. Any idea why?	We also agree that the look-back approach was in principle too naive (and that was our assumption as well at the beginning), although we thought that the random forest's inherent feature selection abilities could naturally pick up the relative importance of the different past episodes. Indeed, looking at the feature ranking in the original manuscript's Table 5 (now Table XXX), we can see how the	More extensive explanations about the "naiveté" of the look-back approach (p. XXXX) and the probable reason for LSTM's low performance (p. XXX) have been added to the manuscript.

	random forest was automatically selecting as more important the features of t0, but still taking into account some of the data of t-1, t-2, etc. to make its predictions. Regarding the low performance of the RNN approach, we believe it is rather a limitation of the dataset used (i.e., relatively low number of data points), as these algorithms have shown their more spectacular results on very large datasets.	
We suggest the author consider the use of graphs instead of or in addition to tables for presenting the results to allow allow graphical understanding for the reader.	We agree that most of the result tables comparing the performance of the different models in the classification tasks, could be communicated more clearly and synthetically using graphs. Given the k-fold cross-validation schemes we were using (which produce a F1 score for each evaluated fold), we have chosen to show the distribution of results using boxplots.	Tables 2, 4, 6-9 from the original manuscript have now been substituted by graphs summarizing the results of the models' evaluations (Figures 3-4 in the new manuscript).
In the discussion section, the authors describe that they used "multimodal features with relatively low semantic value" (p. 19). It is not explained, what exactly the semantic value is. Please provide an explanation.	The "low semantic value" used in the original manuscript refers to the low interpretability of the features. We agree that the term was not very clear, and that talking about "interpretability" (i.e., whether we can make certain inferences about how the model works or makes decisions, by looking at the features and their values) is probably more accurate.	The references to "low semantic value" have been substituted with similar ones about the "low interpretability" of features (pp. XXXX).
Minor: **- Please correct reference to figures and tables: (fig 1/fig 2) (p. 8), table X on page 14** **- Please provide an explanation for the acronym LOSO, LOSitO or LOTO before using them for the first time in the document.**	a) We agree, and thank the reviewer for spotting these mistakes. b) The evaluation schemes have been simplified (as per comments above), and hence most of the references to LOSO/LOSitO and LOTO have been removed anyways.	a) References to figures and tables have been fixed. b) Acronyms LOSO/LOSitO/LOTO have been removed throughout the text. Furthermore, the whole text has been revised to remove typos and other minor errors that slipped the first manuscript submission.

References

(some references anonymized for review)

Chahuara, P., Fleury, A., Portet, F., & Vacher, M. (2016). On-line human activity recognition from audio and home automation sensors: Comparison of sequential and non-sequential models in realistic Smart Homes. *Journal of Ambient Intelligence and Smart Environments*, 8(4), 399-422.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hilsdale. NJ: Lawrence Earlbaum Associates.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (2017). From Individual to Group-level Emotion Recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

Dillenbourg, P., & Jermann, P. (2007). Designing integrative scripts. In F. fischer, H. Mandl, J. Haake, & I. Kollar, F. fischer, H. Mandl, J. Haake, & I. Kollar (Eds.), *Scripting computer-supported collaborative learning: Cognitive, computational and educational perspectives*. Springer Computer-supported Collaborative Learning Series.

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference* (pp. 218–227). Vancouver, British Columbia, Canada: ACM.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., … D'Mello, S. K. (2016a). Automatic Teacher Modeling from Live Classroom Audio. In *Proceedings of the 24th International Conference on User Modeling Adaptation and Personalization* (pp. 45–53). Halifax, Nova Scotia, Canada: ACM.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., ... & D'Mello, S. K. (2016b). Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 177-184). ACM.

Flanders, N. A. 1970. *Analyzing teaching behavior*. Reading, MA: Addison-Wesley. http://www.ascd.com/ASCD/pdf/journals/ed_lead/el_196112_flanders.pdf

Fogarty, J. I., Wang, M. C., & Creek, R. (1983). A descriptive study of experienced and novice teachers' interactive instructional thoughts and actions. *The Journal of Educational Research, 77*(1), 22-32. URL: https://files.eric.ed.gov/fulltext/ED216007.pdf

Fong, A., Hoffman, D., & Ratwani, R.M. (2016). Making sense of mobile eye-tracking data in the real-world: A human-in-the-loop analysis approach. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*, 1569-1573. SAGE Publications.

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *The American Behavioral Scientist, 57*(10), 1439-1459.

Marcos, J. M., Sanchez, E., & Tillema, H. H. (2011). Promoting teacher reflection: What is said to be done. *Journal of Education for Teaching, 37*(1), 21–36.

Morency, L. P., Oviatt, S., Scherer, S., Weibel, N., & Worsley, M. (2013). ICMI 2013 grand challenge workshop on multimodal learning analytics. *In Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 373-378). ACM.

Richards, J. C., & Farrell, T. S. (2011). Classroom observation in teaching practice. In *Practice teaching: A reflective approach*, 90-105.