Developing the FILL+ tool to reliably classify classroom practices using lecture recordings

Journal for STEM Education Research
https://doi.org/10.1007/s41979-020-00047-7

RESEARCH ARTICLE

Developing the FILL+ Tool to Reliably Classify Classroom Practices Using Lecture Recordings

George Kinnear[1] · Steph Smith[2] · Ross Anderson[3] · Thomas Gant[1] · Jill R D MacKay[2] · Pamela Docherty[4] · Susan Rhind[2] · Ross Galloway[3]

Accepted: 6 November 2020Published online: 04 January 2021

Abstract

Lectures are a commonly used teaching method in higher education, but there is significant debate about the relative merits of different classroom practices. Various classroom observation tools have been developed to try to give insight into these practices, beyond the simple dichotomy of "traditional lecturing versus active learning". Here we review of a selection of classroom observation protocols from an ethological perspective and describe how this informed the development of a new protocol, FILL+. We demonstrate that FILL+ can be applied reliably by undergraduate students after minimal training. We analysed a sample of 208 lecture recordings from Mathematics, Physics, and Veterinary Medicine and found a wide variety of classroom practices, e.g. on average lecturers spent 2.1% (±2.6%) of the time asking questions, and 79.3% (±19%) of the lecture talking, but individuals varied considerably. The FILL+ protocol has the potential to be widely used, both in research on effective teaching practices, and in informing discussion of pedagogical approaches within institutions and disciplines.

Keywords Active learning · Classroom observation · Classroom practices · Lecture recording

✉ George Kinnear
G.Kinnear@ed.ac.uk

1 School of Mathematics, University of Edinburgh, Edinburgh, UK

2 Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK

3 School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK

4 Department of Mathematics, Heriot-Watt University, Edinburgh, UK

Introduction

Recent changes to the higher education sector, particularly the continued provision of lecture recording (Gysbers et al. 2011), have renewed debate on the purpose of lecturing as a teaching activity in higher education (MacKay 2019). Lectures are highly prevalent within higher education (Lammers and Murphy 2003) despite being commonly criticised for being ineffective (Gibbs 1981; Freeman et al. 2014), more difficult to access for students from under-represented groups (Leese 2010), and disengaging (Langan and Whitton 2016). Interventions intended to combat some of these issues have shown limited effect size (Huxham 2005).

In its broadest form, a lecture consists of an expert conveying information to non-experts (Trott 1963), although it can also be argued to serve wider purposes, such as offering lecturers the opportunity to model disciplinary practice (McInnes 2013; Pritchard 2010), and providing students an opportunity to integrate feedback into their developing academic identity (Lea and Street 2006). Lectures are undeniably a mainstay of higher education provision (Copeland et al. 2000), but there is considerable debate around how to improve and refine them (Knol et al. 2016). With this ongoing debate, there has been an interest in categorising what goes on inside a lecture, what students do, what staff do, and how this impacts aspects of student learning. For example, Lammers and Murphy (2003) used student observers in live lectures to categorise who was interacting with the material, and found that the majority of lecture time featured the lecturer talking, with the students apparently passively listening. Similarly, a large-scale study of STEM classrooms in North America found that a majority (55%) of the sample were "classrooms in which 80% or more of class time consists of lecturing" (Stains et al. 2018, p. 1469). This type of behavioural categorisation, or ethological study, is one way of exploring what occurs in lectures.

This article makes three novel contributions. First, we use the perspective of ethological study to review a selection of classroom observation protocols ("Review of Classroom Observation Protocols" section). Second, we draw on this review to describe the development process of a new classroom observation protocol, FILL+ ("Method' section'). Third, we show that this new tool can be applied reliably using lecture recordings, and present the results of applying it to a range of lectures at the University of Edinburgh ("Results" section).

Our main goal in analysing these lectures is to establish the reliability of the tool. The analysis that we present is not intended to be comprehensive, and we do not claim to be categorising different teaching approaches. Instead, we aim to highlight different ways that FILL+ could be used to analyse lectures in future studies, as we discuss in the final section ("Discussion" section).

Review of Classroom Observation Protocols

Theoretical Framework

A number of tools have been developed to characterise and analyse classroom practices to give insight into learning attainment (Lund et al. 2015; Smith et al. 2014;

Stains et al. 2018). Many of these use an ethogram (an objective and repeatable description of behaviour, see Martin and Bateson 1993), albeit sometimes unknowingly. There are three important considerations of ethological study that can be used to interrogate the applicability of behavioural classification tools.

The first element is the type of behaviour being measured, which is usually characterised as event vs state behaviours. Events are short duration behaviours best measured in terms of frequency, and states are longer duration behaviours best measured in their duration. Behaviour type is a continuum, and any given behaviour may be considered as a state or an event depending on recording method.

The second element is the recording method, which encompasses both the sampling rule and the recording rule. Recording can be continuous, where all observed behaviours are recorded to produce true frequencies and durations, or time-sampling (sometimes referred to as scan sampling) where behaviour is instantaneously sampled periodically to provide an estimate of the behaviours. Continuous recordings are more likely therefore to pick up short events than time sampled recordings, as events which occur outside the sample windows are not recorded.

The third and final element to consider is the clear and unambiguous description of all behaviours, so each observer can be confident in their categorisation. This is sometimes referred to as the "Martian test": a Martian with no prior experience of the species under observation should still be able to categorise behaviours accurately. These principles of measuring behaviour are designed to produce reproducible and repeatable studies of behaviour, and are the theoretical basis for our work.

There are many tools which aim to characterise classroom behaviour, but we identified three which may be useful for our purpose, and use Martin and Bateson's principles to assess their usefulness.

COPUS

The Classroom Observation Protocol for Undergraduate STEM (COPUS) is intended to let observers rapidly categorise classroom practice (Smith et al. 2013). COPUS was developed with the intent of refining STEM teaching within a Canadian higher education institution and is a modified version of TDOP (Hora and Ferrare 2014). COPUS categorises student and instructor practice into codes, which can be considered as behavioural states. COPUS originally also explored the "cognitive sophistication" of lecturer questions but this was deemed impractical for repeatable coding, an example of the "Martian" test in practice. The COPUS coding form follows a time-sampling methodology in 2-min blocks, with each behaviour being considered a state within that scan. Initial uses of COPUS indicated that students spent the majority of time in lectures listening, and lecturers spent the majority of their time lecturing, with some differences in listening time between classes in the University of British Colombia and the University of Maine. Inter-rater reliability was scored with Cohen's Kappa with generally good agreement between observers. COPUS is a robust tool, and it has been widely used (e.g. Maciejewski 2015; Stains et al. 2018; Semanko and Ladbury 2020). However, the 2-min time sampling methodology may miss short interactions between staff and students, which again are oft-considered important learning opportunities.

PORTAAL

The Practical Observation Rubric To Assess Active Learning (PORTAAL) tool was developed from the "ground up" using evidence from the literature to characterise practices which relate to active learning (Eddy et al. 2015). It features 21 elements spread across four dimensions of best practice in active learning: practice, logic development, accountability, and apprehension reduction. Each element is "supported by at least one published peer-reviewed article that demonstrates its impact on a relevant student outcome" (Eddy et al. 2015, p. 3). PORTAAL reports most behaviours as frequencies, treating individual state behaviours such as "give students practice participating by enforcing participation through cold/random call" as event behaviours, counting the number of these activities, not the duration of them. There are a number of ambiguous phrases which have proved difficult to use effectively for new tool users (Chinnery et al. 2018). However, this is the first tool we explored where the recording method is continuous, theoretically allowing for greater fidelity of behaviour capture. Naive observers using PORTAAL were found to match the expert ratings for 90% of observed states for 19 of the 21 dimensions (Eddy et al. 2015). Similar to COPUS, categorising the cognitive sophistication of questions asked by lecturers proved challenging for observers to score repeatably (Chinnery et al. 2018).

FILL

The Framework for Interactive Learning in Lectures (FILL) was developed by Wood et al. (2016), and aims to explore more nuanced behaviours that may last for less than 2 min. It aims to provide an objective account of observed behaviour, with no claim about the impact these behaviours are having on learning. FILL is influenced by activity theory in learning and therefore includes the contextual aspects of learning also discussed by Hora and Ferrare (2014). It was designed to analyse lectures which were using Peer Instruction cycles, and so focuses on a specific type of classroom practice. FILL was also developed using recordings of lectures, facilitating continuous behaviour recording as the recording could be paused at any point. FILL has 6 codes which can be recorded as states or events given the continuous recording method. Cohen's Kappa was used to describe inter-rater reliability (IRR) between two observers and was calculated at 0.74; however, Wood et al. note that there was some ambiguity regarding how state end points were defined. For example "lecturer question" could include when a lecturer asked a rhetorical question and when a student responded to a question. This variability may influence FILL's ability to be adapted for other disciplines, where we may see more variation in teaching practice, especially where Peer Instruction cycles are less formalised.

Method

The development of a refined classroom observation protocol, FILL+, took place through five main phases, which we describe in detail in the following sections.

Ethical approval for this study was obtained from the Moray House School of Education and Sport Ethics Sub-Committee, Reference 1947.

Phase 1: Selecting a Tool

We decided to focus on FILL and PORTAAL since, unlike COPUS, they can capture detail about activities with short durations. Indeed, a key motivation for the development of FILL was that "coding in 2-min intervals does not give a precise picture of what happens in the classroom, as many activities, particularly the type of interactions (such as questions) that we are interested in for this research, last less than 2 minutes" (Wood et al. 2016, p. 2).

To compare the performance of FILL and PORTAAL, two 50-min lectures each from across four departments (Veterinary Medicine, Agriculture, Mathematics, and Physics) were scored by a single coder (SS) using both tools. The primary objectives were to assess the tools' ease of use, the reliability of the scores obtained, and visualisation and interpretation of results.

Prior to scoring, the coder undertook training with each tool. For PORTAAL, this was based on study of the online manual and training videos, following correspondence with the authors of Eddy et al. (2015). The coder produced answers to 37 items in PORTAAL for each of 14 activities taking place in the training videos; these were in agreement with the authors in 96.1% of cases. For FILL, no training manual was available beyond the details provided by Wood et al. (2016) and instead discussion took place with one of the authors, RG. Early in the coding process, we added a code for "Admin" to the six existing FILL codes; this formalised the decision that "any time spent on administration at the start of the lecture was disregarded in the analysis" (Wood et al. 2016, p. 6), and allowed us to deal with cases where administrative matters arose mid-lecture. Following coding of examples and discussion with RG, when using FILL, the coder was in 100% agreement on the codes used. There were differences in the exact start and end times associated with the codes (with a mean absolute difference of 2 s across all start times). As a further measure of ease of use, the time taken to score a single lecture was recorded. On average, each minute of the lecture took 2 min 30 s to score with PORTAAL, but only 1 min 10 s with FILL.

Approaches to visualisation and interpretation were also compared for the two tools. For FILL, timelines can be produced for each lecture with the various states indicated by different colours (as in Fig. 1a). Summaries can also be produced, showing the proportion of each state in a certain lecture or the average from a set of lectures. For PORTAAL, visualisation of the results was guided by the approach of Eddy et al. (2015), with plots for each of the 21 elements (e.g. elements L3 and L4 are shown in Fig. 1b).

Despite both being continuously sampling tools, FILL and PORTAAL have some important differences. FILL records 6 distinct state behaviours, whereas PORTAAL records 21 state-like behaviours as events. The relatively large number of elements to consider in PORTAAL presents two concerns. First, it is practically difficult for coders to hold all of these elements in mind, and some aspects such as classifying questions according to Bloom's taxonomy require careful thought. This is reflected in the longer time required to score using PORTAAL. A second concern is that some

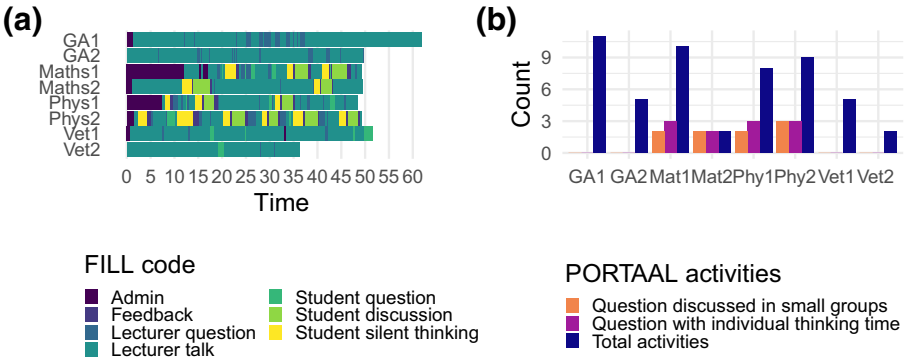


Fig. 1 **a** FILL timelines for each of the 8 lectures, showing the code selected at each time during the lecture. **b** A summary of the PORTAAL activities, with details of when questions were posed with individual thinking time, and with students working in small groups (corresponding to elements L3 and L4 (Eddy et al. 2015, Table 2) in the Logic Development dimension)

elements were little-used or entirely absent; for instance, in our sample, there were no "instances of explicit negative feedback" or cases where the lecturer "explicitly reminds students that errors are natural and useful". This may be a feature of our small sample, but similar redundancies were noted by Eddy et al. (2015, p. 11). In contrast, all six codes in FILL were used at least once in this sample.

Moreover, despite the richness of PORTAAL's elements, they do not account for a number of features that are captured by our implementation of FILL. Most notably, there are no equivalents to FILL's lecturer talk (LT) and administration (AD). In addition, FILL recognises lecturer questions (LQ) even when there is no student response, while PORTAAL disregards these. There could be extensive debate regarding the "learning" that occurs in these types of rhetorical questions, but failing to record them entirely creates a bias within the data. Student thinking time, even when no answer is verbalised, is still an element of teaching and learning. Similarly, FILL records all student questions (SQ) while PORTAAL only scores these if the instructor repeats the question to the class so all can hear.

Most importantly, PORTAAL appeared to be less able to cope with disciplinary differences, in part due to the issues mentioned above. The number of activities recorded by each tool was compared across the eight lectures. In PORTAAL, the number and characterisation of the distinct activities form the basis of the tool. For FILL, there is no equivalent direct metric; however, as start and end times are recorded for each state, the number of distinct states used throughout the lecture can easily be derived. Figure 2 shows the total of these from each tool for each lecture. It is immediately apparent that FILL observes more events than PORTAAL, which may be due to the amount of detail contained within the entity of the "activity", or the fact PORTAAL does not score the lecturer talking (FILL's "LT") or administration (FILL's "AD"). Most importantly, the tools disagreed in some of the lectures (Mathematics and Agricultural Science), demonstrating that disciplinary differences may be observed. FILL was better able to provide more detail about teaching practice in all disciplines compared to PORTAAL.

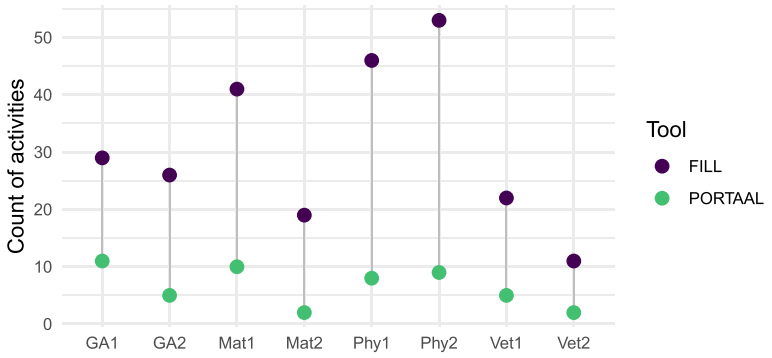


Fig. 2 Number of activities recorded using each tool. For PORTAAL, the activities are counted as part of the scoring process. For FILL, we have treated each change of state as a new activity

Given its relative ease of application and interpretation of results, FILL was our preferred tool. Furthermore, FILL aligns well with the ethology approach and our goal to document the observed behaviour without "the a priori equation of instructional quality with certain teaching methods" (Hora and Ferrare 2014, p. 37). That said, there at least two reasons to expect that FILL data could shed light on important aspects of teaching practice. First, there is consonance between some of the codes in FILL and in PORTAAL, which does draw on "research-based best practices for active learning" (Eddy et al. 2015, p. 2). Second, the approach used to develop FILL was based on characterising interactive engagement practices, since "there is now substantial evidence that these teaching approaches lead to better outcomes compared to traditional methods" (Wood et al. 2016, p. 1).

While FILL is our preferred tool, this pilot work highlighted some issues with applying FILL beyond its original context (physics lectures based around Peer Instruction). We address these in the next section.

Phase 2: Developing FILL+

There were two main issues we identified in applying FILL beyond its original context. We sought to address these by creating a modified tool, FILL+. In doing so, our aims were to minimise the number of new codes in FILL+ (to avoid making it overly complex), and to try to retain "backwards compatibility" with previous FILL codes so that results could be compared with previous work.

The first issue we identified is that FILL is based on two coding schemes (Wood et al. 2016, Tables I and II), one of which is specifically designed with Peer Instruction in mind and is therefore not applicable to other lecturing approaches. To address this, we combined the two coding schemes by augmenting the main set of codes (from Table I) so that they can be used to capture the stages of Peer Instruction (from Table II). In doing this, we introduced the code CQ for "class question", meaning any question directed at the entire class that expects a response from the majority of students. This response may be via an audience response system (e.g. clickers, as considered by Wood et al. 2016), but could also be by show of hands, flash cards, or

otherwise. This enables FILL+ to meaningfully capture whole-class interactivity in lectures, whether or not this is based on Peer Instruction.

The second issue was that we observed alternative modes of question and response in lectures which do not make use of Peer Instruction, and these were not being captured by FILL. With FILL, the LQ and SQ codes cover both the question and the response; with FILL+, we only use LQ and SQ to cover the time when a question is being posed, and we use new codes to cover any student response (SR) or lecturer response (LR). This increased resolution enables more detailed analysis of the way that questions are used in lectures. We also incorporated more "non-verbal" interactions, as suggested by Wood et al. (2016, p. 13), to help distinguish monologic lecturing from a more interactive back-and-forth style. For example, student laughter in response to something the lecturer or another student has said would be coded as SR since the laughter indicates that students heard and processed what was said. Finally, we added a code for administration (AD). Time spent on administration is acknowledged by Wood et al. (2016) but is not recorded by FILL as it was not considered by the authors to be an activity that contributed to learning. Following the same sociocultural approach, we designed FILL+ with a focus on the student experience and felt that administration contributed to this in a manner distinct to that recorded by any of the pre-existing codes.

The resulting set of 10 codes used in FILL+ is summarised in Table 1. Exactly one code is assigned at any given time and activities are coded contiguously. A new code is only introduced when it is clear from the student perspective that a new activity is taking place, e.g. if after a period of lecturer talk (LT), the lecturer pauses to change slide and then asks a class question (CQ), the pause where they were changing slide would be recorded as LT. In instances where multiple codes occur simultaneously, the code used is the activity that the majority of students were experiencing at that time, e.g. if an individual student asked a question to a lecturer during ST then the period where the question was being asked and the lecturer was responding to the individual student would still be recorded as ST. If, however, the lecturer deemed the question important enough to pause the class activity and address everyone, the exchange would be recorded as "SQ" (student question), then "LR" (lecturer response). Further details of the approach to coding are given in the training manual (Smith et al. 2020).

Phase 3: Establishing Reliability with FILL+

Overview

The reliability of FILL+ was assessed over a ten week period in summer 2019 by three coders (SS, RA, and TG), who gathered data from a total of 208 recorded lectures.

An iterative process was used to measure and improve reliability. First, lectures were independently coded. Then an inter-rater reliability metric was calculated and any disagreements in coding were discussed and resolved. Finally, the ethogram was updated to reflect any changes to coding made following the discussion. The coders completed this process three times, albeit with slight differences in the exact method used on each occasion. The three iterations of this process and differences between

Table 1 Summary of FILL+ ethogram

Code	Activity	Description	Interactivity level
AD	Administration	Discussion of non-subject material, such as assessment deadlines.	Non-interactive
LT	Lecturer talk	Lecturer talking to students about subject material with no expected interaction from students.	Non-interactive
LQ	Lecturer question	Lecturer asking a question that expects a response from an individual student.	Vicarious interactive
SR	Student response	Student(s) responding to something the lecturer has done, e.g. responding to an LQ and laughing	Vicarious interactive
SQ	Student question	Student asking a question (prompted or unprompted) pertaining to subject material that expects a response from the lecturer.	Vicarious interactive
LR	Lecturer response	Lecturer responding to input from an individual student.	Vicarious interactive
CQ	Class question	Lecturer posing a question to the entire class that expects a response from the majority of students via some form of audience response system, e.g. clickers, show of hands, and flash cards	Interactive
ST	Student thinking	Students individually thinking about and answering a CQ.	Interactive
SD	Student discussion	Students discussing a CQ or other subject related problem with each other, i.e. student-student interaction.	Interactive
FB	Feedback	Feedback on/discussion of student responses to some activity such as a CQ or a weekly quiz (completed out of class).	Interactive
NA	Not applicable	Code used when none of the previous codes apply and there is no expected student engagement, e.g. when the lecture has not started, or if students are given a break.	–

them are outlined in the "First Reliability Check" to "Final Reliability Check" sections. The need to undergo this process three times was due to FILL+ still being in development during this period. Multiple reliability checks were included to ensure that the three coders understood and were able to apply any changes that were made.

To compare FILL+ timelines from different coders, a conservative approach was taken, with agreement measured on a second by second basis. This encapsulates both differences in code and differences in code duration. In addition to measuring coder agreement, we measured inter-rater reliability (IRR) by calculating Krippendorff's alpha (Krippendorff 2004). With this coefficient, perfect agreement is represented by the maximum value of $\alpha = 1$ and perfect disagreement corresponds to $\alpha = 0$. A widely used threshold for acceptable reliability is $\alpha \geq .800$ (Krippendorff 2004, p. 429).

First Reliability Check

The first reliability check was carried out in the first week. The aim was to ensure that the new coders (RA and TG) were able to code accurately and consistently with respect to the intentions of FILL+'s authors.

To assess reliability, the three coders each scored the same four lectures independently. The lectures were from introductory courses in mathematics and physics, with two lectures from "Introduction to Linear Algebra" and two lectures from "Physics 1A". These results are not included in our final data set, although one of the mathematics lectures was re-coded by one of the coders at a later date and does appear in the final data set. After independent coding, the coders compared the resulting timelines and consistent differences in approach were discussed until agreement was reached. This process was used to develop an ethogram (referred to by the coders as FILL+ v1.0) which was subsequently used as a reference for the coders to independently re-code those lectures with the lowest percentage agreement.

Second Reliability Check

The second reliability check was done 4 weeks later. By this point, using the FILL+ v1.0 ethogram developed after the first reliability check, RA and TG had independently coded 62 lectures, covering 3 distinct courses and 6 lecturers. From this subset, the two coders selected six lectures that they felt were challenging (one each per course). These were then coded by SS. After the percentage agreement was calculated, the two coders discussed and resolved differences in coding.

After this discussion, a significant change to the definition of the "FB" (feedback) code was made. Prior to this change, the feedback code covered both the direct response to the activity that students had completed and, if the activity had correct answers, why some answers were correct and others were not (alternative answers). The change made was to remove the second part of the previous sentence from the definition of feedback. The reason for this was that with the previous definition, lengthy explanations of questions that had been answered would be coded entirely as feedback (an interactive code), when they were for the most part non-interactive. This contradicted a key aim of FILL+, which was to provide an accurate representation of what students were experiencing. In addition, this change made the transition from feedback to lecture talk (an extremely common transition in lectures that use Peer Instruction) far less ambiguous for the coders and it was hoped that inter-rater reliability would improve as a result.

The 6 lectures were then re-coded by all 3 coders using the new ethogram (FILL+ v1.1) along with an additional 2 lectures. This yielded a mean improvement in percentage agreement of 3.5% (87.3 to 90.9%) for the 6 re-coded lectures and a mean percentage agreement of 89.3% across all 8. Due to the conservative nature of the metric and the fact these lectures were chosen to represent the most challenging subset, this was agreed to be an acceptable score to move forward with FILL+ v1.1.

Final Reliability Check

The final reliability check was done approximately 4 weeks after the second reliability check. By this point, coders had re-coded the first 62 lectures in accordance with the FILL+ v1.1 ethogram and had coded a further 121 lectures. For a final measure of inter-rater reliability, approximately 10% of the coded lectures were independently re-coded. This took the form of 45 randomly selected 20-min segments, to provide a snapshot of a typical coding experience and a diverse set of lecturer-course combinations. The segments were chosen by randomising the list of lectures coded by each individual coder and selecting the top 15 from each list. Each chosen lecture was then assigned a random start time between 5 and 25 min allowing for variation in start/end time of a standard 50-min lecture. In a similar way, each coder independently re-coded three randomly selected 20-min segments that they had coded previously, so that intra-rater reliability could be estimated. The results of the final reliability check are discussed in the "Reliability" section.

Phase 4: Applying FILL+ Across Disciplines

Recruitment

In summer 2019, lecturers were invited to participate in the study from the three schools, the School of Physics and Astronomy (to represent the Physics & Astronomy discipline), the School of Mathematics (to represent the Mathematics discipline), and the Royal (Dick) School of Veterinary Studies (to represent the veterinary medicine and agricultural science disciplines). Lecturers were emailed directly by a member of the research team within their school and informed about the project. They were invited to respond to a survey (not reported here) and within that survey they gave their consent for their lectures to be scored. Participants were purposefully sampled to cover a breadth of pedagogies, year groups, and class sizes within the schools. We were also mindful that lecture recording can be an emotive and concerning topic for lecturers (MacKay 2019), so we invited participants who we knew to be comfortable with lecture recording, considering this form of sampling to be least distressing for participants. In addition, as the recordings were being taken from the 2018–2019 academic year, the participants could not change their recordings now they were aware we were observing them, which limits the potential for bias. In total, 51 lecturers were invited to participate across the three schools, and 37 responded. This sample includes three of the authors of this paper, but they were not directly involved in the application of FILL+.

Sampling of Lectures

Across the three schools and 37 lecturers, we coded 208 lectures with the FILL+ tool. Participants were invited based on a specific course they had taught, and in two cases, the same lecturer participated for two different courses, so there were 39 lecturer-course combinations. The number of lectures analysed in each discipline is shown in Table 2.

Table 2 Summary of lectures coded with FILL+, by discipline

Discipline	Lecturers	Lecturer-course combinations	Lectures used in analysis
Mathematics	20	21	98
Physics	9	9	60
Veterinary medicine	8	9	50
Total	37	39	208

For the majority of lecturer-course combinations, exactly four lectures were scored based on the finding that "at least four observations are necessary for reliable characterization of teaching" (Stains et al. 2018, p. 1496). These four lectures were selected at random from all lectures given by that lecturer in the course, excluding lectures in the first and last week of term (since these are often atypical due to introductory administration or revision before exams).

While most lecturer-course combinations had exactly four lectures scored, all lectures in weeks 2–10 of the mathematics course "Introduction to Linear Algebra" (ILA) and the physics course "Physics of Fields and Matter" (PFM) were scored. These two courses were looked at in depth to accumulate additional data for use in further research projects.

Missing Data

We encountered two problems with missing data in recorded lectures. The first problem was that some lecture recordings were audio-only. To check that these could still be coded reliably, we took a previously coded lecture which had both audio and video, and re-coded it using only the audio. The coder then calculated "intra-rater reliability" with a percentage agreement of 99.2% and a Krippendorff's alpha of 0.973. This was agreed to be satisfactory and as a result lecture recordings with complete audio were coded, even if the visual component of the lecture was poor or non-existent.

The second problem was that some recordings frequently had missing sections due to the lecturer pausing recordings, typically during student-student discussions. In cases where one of the selected lecture recordings turned out to be unusable (e.g. missing audio or large sections where the recording had been paused), we substituted another lecture chosen at random. In three cases (L12 and two courses for L20), sections were missing in all lectures so substitution was not possible; the recordings were still used in analysis, as the coders were confident that they could make an accurate estimate on the length of the missing sections and pinpoint the code that was applicable during those sections. A fourth case had large, frequent pauses in the recordings where the duration and the activity taking place regularly could not be identified and so it was not included in the final data set.

Phase 5: Testing FILL+ Training Materials

As a further test of the reproducibility of FILL+ coding, three new undergraduate coders undertook training and re-coded a sample of previously coded lectures. The three new undergraduate coders were mathematics students working on a final-year undergraduate project involving analysis of FILL+ data. An iterative process was used, similar to that described in the "Phase 3: Establishing Reliability with FILL+" section. First, the new coders read the training manual (Smith et al. 2020) and calculated measures of agreement with the established coding of the training videos. This was followed by group discussion of any disagreements. Next, the coders each independently coded a previously coded lecture; measures of agreement were calculated and any disagreements discussed. This was done twice: first with three ILA lectures and then with three PFM lectures. Finally, all three coders independently coded a further PFM lecture, and measures of agreement were calculated.

Results

Reliability

For the set of 45 20-min segments, all three coders supplied FILL+ codes independently (with one set of codes coming from the original data collection). This gave a 3×54000 table of codes, where each row corresponds to 1 s of lecture and each column corresponds to one of the coders. Similarly, each coder re-coded three segments from lectures they had previously coded, giving a table of 2×10800 codes with the columns corresponding to the first and second attempt at coding. Reliability measures for this data were computed using the `irrCAC` R package (Gwet 2019) and are summarised in Table 3.

We found high inter-rater reliability in applying FILL+, with 95.7% agreement and a Krippendorff's alpha value of 0.852 (95% CI (0.847, 0.857)). Furthermore, the distribution of codes is extremely skewed (with over 80% of the time coded as LT) and recent work has found that Krippendorff's alpha is among a class of IRR coefficients that "underreport IR values in skewed frequency distributions" (Quarfoot and Levine 2016, p. 383). Based on this, we also used `irrCAC` to compute Gwet's AC1, which has been proposed as an IRR measure that can cope with skewed distributions (Gwet

Table 3 Summary of reliability measures, both for inter- and intra-rater reliability, using the assigned FILL+ codes

Measure	Percent agreement	Krippendorff's alpha	AC1
Inter-rater	0.957	0.852 (0.847, 0.857)	0.956 (0.954, 0.957)
Intra-rater	0.965	0.849 (0.835, 0.864)	0.965 (0.961, 0.968)

Values for Krippendorff's alpha and AC1 are shown along with 95% confidence intervals

2008). The AC1 value was 0.956 (95% CI (0.954, 0.957)), giving further evidence of high inter-rater reliability.

Intra-rater reliabilty was also high, with 96.5% agreement on FILL+ codes between the two attempts, and similarly high values of Krippendorff's alpha and of AC1 (see Table 3).

The three new undergraduate coders also achieved a high level of inter-rater reliability. On the training videos, their codes agreed with the model answers 88% of the time. After further coding practice, their codes for a PMF lecture agreed with the model answer 93% of the time. Furthermore, their inter-rater reliability at this point was high (Krippendorff's alpha 0.820, with 95% CI (0.806, 0.834)).

Analysis of Classroom Practices

Across the 37 lecture-course combinations, FILL+ characterised a range of classroom practices. The diversity of approaches is visually apparent in the timelines shown in the Appendix, and can also be seen in the variation in the proportion of time spent in each state across lecturer-course combinations, as shown in Fig. 3 and summarised in Table 4. For instance, we see that across our sample of 208 class sessions, lecturers spent an average of 79.3% (± 19%) of the time talking—but the detailed timelines show that this ranged from 42.2% in one lecturer-course combination (Mathematics, C21 L28) to 98.7% in another (Veterinary Medicine, C1 L1).

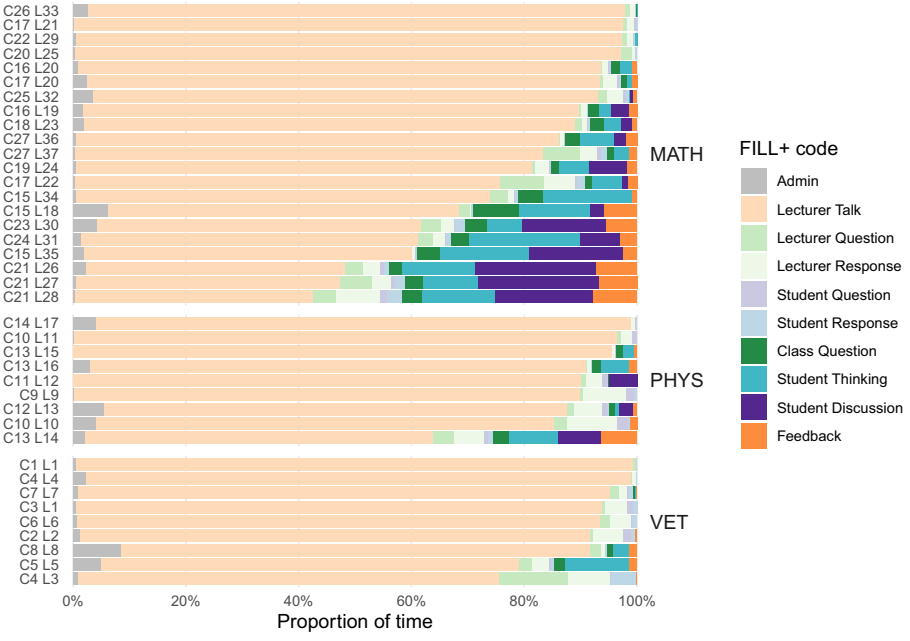


Fig. 3 Proportion of lecture time allocated to each FILL+ code, for each lecturer-course combination. These summarise the detailed timelines shown in the Appendix

Table 4 Mean and standard deviation of the proportion of class time spent in each state, across all lectures in each discipline

FILL+ code	Mathematics	Physics	Vet medicine	Overall
AD	1.5 (± 2.3)	2.1 (± 2.8)	1.8 (± 5.0)	1.8 (± 3.3)
LT	71.9 (±21.2)	81.5 (±15.6)	91.2 (± 9.5)	79.3 (±19.0)
LQ	2.6 (± 2.5)	1.5 (± 1.8)	1.8 (± 3.3)	2.1 (± 2.6)
LR	2.5 (± 3.0)	3.6 (± 4.2)	2.7 (± 3.2)	2.9 (± 3.4)
SQ	0.4 (± 0.5)	0.7 (± 0.8)	0.4 (± 0.7)	0.5 (± 0.7)
SR	0.9 (± 1.2)	0.3 (± 0.5)	0.8 (± 1.2)	0.7 (± 1.0)
CQ	2.3 (± 2.3)	1.4 (± 1.4)	0.2 (± 0.7)	1.6 (± 2.0)
ST	7.1 (± 8.3)	3.7 (± 4.1)	0.8 (± 4.1)	4.6 (± 6.9)
SD	7.7 (±10.0)	2.9 (± 4.9)	0.0 (± 0.0)	4.4 (± 8.0)
FB	3.1 (± 3.2)	2.3 (± 3.2)	0.3 (± 0.7)	2.2 (± 3.0)

The rich detail available in FILL+ data enables deeper analysis of particular classroom practices. One example of this rich detail is found in the periods of lecturer talk (LT). Stains et al. (2018) found that lecturer talk took up "an average of 74.9 ± 27.8% of the total 2-min intervals of a given class" (p. 1469); our headline findings mirror this, but we can go further and consider the precise duration of each turn of LT. Figure 4 shows the distribution of these durations for each course-lecturer combination, along with the mean LT duration for each lecturer. This shows that there is considerable variation in practice, with some lecturers often speaking for more than 10 min at a time, while others do so rarely. Furthermore, this pattern is observed across the three disciplines, and across lectures with different levels of interactivity.

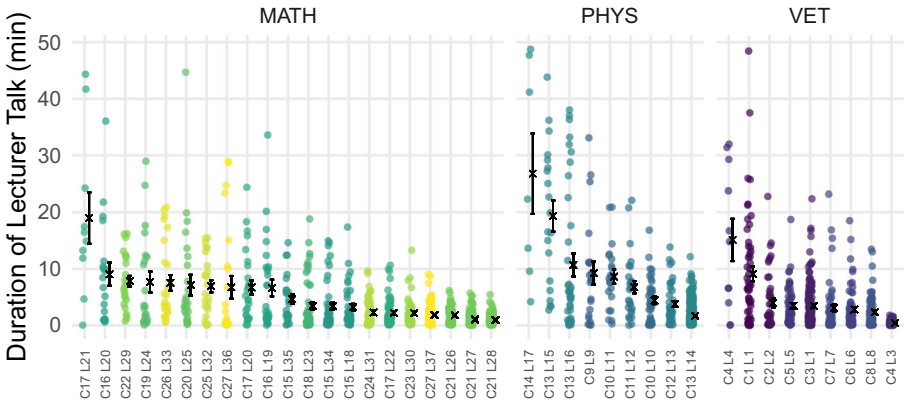


Fig. 4 Distribution of the duration of Lecturer Talk (LT) periods, for each course-lecturer combination. Data points are shown coloured by course, with mean and standard error shown in black. Course-lecturer combinations are ordered by mean duration

A second example of the rich detail in FILL+ data comes from lecturer questions (LQ). Previous work has found large variations in the use of questions by different lecturers (Larson and Matthew 2013; Paoletti et al. 2018), and the same is true in our sample (see Fig. 5). In one particularly striking case (C4 L3), the lecturer typically asks over 100 questions per class session (mostly variations on "is that OK?" with a genuine pause for a response), while the overall mean number of questions per session is 10.7.

Similarly, the data enables an analysis of student questions (SQ). This is a relatively un-studied aspect of classroom practice, though previous work on lecturer questioning did also note student questions, finding a mean of one question asked per 7.5 min (Duell et al. 1992, p. 485). In our sample, we observed a substantially lower rate of one question per 21.7 min (477 student questions across 10,357 min). As with lecturer questions, there was considerable variation in the number of student questions asked in different classes (Fig. 6), which may be due to students' perceptions of the lecturer's openness to student questions (Micari and Calkins 2019; Gasiewski et al. 2012). This does seem to depend on the lecturer in particular; for instance, the mathematics course C27 has two different lecturers and the number of SQs per session is quite different for each lecturer ($m = 1.25, se = 0.48$ for L36 versus $m = 3.75, se = 1.49$ for L37).

Discussion

Our results show that FILL+ is a highly repeatable and reliable tool for characterising classroom practices; with minimal training, novice observers were able to reach high levels of inter-rater reliability (Krippendorff's alpha 0.82). Furthermore, we have demonstrated that the rich detail in FILL+ timelines can be used to give deeper insight into certain teaching practices than existing tools such as PORTAAL and COPUS. In

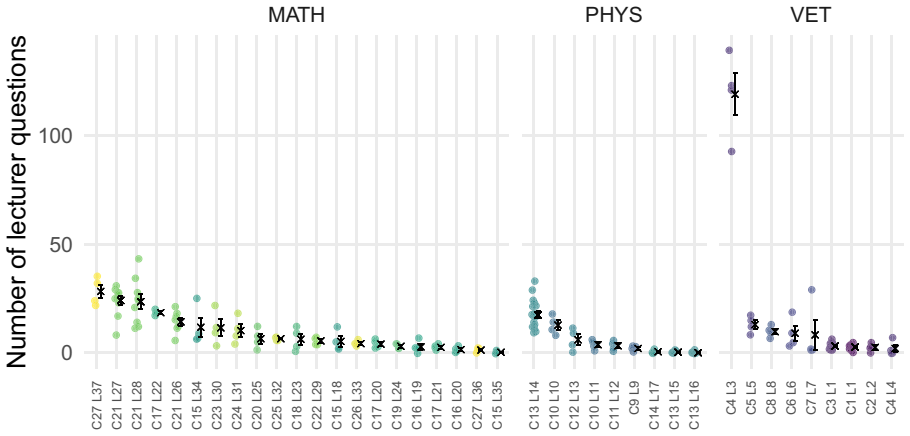


Fig. 5 Distribution of the number of lecturer questions asked per lecture, for each course-lecturer combination. Data points are shown coloured by course, with mean and standard error shown in black

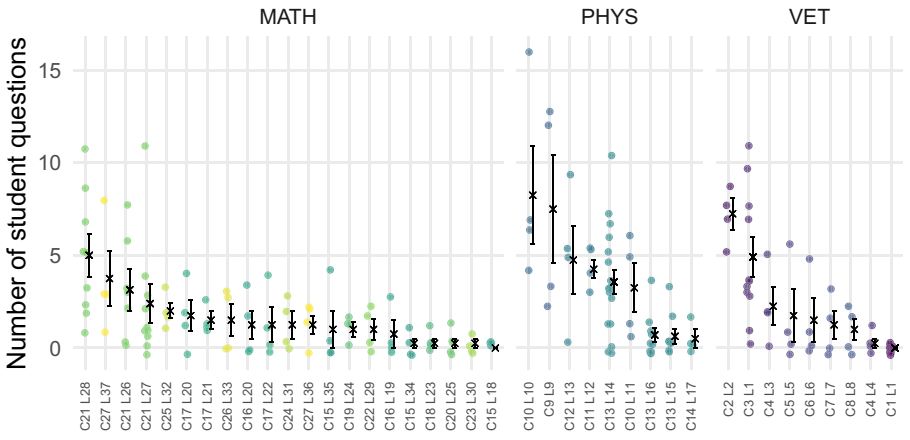


Fig. 6 Distribution of the number of questions asked by students per lecture, for each course-lecturer combination. Data points are shown coloured by course, with mean and standard error shown in black

the remainder of this section, we summarise these insights, discuss limitations of the results, and suggest how FILL+ could be used in further work to analyse teaching practices.

One insight was the variability in the duration of periods of lecturer talk. The mean proportion of time spent on lecturer talk was 79.3% across our sample, but there was considerable variation in the length of individual periods of uninterrupted lecturer talk. This may represent qualitatively different styles of lecturing: turn-based with regular breaks for questions, or largely monologic with other activities happening less frequently. While there is little evidence for claims that students have short attention spans, this is a regular feature of advice to lecturers (Bradbury 2016). Data from FILL+ timelines, showing durations of periods of lecturer talk, could help to investigate this issue by showing the relative prevalence of long periods of lecturer talk and comparing with measures of student attention.

A further insight gained from FILL+ data was the variability in prevalence of questions, both from the lecturer and from students. We found that some lecturers regularly asked more than 20 questions per lecture, while others were observed asking a question only once or twice across all the lectures in our sample. This replicates findings from previous work on the variability of lecturers' questioning practices (Larson and Matthew 2013; Paoletti et al. 2018). These previous studies also investigated the content of questions (e.g. their cognitive level) and wait time; the FILL+ timelines do not address these directly but would speed up a secondary analysis which could be the subject of further work.

Limitations

There are three main limitations on our results, connected with issues of validity. The first limitation is that we purposively sampled lecturers who we felt would be

accepting of being observed. This may have constrained our development of FILL+ to only consider practices of a particular kind, though this does not appear to be the case given the range of profiles observed in our results (see Fig. 3). Furthermore, our non-representative sample means that we cannot draw conclusions about the wider population of lecturers, and in particular our results cannot be used to make comparisons between disciplines. Further use of FILL+ outside of our sphere of influence, and with more systematic samples, will test these concerns.

A second threat to validity comes from our use of lecture recordings. On the one hand this is a strength, as it eliminates the potential for participants to alter their practice during observation, and enables coders to pause and rewind as needed. However, as noted in the "Missing Data" section, some recordings were affected by missing video or audio, and in some cases, sections of the recording were missing due to the lecturer pausing the recording.[1] This may make some proportions of student discussion time an underestimate. The ability to pause recordings is an important aspect of affording the academic control over the learning environment and we would not advocate for its removal for the sake of characterising lecture practices. The creation of a "paused time" behavioural code could allow for recording of this when the total lecture time is known.

The third concern about validity is that we have no external measure of the behaviours scored by FILL+ for comparison with our results. We took an ethology-led approach to the development of the FILL+ ethogram, in particular aiming to pass the "Martian test". The ability of the observers to use the ethogram reliably with little discussion suggests that these behavioural codes are themselves independently recognisable. However, further work is needed to establish the external validity of the codes—for instance, applying FILL+ and other protocols such as PORTAAL and COPUS to the same set of lectures to check convergent validity. Another measure of convergent validity will be provided by forthcoming work which will explore relationships between the FILL+ data presented here, and lecturers' perceived practice as measured by the Teaching Practices Inventory (Wieman and Gilbert 2014).

There is an important caveat when we discuss the validity of these measures, however. This work is simply characterising classroom practices and makes no claim about what practices may be more or less effective (unlike tools such as PORTAAL which aim to identify practices consistent with recommendations from research). The FILL+ profiles (Fig. 3) are not intended to make any qualitative judgement about the experience or learning of those in the lecture. Indeed, Hora and Ferrare's (2014) main critique of a number of classroom taxonomy tools was that they made a priori assumptions about teaching; we believe that the ethology-led approach of FILL+ avoids this. Still, we are conscious of Goodhart's Law as described by Strathern (1997) that "When a measure becomes a target, it ceases to be a good measure", and we caution against any use of FILL+ data in setting targets.

[1] At this institution, lecturers may (and do) pause the recording at will. In policies and materials surrounding lecture recordings, we advise this as a measure to ensure candid discussion, to remove the concern of the recording limiting discussions, (MacKay and Bovill 2020).

Further Work

Instead, we see two possible roles for the use of FILL+ in relation to improving teaching and learning. First, the output from FILL+ could be used as a reflective tool. We have returned to some lecturers with examples of their practice, and they have found this helpful in reflecting on their teaching approach. This approach could be utilised more widely as a method of self-evaluation, perhaps in combination with a questionnaire like the Teaching Practices Inventory (Wieman and Gilbert 2014). This process would need to be handled sensitively, with appropriate guidance on interpretation; academics already feel "monitored" within lecture spaces (Gonzales et al. 2014) and lecture recording may exacerbate this (MacKay 2019).

A second use of FILL+ is in education research. As noted above, one avenue for further work is to explore wider application of FILL+ across different contexts. This could enable comparisons across disciplines, institutions, and even cultures, through a large-scale collaboration similar to that of Stains et al. (2018). Our findings suggest that observers can reach high levels of reliability with minimal training, and coding can eventually be done in almost real time, so FILL+ could be a relatively inexpensive way to achieve this. Furthermore, by providing richer data than similar tools such as COPUS, FILL+ provides a means to "explore which aspects of instructor behaviour are most important for achieving the greatest gains with active learning" (Freeman et al. 2014, p. 8413), by giving a reliable measure of instructor behaviour that can be compared with student outcomes.

Conclusions

In conclusion, we have demonstrated a novel use of lecture recordings, in that they can be used in combination with lecture taxonomy tools to characterise classroom practices. This can enable us to make more informed judgements about what teaching looks like in our institution and informs the discussion of pedagogy within and across our respective STEM disciplines.

Acknowledgements We thank the two anonymous reviewers, whose constructive input helped us to make improvements to the article. We are also grateful to our colleagues for their participation in the research.

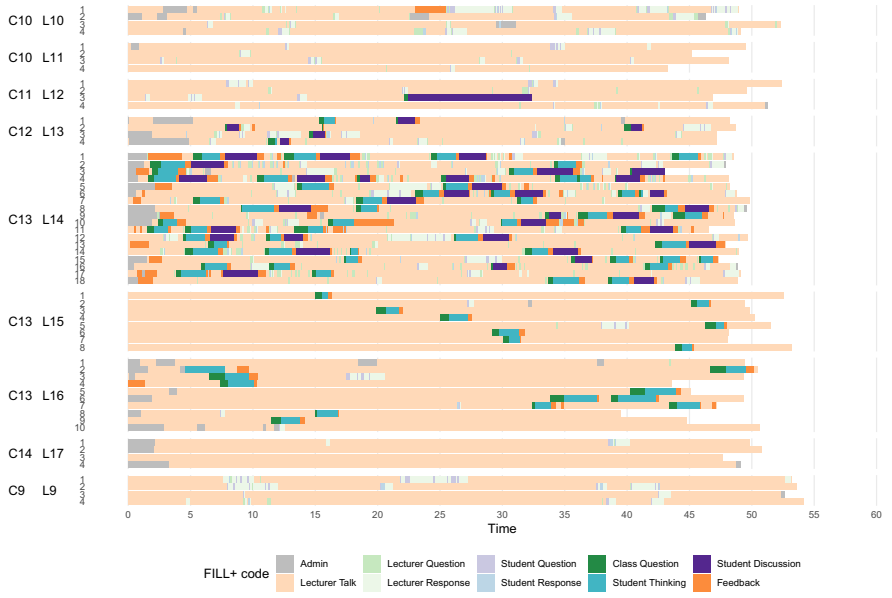
Funding This project was supported with funding from the Principal's Teaching Award Scheme at the University of Edinburgh.

Data Availability The data that support the findings of this study are openly available at https://doi.org/10.17605/OSF.IO/63Z29

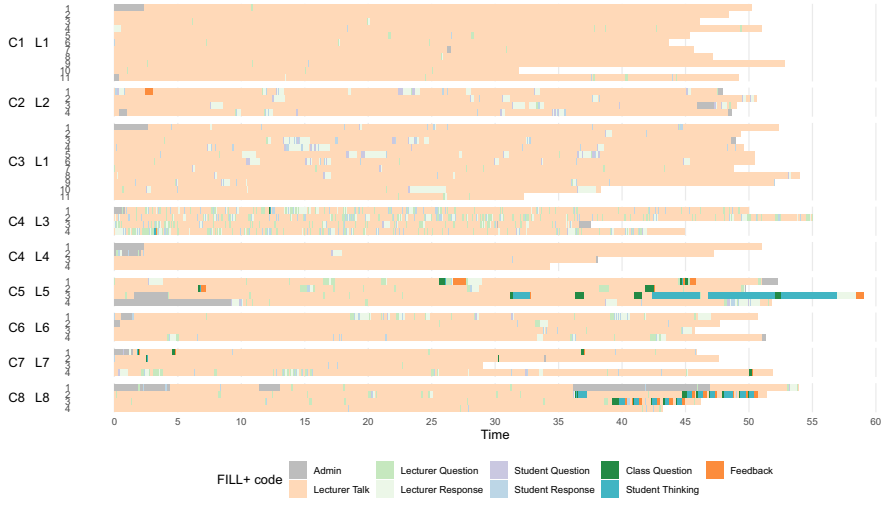
Appendix

Mathematics Lectures

Physics Lectures



Veterinary Medicine Lectures



Note: The lecture C5 L5 session 3 was 2-h long; only the first hour of the timeline is shown here.

References

Bradbury, N.A. (2016). Attention span during lectures: 8 seconds, 10 minutes, or more?. *Advances in Physiology Education*, *40*(4), 509–513. https://doi.org/10.1152/advan.00109.2016. ISSN:15221229.

Chinnery, S., Hughes, K., MacKay, J.R.D. (2018). The active lecture? Exploring engagement in the veterinary lecture through the PORTAAL tool. In *VetEd: international symposium of the veterinary schools council*.

Copeland, H.L. et al. (2000). Successful lecturing. *Journal of General Internal Medicine*, *15*(6), 366–371. ISSN:0884-8734. https://doi.org/10.1046/j.1525-1497.2000.06439.x. http://www.pmc/articles/PMC1495460/?report=abstract.

Duell, O.K. et al. (1992). Wait-time in college classes taken by education majors. *Research in Higher Education*, *33*(4), 483–495. ISSN: 03610365. https://doi.org/10.1007/BF00973768.

Eddy, S.L., Converse, M., Wenderoth, M.P. (2015). PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE–Life Sciences Education, 2*(14). Ed. by Jeff Schinske, ar23. https://doi.org/10.1187/cbe.14-06-0095. http://www.ncbi.nlm.nih.gov/pubmed/26033871.

Freeman, S. et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–5. ISSN:1091-6490. https://doi.org/10.1073/pnas.1319030111.

Gasiewski, J.A. et al. (2012). From gatekeeping to engagement: a multicontextual, mixed method study of student academic engagement in introductory STEM courses. *Research in Higher Education*, *53*(2), 229–261. ISSN:03610365. https://doi.org/10.1007/s11162-011-9247-y.

Gibbs, G. (1981). Twenty terrible reasons for lecturing. ISBN:9780128009598. https://doi.org/10.1016/B978-0-12-800959-8.00021-3.

Gonzales, L.D., Martinez, E., Ordu, C. (2014). Exploring faculty experiences in a striving university through the lens of academic capitalism. *Studies in Higher Education*, *39*(7), 1097–1115. ISSN:1470174X. https://doi.org/10.1080/03075079.2013.777401.

Gwet, K.L. (2019). irrCAC: computing chance-corrected agreement coefficients (CAC). https://cran.r-project.org/package=irrCAC.

Gwet, K.L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. ISSN:00071102. https://doi.org/10.1348/000711006X126600.

Gysbers, V. et al. (2011). Why do students still bother coming to lectures, when everything is available online? *International Journal of Innovation in Science and Mathematics Education*, *19*(2), 20–36. ISSN:2200-4270.

Hora, M., & Ferrare, J. (2014). Remeasuring postsecondary teaching: how singular categories of instruction obscure the multiple dimensions of classroom practice. *Journal of College Science Teaching*, *43*(3), 36–41. ISSN:0047-231X. https://doi.org/10.2505/4/jcst14_043_03_36.

Huxham, M. (2005). Learning in lectures: do interactive windows help? *Active Learning in Higher Education*, *6*(1), 17–31. ISSN:14697874. https://doi.org/10.1177/1469787405049943.

Knol, M.H. et al. (2016). Measuring the quality of university lectures: development and validation of the Instructional Skills Questionnaire (ISQ). *PLoS One*, *11*(2), 1–21. ISSN:19326203. https://doi.org/10.1371/journal.pone.0149163.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. ISSN:0360-3989. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Lammers, W.J., & Murphy, J.J. (2003). A profile of teaching techniques used in the university classroom. *Active Learning in Higher Education*, *3*(1), 54–67. ISSN:1469-7874. https://doi.org/10.1177/1469787402003001005.

Langan, A., & Whitton, N. (2016). Mark understanding learner disengagement: why do students pay 9,000 a year not to attend lectures? In *Learning and teaching in action 11*, (Vol. 2 pp. 56–70).

Larson, L.R., & Matthew, D.L. (2013). Evaluating the efficacy of questioning strategies in lecture-based classroom environments: are we asking the right questions?. *Journal on Excellence in College Teaching, 24*(1).

Lea, M.R., & Street, B.V. (2006). The academic literacies model: theory and applications. *Theory Into Practice*, *45*(4), 366–377. ISSN:00405841. https://doi.org/10.1207/s15430421tip4504.

Leese, M. (2010). Bridging the gap: supporting student transitions into higher education. *Journal of Further and Higher Education*, *34*(2), 239–251. ISSN:0309-877X. https://doi.org/10.1080/03098771003695494.

Lund, T.J. et al. (2015). The best of both worlds: building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE - Life Sciences Education, 14*(2). Ed. by Jennifer Momsen, ar18. https://doi.org/10.1187/cbe.14-10-0168.

Maciejewski, W. (2015). Flipping the calculus classroom: an evaluative study. *Teaching Mathematics and its Applications*, *19*(4), hrv019. ISSN:0268-3679. https://doi.org/10.1093/teamat/hrv019.

MacKay, J.R.D. (2019). Show and tool: how lecture recording transforms staff and student perspectives on lectures in higher education. *Computers and Education*, *140*, 103593. https://doi.org/10.1016/j.compedu.2019.05.019.

MacKay, J.R.D., & Bovill, C. (2020). engagEd in... teaching with lecture recording. https://indd.adobe.com/view/dc75e5a9-903b-40d2-9853-94f193265c14.

Martin, P., & Bateson, P. (1993). *Measuring behaviour. An introductory guide*. Cambridge: Cambridge University Press. ISBN:0 521 44614 7.

McInnes, D. (2013). The performance of academic identity as pedagogical model and guide in/through lecture discourse. *Teaching in Higher Education*, *18*(1), 53–64. ISSN:13562517. https://doi.org/10.1080/13562517.2012.678327.

Micari, M., & Calkins, S. (2019). Is it OK to ask? The impact of instructor openness to questions on student help-seeking and academic outcomes. *Active Learning in Higher Education*. ISSN:1469-7874. https://doi.org/10.1177/1469787419846620.

Paoletti, T. et al. (2018). Teacher questioning and invitations to participate in advanced mathematics lectures. *Educational Studies in Mathematics*, *98*(1), 1–17. ISSN:0013-1954. https://doi.org/10.1007/s10649-018-9807-6.

Pritchard, D. (2010). Where learning starts? A framework for thinking about lectures in university mathematics. *International Journal of Mathematical Education in Science and Technology*, *41*(5), 609–623. ISSN:0020-739X. https://doi.org/10.1080/00207391003605254.

Quarfoot, D., & Levine, R.A. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, *70*(4), 373–384. https://doi.org/10.1080/00031305.2016.1141708.

Semanko, A.M., & Ladbury, J.L. (2020). Using the reasoned action approach to predict active teaching behaviors in college STEM courses. *Journal for STEM Education Research*, 1–16. ISSN:2520-8705. https://doi.org/10.1007/s41979-020-00038-8. http://link.springer.com/10.1007/s41979-020-00038-8.

Smith, M.K., Jones, F.H.M., et al. (2013). The classroom observation protocol for undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. *CBE Life Sciences Education*, *12*(4), 618–27. https://doi.org/10.1187/cbe.13-08-0154.

Smith, M.K., Vinson, E.L., et al. (2014). A campus-wide study of STEM courses: new perspectives on teaching practices and perceptions. *CBE Life Sciences Education*, *13*(4), 624–35. ISSN:1931-7913. https://doi.org/10.1187/cbe.14-06-0108.

Smith, S. et al. (2020). FILL+ training manual. https://doi.org/10.17605/OSF.IO/27863.osf.io/27863.

Stains, M. et al. (2018). Anatomy of STEM teaching in North American universities. *Science (new york, n.y.)*, *359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892.

Strathern, M. (1997). Improving ratings: audit in the British University system. *European Review*, *5*(03), 305–321. ISSN:1062–7987. https://doi.org/10.1017/s1062798700002660.

Trott, J.R. (1963). Lectures, lecturers, and the lectured. *Improving College and University Teaching*, *11*(2), 72–75. ISSN:0019-3089. https://doi.org/10.1080/00193089.1963.10532218.

Wieman, C., & Gilbert, S. (2014). The teaching practices inventory: a new tool for characterizing college and university teaching in mathematics and science, (Vol. 13 pp. 552–69). ISSN:1931-7913. https://doi.org/10.1187/cbe.14-02-0023.

Wood, A.K. et al. (2016). Characterizing interactive engagement activities in a flipped introductory physics class. *Physical Review Physics Education Research*, *12*(1), 010140. ISSN:2469-9896. https://doi.org/10.1103/PhysRevPhysEducRes.12.010140.