Paper prototype tests @ Leysin American School

Internal report

Luis P. Prieto and Paul Magnuson

08 March 2016

Abstract

This document describes the main results from the TAKK paper prototype tests performed at the Leysin American School the 8-24 February 2016, involving 9 teachers. The different quantitative and qualitative data show that it is feasible to obtain a respectable amount of data about both student experience and teacher practice, in a relatively short time. We also found that both teacher- and student-driven data are interesting for teachers (and probably should be used simultaneously), and that we might have to rethink slightly the app implementation strategy, in terms of the actions/experiences to be tracked, or the app's main form factor (as the desktop experience was preferred over the mobile one).

Introduction

This report is part of a joint effort between Leysin American School (LAS) and the CHILI Lab of EPFL. The main goal is to develop **technologies and approaches to generate data about everyday teaching practice** in a non-scary way, and to foster **evidence-based teacher reflection**, especially in the context of professional development conversations as encountered in that school.

LAS had already done a first paper prototype pilot with six teachers over two weeks. During that time, up to 173 teacher annotations were recorded, with collaborative group work being the most often noted (31% of the annotations). The participant teachers provided encouraging feedback on its value as a reflective tool, but warned that such annotations were hard to do *during* the lessons. This gave an initial idea that the approach was feasible (at least on the short term), but still did not provide much detail regarding which form the application should take, and how it should be rolled out to best fit the needs and life of the school. For instance, the question had been posed as to whether the *students* could make the annotations (instead of the teachers), or whether an "automated recording/annotation" system (as used in some of CHILI's previous work) would be better valued.

In this second, more formal trial, we aimed at responding to several such questions, which would help to guide the software development of the application (e.g., for priorization of features), and its eventual roll-out at LAS. The main questions we will respond to here are:

- **Q1**: What kind of practices/experiences did the teachers record (i.e., what data was gathered with the new paper prototypes)?
- **Q2**: Did teachers use the prototypes? (including the use and effectiveness of reminders)
- **Q3**: Did teachers prefer the teacher-driven approach, or the student-driven use of the application?
- **Q4**: In the case of the teacher-driven app, which form factor was preferred (mobile, desktop, poster...)? How did they use it?
- **Q5**: When did teachers make the annotations (e.g., during the lesson, or after it)?
- **Q6**: Would teachers record media when doing the annotations? If so, what kind of media would they use?
- **Q7**: What kind of reporting and statistics would teachers like to have?

- **Q8**: How often and for how long would teachers use the app?
- **Q9**: Are there problems of trust and data ownership in the use of the app?
- **Q10**: Do teachers perceive an added value to the use of the app? (including whether the current icons are the right ones)
- **Q11**: Would teachers like to have an "automated recording/annotation" system?
- **Q12**: What other emergent themes and issues were observed during the study?

Method: To answer these questions, we worked with 9 teachers, to whom we provided paper prototypes to be used (either by them or the students) for approximately one week (4 school days). Short interviews were done at the end of the two weeks to get the teachers' impressions, and then a personalized report was generated for each teacher with the data they provided (some teachers have also provided additional feedback on the basis of this personalized report).

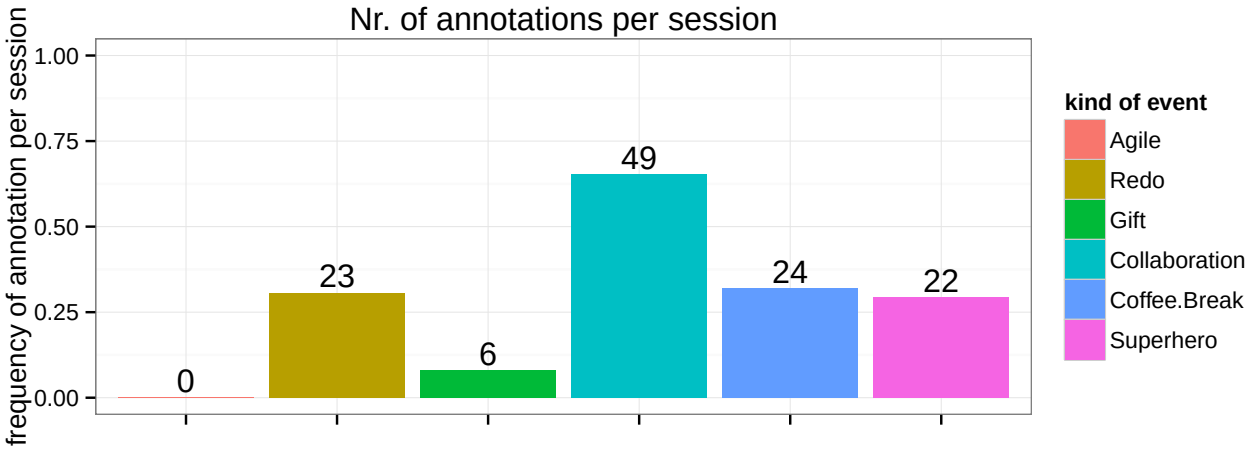
The rest of the document reviews each of the previous questions, providing responses in light of the quantitative and qualitative data gathered in the study, which will help plan the software development of the app and its eventual roll-out in the school.

Research results

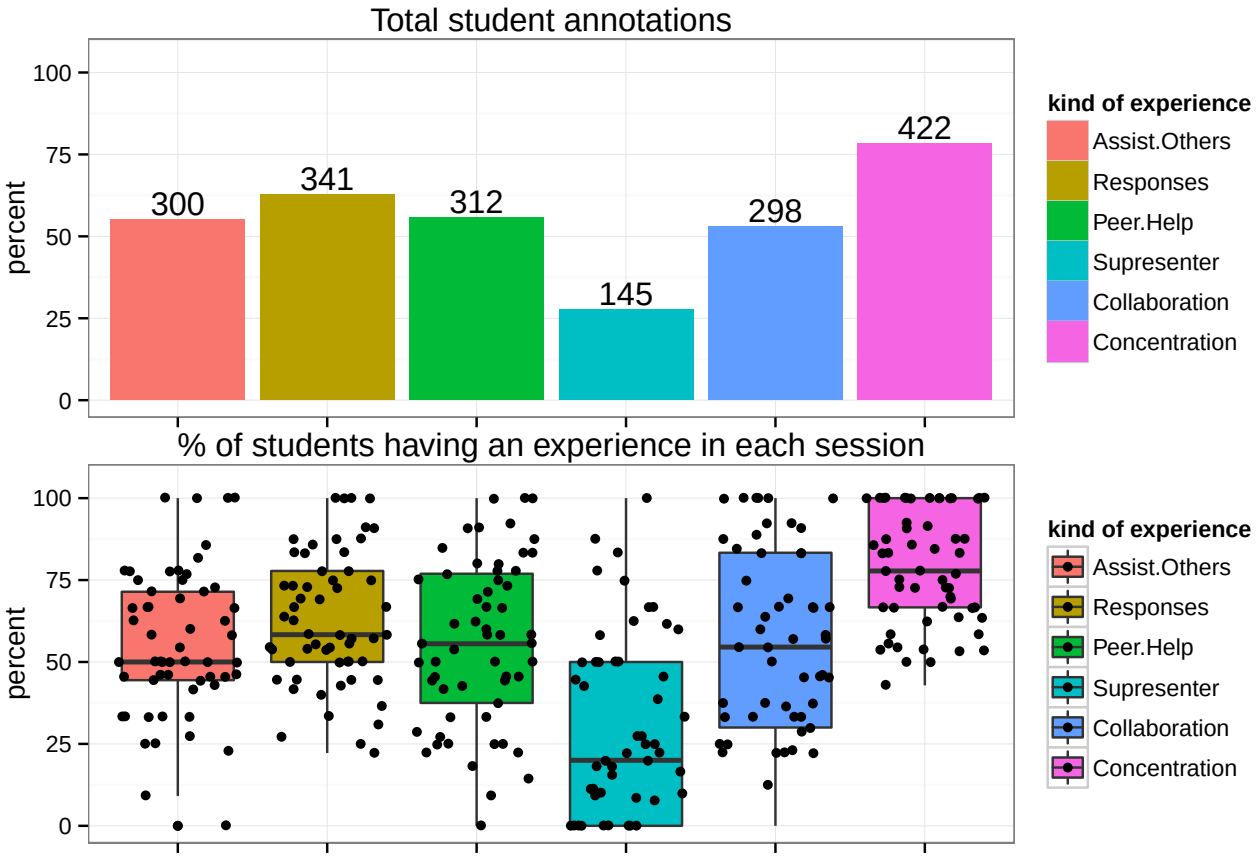
From the interviews and meetings with the teachers, we can gather that the teachers come from a wide variety of backgrounds, and teach many different subjects, from physics to economics or languages. The fact that most of them are quite experienced (with 15 or more years of experience), and that they volunteered for this study, shows that they may not be necessarily representative of the rest of the LAS teachers – hence, any school-wide extrapolations should be handled with care. Other than this bias towards the experienced and "volunteerist" side of the spectrum, teachers showed a variety of different preferences in terms of how professional development should be, and on which parts of their teaching practice they find more challenging (from classroom management or use of technology, to time pressures or student heterogeneity/differentiation).

Q1: What kind of practices/experiences did the teachers record (i.e., what data was gathered with the new paper prototypes)?

Teacher-driven approach. Throughout the two weeks of the study, teachers made up to **132** teacher-driven annotations from **75** different sessions, recorded over 8 different school days. The graph below shows the total distribution of annotations, as well as their average frequency per session (confirming roughly the results of the previous pilot, in terms of collaboration being the most frequent annotation).



Student-driven approach. Throughout the two weeks of the study, teachers gathered **551** student-driven annotations from **57** different sessions. The graph below shows the percentage of students in a certain session that marked each of the six kinds of student experiences, on average. The points show the values for each of the recorded sessions, giving an idea of the variance and spread of the student experiences across all participant teachers:



Indeed, even in the same session, student experiences also varied from one student to another (e.g., not *all* students will probably have the chance to act as a presenter, in a single session). For reference, the average similarity in the data from all teachers was 0.69[1].

Q2: What was the overall response to the study? Did teachers use the prototypes? (including the use and effectiveness of reminders)

When asked point-blank about how the experience of the study had been, teachers almost unanimously responded *positively*. Interestingly, those that jumped right into the difficulties of using the app, were mostly referring to the teacher-driven version of the prototype (and especially in doing annotations in the middle of a lesson). The major complaint points were about the *lack of representativity of (some of) the icons/actions* (more on that in Q10 below), and the *remembering to do the annotations* (which goes along the hypothesis that teachers are already by default cognitively overloaded, and additional tasks can be easily forgotten or overlooked).

On this issue of remembering to do the annotations, during the study we offered the option to setup *reminders* for the teachers using different means (email, whatsapp, calendar, etc.). Eight out of the 9 teachers chose to have reminders set for them (and the remaining teacher requested such reminders too after the first

[1] A similarity of 1 means totally uniform experience among all students (quite unlikely), and a value of 0 means totally different experience for every student (also unlikely).

week of the study). Most teachers (77%) asked for *daily* reminders (often, first thing in the morning), and most of them (55%) chose the reminders to be delivered by email, with the rest of them asking either a (Google) calendar alert or a WhatsApp message. In the interviews, teachers invariably found them very useful throughout the study. However, even using these reminders, several teachers stated that they had gathered data/annotations of around 80-90% of the sessions (e.g., one teacher said that one of the days she started the lessons without having read email, and hence forgot to do annotations until the end of the day).

Q3: Did teachers prefer the teacher-driven approach, or the student-driven use of the application?

This question is harder to answer unambiguously: On the one hand, teachers gathered many more "data points" using the student-driven prototype than with the teacher-driven one (551 vs. 132 – quite logically since there are many more students than teachers). Not only there is "comfort in the big numbers", but also many of the teachers declared that it was easier to remember to do, and to execute the student-driven version (maybe because both students and teachers are used to the dynamic of handing out papers and doing assessments). But, which of the two approaches is most interesting/valuable? When asked directly in the interviews, some teachers considered the student view more interesting, while others valued more the teacher view for the component of reflecting on their own practice, and yet others considered both equally interesting.

However, some teachers may have found a "middle path" solution, by suggesting that both sides are interesting to know (and even, to compare), and that actually the most convenient way to do such annotations is through *simultaneous use of both teacher- and student-driven app* (i.e., at the end of the lesson, while students take two minutes to fill in their experiences, teacher can do a quick annotation from his perspective). This was further confirmed during the interviews: when asked directly, all teachers agreed that such a use "would be interesting", or that it would "make sense".

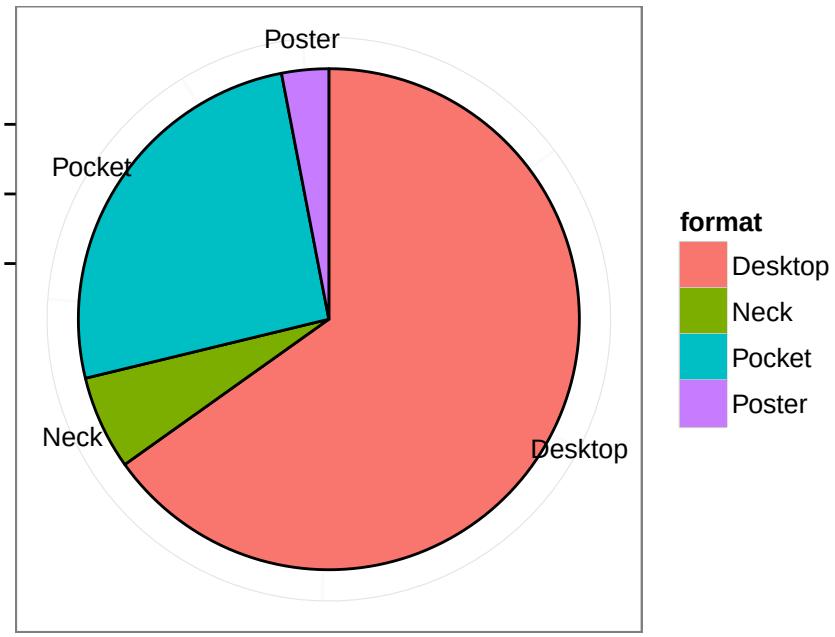
Also related to this issue is the fact that, as some teachers noted, sometimes the students may have filled out their annotations with a certain lack of attention, or under misunderstandings about what each item means (or because they like or dislike the teacher), hence raising the issue of how *reliable* the student annotations are. But then again, we saw in some of the interviews and in the previous pilot that there is margin for confusion and interpretation when teachers did their annotations, too. It becomes clear that the experience/event descriptions have to be polished, and that the teachers have to emphasize their meaning to students when passing them around.

Q4: In the case of the teacher-driven app, which form factor was preferred (mobile, desktop, poster...)? How did they use it?

Interestingly, the *desktop* format (an A4 sheet depicting a timeline for each lesson, where the annotations could be placed) appeared to be the most favored teacher annotation means (see also the figure below): 86 annotations were made (out of 132) in this format, by 6 teachers. The second most preferred was the original pocket/mobile version (34 annotations by 3 teachers).

Indeed, several teachers said, on the topic of the desktop format, that they found interesting to record and reflect on *how much time they spent doing the different kinds of activities* – more even than the "event-based" annotation of the classic TAKK application (a quantification that some teachers found a bit over-simplistic).

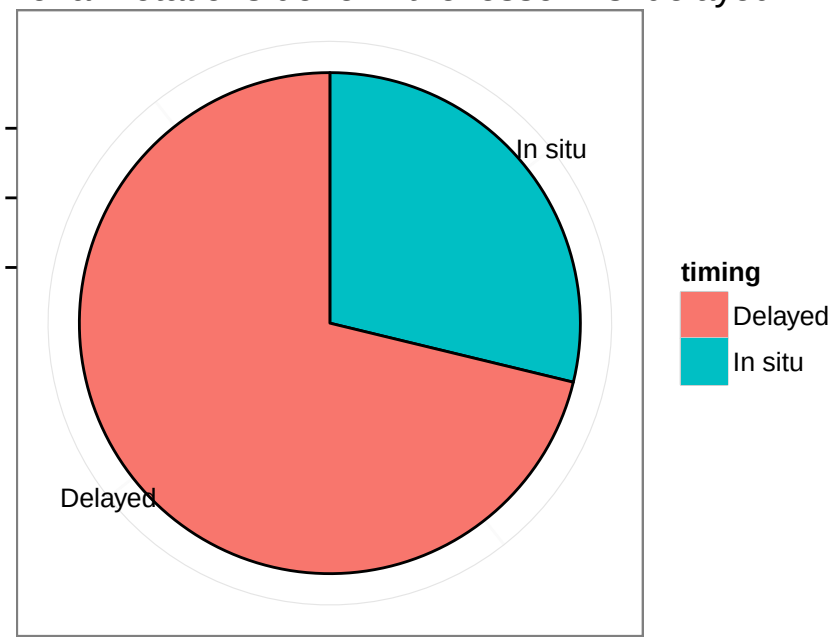
Number of annotations done in the different teacher−driven form factors



Q5: When did teachers make the annotations (e.g., during the lesson, or delayed until the end of it)?

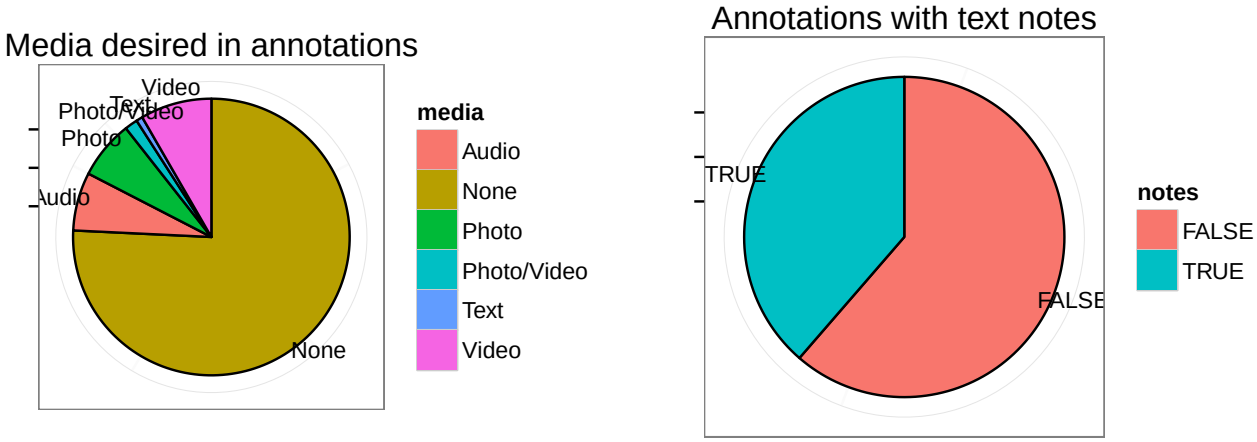
Here again, there was a clear preference for a delayed annotation (94 vs. 38 annotations, see the figure below). Regarding the issue of *how much* the annotations were delayed from the actual lesson, responses varied: teachers often said they would do it just after the lesson, but there were also several cases of teachers delaying the annotation until the end of the day (especially for the cases when they had forgotten to do it before, and often using their own lesson notes/plans as a memory aid).

Number of annotations done in the lesson vs. delayed



Q6: Would teachers record media when doing the annotations? If so, what kind of media would they use?

This feature of being able to record videos, photos, etc. of an episode as it happened, and which was quite present since the inception of the project, received a rather cold response, both in the interviews and in the amount of annotations that stated additional media to be captured (only 32 out of the 132 annotations indicated the need for some kind of recording, see figure below). This may be in part connected with the popularity of the delayed mode of use, which makes such recording features rather pointless. However, adding a short text note to the annotation was used relatively more often (51 out of the 132 annotations, see below).



Q7: What kind of reporting and statistics would teachers like to have?

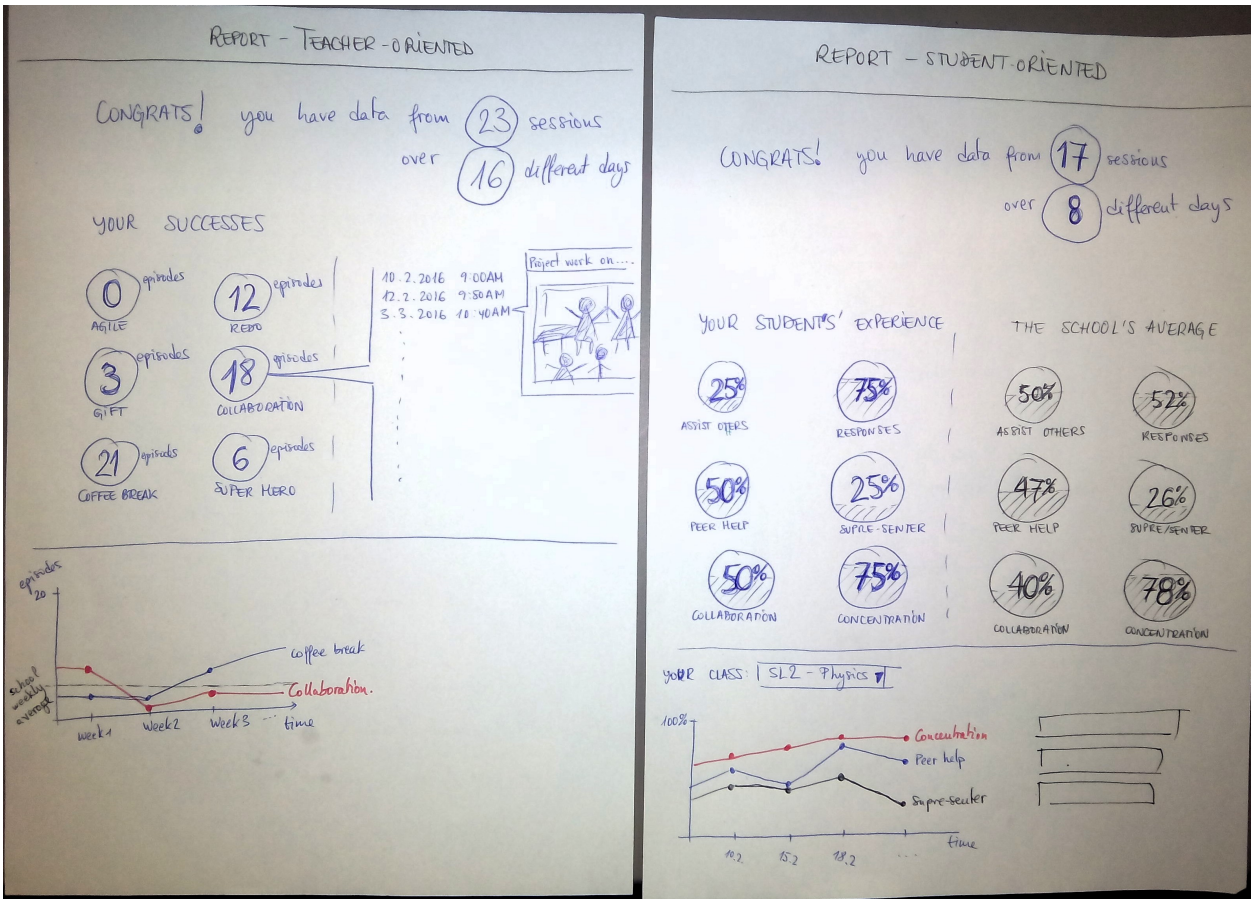
In order to have a first idea of what kind of reporting of the recorded data would be interesting for the teachers, a sketch of the statistics for both the student- and the teacher-driven annotations was shown to the teacher during the interview (see image below).

In general, the statistics shown in the sketch report were considered clear and interesting. Some of the most common suggestions received about the reports include:

- Different teachers found the *comparison* with the average statistics of the whole school useful, to frame the statistics. Others suggested to compare with the department-level, or with explicitly-stated school-defined goals. There were also voices rejecting the notion of comparisons altogether (especially since the student experience profile can be very dependent on the subject matter).
- Also, some teachers suggested that seeing the *teacher- and student-driven statistics side-by-side* could be useful (if a one-to-one matching between both sets of icons could be found)
- Quite a few teachers mentioned that it would be useful to see the statistics *gathered by class* (group of students), even more so than the overall/average student experience profile
- Although a couple of teachers suggested that seeing the information of *individual students* could be useful, most other teachers thought that would probably be excessive detail (and it could make the student responses less reliable, if they are not anonymous).
- Regarding the temporal dimension (e.g., showing the weekly evolution), while some teachers found it acceptable, others suggested that it was not very meaningful given that each week can have very different needs. Other teachers suggested rather doing the temporal aggregation by *curriculum unit* (since activities in each unit often will have a certain balance).
- Several teachers mentioned (spontaneously) the aggregation of the different teacher annotations in terms of time spent in them, rather than as single events (probably prompted by the use of the "Desktop" format of the app, which defines a timeline for the lesson, and time alloted for the different annotations). A teacher further suggested that, if working with temporal aggregation by curriculum unit, the teacher

could self-define the goal for the different mix of events/experiences to be attained in the unit (instead of some fixed school-level value), and then compare the statistics against that goal.

- Some teachers suggested that it could be useful to show (upon request) the detailed timeline of each lesson, including also a small text comment to give a bit of context about why the lesson timeline was like it was.
- Graphical enhancements like color-coding the different values, showing the comparison right next to the personal value, etc.



Q8: How often and for how long would teachers use the app?

During the interviews, teachers discussed what usage dynamic they saw for the app (e.g., whether they would use it all year around, or rather at specific times). Most of the teachers stated that using it continously during the whole academic year would quickly become tiring and lose its usefulness (although some teachers suggested that a select group of teachers could do just that, while the rest of the school did it only at specific designated times). Multiple teachers mentioned sampling the teaching/learning experience *a few times per year* (e.g., doing it for a few weeks at the beginning, in the middle and at the end of the course). Other teachers suggested different sampling strategies (e.g., every Monday at 9am).

Q9: Are there problems of trust and data ownership in the use of the app?

When asked about the issue of data ownership (who owns or can see the data associated to a particular teacher) and how it would affect the usage of the tool, teachers provided a wide variety of responses: most of the teachers said they would not mind the data being shown to a supervisor (for accountability) as long as

the *rules and process* of how the data will be used are *clearly laid out* (what are the expectations, the goals, etc.). However, a few teachers also stated that they would feel more comfortable having the power to decide who to share this data with, and the fact that public or evaluative use of the data can lead to more unreliable uses of the app (thus losing its power); in this sense, there were suggestions for example about sharing the statistics at the *department-level* instead of person-by-person (although there were also suggestions of heads of department having access to each of the teachers' data working with them).

Q10: Do teachers perceive an added value to the use of the app? (including whether the current icons are the right ones)

Probably the most commonly-voice difficulty or complaint among the teachers that participated in the study, as to whether the application provided enough added value, was the fact that some of the icons used (both for the student and for the teacher-driven annotations) were *not relevant, descriptive, or applicable* to their daily teaching practice. This includes items that come up so often in the class dynamic that it makes no sense to track them, or the lack of items related to the subject matter and subject-specific skills, or other aspects of teaching/learning that are of special interest to each specific teacher.

Nevertheless, at different points in the interview almost all teachers expressed that they saw the value of reflecting on what they do, just by the fact of looking at those items and annotating them (or that it would if the actions to be tracked were more relevant to them). There were a few teachers that also pointed out that the value (and the success) of the approach may depend on the local school culture – it may be more relevant for more novice teachers, or in schools with less active or less pedagogically-developed staff, or with a strong top-down enforcement (implicitly suggesting that maybe at LAS its value might be limited).

All these issues point to the importance of being able to *personalize* in some way the set of icons/actions/experiences in the app, not only at the school level, but also possibly at the department and even at the personal level (something that matches LAS's strategy for teacher assessment). However, this kind of personalization would also pose some drawbacks or tensions, such as: a) making comparisons (of statistics) across the whole school difficult; b) being more complex to setup (as schools/departments/individual teachers would have to agree on concrete sets of icons and come up with good definitions for them); or c) being more difficult to understand and fill in correctly for students, who would might see a different set of questions for in every subject/course.

Q11: Would teachers like to have an "automated recording/annotation" system?

During the interviews we also explored the idea of having a technology-driven annotation system, that could gather data about the lesson without explicit teacher/student intervention. Although this kind of system could enable, for example, the measuring and reporting on amounts of time spent in different activities (mentioned in Q7 above), the idea was majoritarily *rejected* by the participant teachers (especially, the notion of having a phone hung from the neck), for a variety of reasons: the lack of interest in gathering additional media about the classroom events, a lack of time for later reviewing such footage, privacy concerns and about how the footage could be used, the effect of having a recording device on the spontaneity of students, etc. Also, one teacher pointed out that such automated means would partly defeat the purpose of the tool, which is to prompt reflection just by asking students or the teacher to remember/describe the lesson.

However, despite this general skepticism, there were teachers that showed interest in having such a system, not so much as a recording device, but rather as an element that could automatically gather useful information that could then be summarized for quick access. As an alternative format to the wearable/mobile application, some teachers suggested that the *teacher's laptop* (which is already part of the classroom environment) could be a viable means for recording, if needed.

Q12: What other emergent themes and issues were observed during the study?

Aside from the aforementioned questions that we had targeted with the study, there are a few other interesting ideas or issues that have arisen throughout the conversations with the LAS teachers:

- *Observing/annotating others*: The app feature for teachers to be able to annotate classroom events for other teachers (e.g., in the context of observing another teacher's lesson), was notoriously absent from the conversations with the teachers, except in those cases where the teacher was already following an observation program at the department level. When considering the app feature roadmap we should consider whether these department-level observation schemes offer a good "testing ground" for a first version of the app (as they might be especially motivated to use the app regularly), but also what new/different functionalities would be needed in the app (and whether those are convenient to do in a first stage) – probably a conversation with Aaron is needed to make things more concrete.
- *Timelines vs. events*: As it has been mentioned above, many teachers found the "timeline metaphor" of the desktop prototype interesting and useful. This kind of graphic metaphor puts an emphasis on time and duration of activities, but... is that focus a desired one, to be followed to let teachers reflect and analyze in what they spend their time, or does it distract from other important aspects of the teaching practice? (and which ones?)
- *Modulate the app according to teaching experience*: Another idea that has arisen several times throughout the study is that teachers with different levels of experience may not benefit so much from the app depending on how we design it. Hence the question is: should we try to target/cover *all* teachers' needs and demands, or do we try to focus on those that may benefit most and will oppose less resistance to the idea?
- *Transversal pedagogy vs. subject matter*: Although this is a classic tension in much of educational research, so far the design of the app – especially, the different actions– were broad pedagogical items, rather than focused on a specific subject matter such as math or languages (to make it more widely appealing independently of the subject matter taught). However, it is clear that some teachers see the subject matter as central to their practice and their professional development. Then, should we keep the broad pedagogical focus of the app to facilitate comparisons and a uniform experience, or should we pursue the path of customization and "subjectification" of the app? "Middle ground" alternatives are also possible, such as packages of subject-specific icons to be made available in the application's library.
- *School nomenclature, timetables*: One interesting (albeit admittedly small) detail in the data gathering was the confusion about how to name the groups, the periods, and the general complexity of the school and the teacher's schedule. This small detail might be important later on for assigning the different annotations and events to their right categories for the statistics. Since it seems like there is a unified, unambiguous way of naming groups of students and periods (through the school's backend system, PowerSchool), at some point the connection of the app with this school backend systems will probably have to be implemented, so that annotations and events are assigned to the right student, teacher and cohort.
- *Systems integration*: Also along these lines, we are already seeing the need and benefits of integrating the app with other existing technologies at LAS, from the teachers' calendars or their emails (for providing much-needed reminders), to suggestions about integrating the app with the existing attendance application, to avoid having the "yet another platform" syndrome. Given the general opinion that in LAS there are "too many platforms" (something that may or may not change in the near future), it is still very soon to consider very complex bridges to other systems which can be quite costly while adding a relatively small value; it is, however, an aspect we should not completely forget about.

Appendix A: Example individual report provided to teachers[2]

(see the next page)

[2]To give an idea of the kind of statistics that we have provided to teachers, and which will make up the first version of the app's dashboard (in the sense of the statistics provided – not the graphical aspect, which will probably be quite different).

Teacher 1

First of all, thanks a lot for your participation, This short report summarizes the data you gathered about your daily teacher practice and student experience, in the hope that you are curious about what you and your students did over these two weeks. This report illustrates the kind of statistics that can be generated using the app we are developing, so if you have any feedback or suggestions about them, please do not hesitate to contact me at `luis.prieto@epfl.ch`.

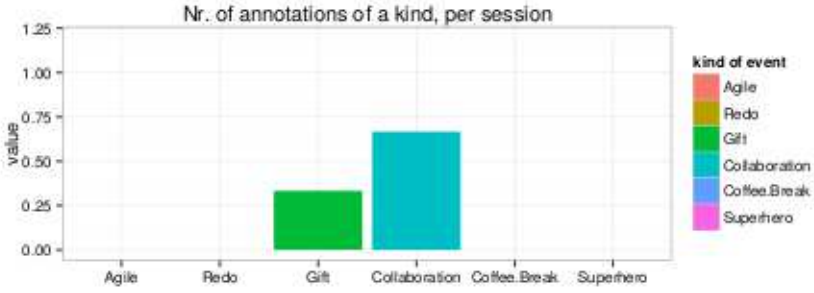
Teacher annotations

Congratulations! You have a total of 6 teacher-driven annotations from 6 different sessions, recorded over 4 different school days.

The kinds of events to record[1] were:

- **Agile**: "time-boxed sprints with visible goals, stand-up planning meetings, scrum boards"
- **Redo**: "students turn failure into success by revisiting work that didn't meet standards"
- **Gift**: "use colleagues' teaching idea in class or colleagues using your ideas in class"
- **Collaboration**: "students do group work for more than 5 mins without teacher interruption"
- **Coffee break**: "students continue on task while you (could) walk out of the class (more than 5 mins)"
- **Superhero**: "a student leads more than 5 mins of a class without teacher interruption"

[1] These were just hypothetical kinds of episodes that might be desirable, and are subject to change when/ if the application is rolled out in all or part of the school.



Naturally, this does not mean that you should teach more like the mean of teachers, or that all values should be at 1 or above. However, we hope that this view gives you a (rough) idea of what kind of things you do more often in your teaching (at least for that week), and maybe provides ideas for things you might want to try more (or less) in the future.

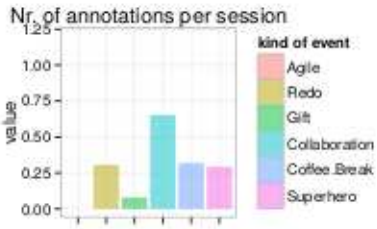


Figure 1: ALL TEACHERS' data

Student annotations

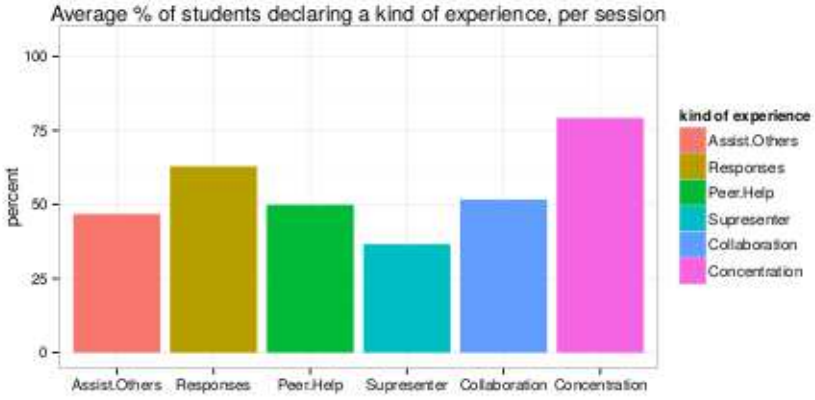
Overview

Congratulations! You have a total of 89 student-driven annotations from 9 sessions, recorded over 5 different school days.

The kinds of events to record[2] were:

- **Assist others**: "I had the chance to help a classmate"
- **Responses**: "I had the chance to ask my individual questions - and I understood the answer"
- **Peer help**: "I had the chance to get help from a classmate"
- **Supre-senter**: "I had time to present in front of a small group or the whole class"
- **Collaboration** :"I had time to work collaboratively with other classmates (more than 5 mins)"
- **Concentration**: "I had time to work individually, which I used productively"

[2] These were just hypothetical kinds of episodes that might be desirable, and are subject to change when/ if the application is rolled out in all or part of the school.



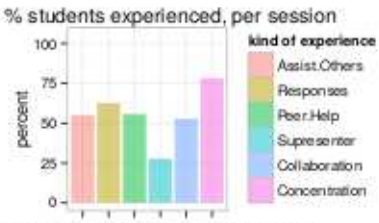
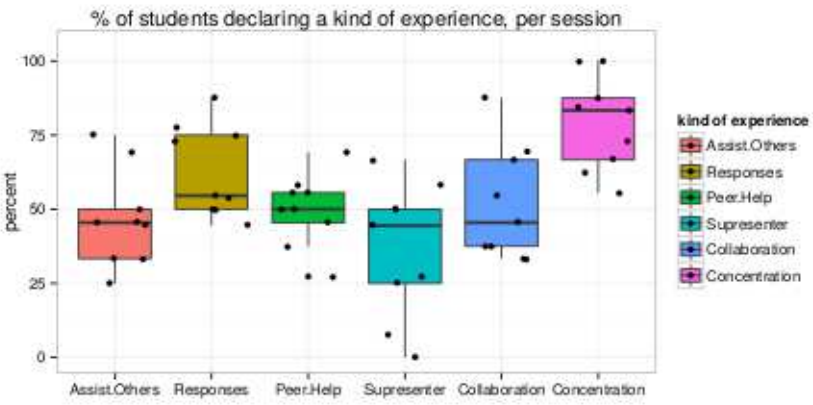


Figure 2: ALL TEACHERS' data

Detailed view

You had a total of 89 student-driven annotations from 9 sessions, recorded over 5 different school days.

As expected, in different sessions the student experiences varied a bit. Below you can see the average student experience for each session (the black points), and a representation of the distribution experiences for each session (the colored boxes) and the median[3] value (the thick black line in the middle of each box). To the right, you can see the distribution of experiences for all the teachers in the school that participated in the study.



Indeed, even in the same session, student experiences also varied from one student to another (e.g., not all students will probably have the chance to act as a presenter, in a single session). In your case, the student experience annotations have an average similarity of 0.56[4].

Naturally, this does not mean that you should teach more like the mean of teachers, or that all values should be 100%. However, we hope that this view gives you a (rough) idea of what kind of experiences your students have, and maybe gives you some ideas for things you might want to try more (or less) in the future.

Note: In the real app we may also include things like the evolution over time, and the data per class/group of students. However, there is not enough data from just one week of use to make such statistics meaningful here.

[3] Please note that the median value and the average shown in the previous page may differ.

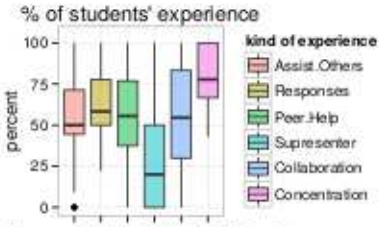


Figure 3: ALL TEACHERS' data

[4] A similarity of 1 means totally uniform experience among all students (quite unlikely), and a value of 0 means totally different experience for every student (also unlikely). For reference, the average similarity in the data from all teachers was 0.69