

Analysis of Teaching Practice

Pilot outcomes
November 2026

Nuffield Foundation

NATIONAL INSTITUTE of Teaching

Funding acknowledgement

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds and undertakes rigorous research, encourages innovation and supports the use of sound evidence to inform social and economic policy, and improve people's lives. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. This project has been funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily the Foundation.

Find out more at: nuffieldfoundation.org
Bluesky: @nuffieldfoundation.org
LinkedIn: Nuffield Foundation

Contents

Executive summary 4

1. Background 7

2. Objectives 9

3. Data and methods 9

4. Findings 17

5. Discussion 29

6. Limitations 32

7. Conclusion and next steps 33

8. References 34

9. Supplementary materials 37

Executive summary

Background

The Analysis of Teaching Practice (ATP) project is a strand of the Teaching Improvement through Data and Evaluation (TIDE) project, funded by the Nuffield Foundation. ATP aims to find out whether automated feedback on lessons can support the mentoring of early career teachers (ECTs), and which teaching practices are associated with pupils' progress. It brings together a secure tool for ECTs and mentors to record lessons, a dataset of anonymised lesson transcripts labelled for teaching practices, and an automated feedback tool developed from these labels. A trial in 2027–28 will test the impact of AI-supported mentoring on ECTs' practice and pupils' attainment. All of this depends on a recording tool that works in real classrooms, safeguards that protect teachers and pupils, and labels that experts can apply consistently. Before scaling up, we tested these foundations in a pilot.

Objectives

The pilot ran from April to June 2026. It tested acceptability (whether the recording tool could work in real classrooms), safety (how well our safeguards worked in practice) and feasibility (how consistently experts could identify teaching practices from transcripts, and whether hearing or watching the lesson helped).

Data and methods

Four ECT and mentor pairs took part, drawn from one primary school, two secondary schools and one special school. Twenty-eight experienced teacher educators[1] labelled 34 lesson transcripts, using a framework of 29 practices aligned with the Initial Teacher Training and Early Career Framework, in four domains: managing behaviour and high expectations, explanation and modelling, feedback and questioning. Sixteen lessons were labelled again with audio, video or both, so that we could compare agreement with and without them. We measured agreement between labellers from 0 (none) to 1 (complete) and held workshops that explored why they agreed or disagreed.

We deliberately kept labeller training and calibration to a minimum, accepting lower agreement to see where expert judgement differs. Disagreement is useful at this stage: it shows where practice definitions need sharpening, and where experts genuinely see things differently, which shapes what automated tools should try to identify.

[1] Teacher educators are professionals who teach or coach student or practising teachers to support their professional development.

Findings

The recording tool worked in real classrooms. Two-thirds of uploads succeeded, and we have since made uploads of long lessons—the main cause of upload failures—more reliable. Teacher feedback gathered via email throughout the pilot indicated that ECTs and mentors found the tool easy to use. Their feedback also highlighted the benefits of the tool for mentoring. For example, one ECT said: "My mentor used one of those videos for my mentor feedback this week – it's already paying off!" At the same time, mentors asked for additional features: to see transcripts alongside recordings and to be able to tag moments in a recording to support efficient review with their ECT(s) in meetings.

Our safeguards protected teachers and pupils, and only one parent opted their child's class out of the research. Camera positioning kept most pupils out of view. In the main phase, where pupils do appear, their privacy will be protected in other ways: only the ECT and mentor will have access to the recordings, which will be automatically deleted after 30 days. We also identified some areas for improvement. Automated redaction removed most personal data but missed some pupils' first names, so we have added further checks. The opt-out process also slowed set-up and deterred some schools from taking part. We have therefore streamlined the sign-up process to make it easier for schools, while retaining parents' and pupils' right to opt out of research use of transcripts.

Agreement between labellers ranged from 0 to 0.56, and was highest for practices with clear verbal markers, such as *courteous or warm language* or *correction or sanction*. Although every transcript contained errors and pupils' talk was often missing, labellers were able to identify most practices. When listening to recordings, labellers generally found the teacher audible despite background noise, and hearing or seeing the lesson resolved some ambiguities—for example, whether a pupil or teacher was speaking—but did not clearly improve agreement. Some disagreement indicated that definitions will need sharpening before the main phase, while some reflected differing expert views, for example on what counts as *narrating thinking*.

Limitations

The pilot was small, and the sample is unlikely to be representative of schools in England since schools actively opted in to an experimental pilot. Labeller training used synthetic examples of each practice because the labelling platform and real transcripts were not available for the training period. Real classroom transcripts are considerably messier than these synthetic examples, so labellers may have been less prepared for ambiguous cases, which may have lowered agreement. Few lessons were labelled with audio or video, so the lack of a clear effect on agreement should be treated with caution. Finally, agreement measures how consistently labellers

apply the framework, not whether their labels are correct: two labellers can agree with each other and still both apply a label incorrectly.

Discussion

ATP's core components worked in real classrooms. However, consistent labelling will need more than a shared framework: it will need further training, calibration, and agreed examples from real data. Even then, some disagreement may remain, since it can reflect genuine differences in professional judgement, and how far it can, or should, be removed is an open question. Automated tools may be most useful if they identify straightforward signals consistently, so mentors can quickly see where to look in a recording, while leaving judgements about what worked to them. The pilot also highlighted some trade-offs: recording on teachers' own devices makes the tool accessible but lowers transcript accuracy, and safeguards add to schools' workload. Finally, labelling prompted rich professional conversations, suggesting it may have value for professional development.

Conclusion and next steps

We will revise the framework and codebook and test the revisions on a small sample before labelling at scale. In spring 2027, we will use the labelled data to develop and test the automated feedback tool, including across subjects and phases. In parallel, we will prepare for the 2027–28 trial, working with schools on workable safeguards and with ECTs and mentors on how feedback is presented.

1. Background

The Analysis of Teaching Practice (ATP) project is a strand of the Teaching Improvement through Data and Evaluation (TIDE[2]) project, led by the National Institute of Teaching (NIoT) and funded by the Nuffield Foundation. TIDE asks whether we can estimate teachers' impact on pupil attainment—known as teacher value-added (TVA)—and explores the characteristics and practices of higher-TVA teachers, so that innovations can be made in teacher development that raise pupils' attainment.

Research from the United States has found that teachers differ markedly in their impact on attainment (Chetty et al., 2014; Hanushek and Rivkin, 2010), but we know relatively little about what more effective teachers do differently in the classroom. Despite growing national datasets on pupil attainment, deep expertise in the teacher development sector and a rich academic literature on effective pedagogy, there is currently no practical way to link what teachers do in classrooms to what pupils achieve, continuously and at scale.

The largest attempt to link observed practice to teacher impact, the Measures of Effective Teaching (MET) project, scored thousands of lessons using trained observers and several established frameworks. It found that observation scores were associated with TVA, but only modestly. It also found that multiple observations by more than one observer were needed to describe a teacher's practice reliably (Kane and Staiger, 2012). Observation on this scale is costly; as a result, progress in linking what teachers do to their impact on pupils has been limited.

Since then, lesson recordings and transcripts have been shown to contain rich, recoverable signals of teaching quality. Researchers have developed measures of teaching from lesson transcripts (Liu and Cohen, 2021), and trained models to identify specific practices, such as questions without a predetermined answer (Kelly et al., 2018), 'talk moves' that support classroom discussion, such as prompting pupils to explain their reasoning (Suresh et al., 2021), and verbal or material sanctions (Tan and Demszky, 2026). There is also evidence that feedback derived from such analysis can lead to changes to teachers' practice. Randomised controlled trials (RCTs) have shown that automated feedback can increase teachers' uptake of pupils' ideas (Demszky et al., 2024), the share of talk by pupils (Demszky et al., 2023), and teachers' use of questions that focus pupils' thinking (Demszky et al., 2025). Such feedback could also help coaches or mentors quickly find where to look

[2] Teaching improvement through data and evaluation (TIDE) – The National Institute of Teaching

in a recording, and help teachers spot patterns across lessons, classes, or over time (Malamut et al., 2026).

ATP has built a lesson recording tool that allows early career teachers (ECTs) and their mentors to securely record lessons from any device and share them with each other. Where research use is permitted, we will also use the tool to build a large dataset of anonymised lesson transcripts. These transcripts will be labelled for teaching practices aligned with the Initial Teacher Training and Early Career Framework (ITTECF; DfE, 2024) and used to develop and test automated methods for identifying these practices. These methods will underpin an automated feedback tool and will allow us to explore at scale which practices are most strongly associated with pupil attainment. To do this, we will use the NIoT's Teacher Education Dataset (TED), which links pupils' assessment results to the teachers who taught them. A trial in 2027–28 will then test whether the feedback tool changes ECTs' practice and improves their pupils' attainment.

ATP therefore addresses several important gaps. First, almost all of the research described above has been conducted in the United States, mostly in maths and English. ATP works with teachers in England, across a broader range of subjects and phases. Second, combining automated identification of practices with TED could make it possible to study the link between practice and impact at scale and on an ongoing basis, without observing every lesson or collecting new attainment data for each study. Third, we will test a range of automated methods, including large language models (LLMs), against labels grounded in expert teacher judgement and aligned with the ITTECF. For many tools already available to schools, it is not always clear what powers them or how they have been tested. We will share our methods and findings, so that others, including teachers and schools, can scrutinise and build on our work. Lastly, we will evaluate in an RCT whether providing mentors with automated feedback improves ECTs' practice and their pupils' attainment, and whether it has any unintended consequences, such as on the relationship between ECTs and their mentors.

The first year of ATP (2025–26) focused on building the recording tool, its data governance, and a framework of teaching practices; the second (2026–27) will focus on collecting and labelling transcripts and developing automated methods and the feedback tool; and the third (2027–28) will be the trial. Both the feedback tool and the research depend on high-quality labels, which in turn depend on a framework that labellers can apply consistently. Before rolling out the recording tool and labelling at scale, we ran a pilot to test these foundations in practice.

2. Objectives

The pilot ran from April to June 2026. It tested:

1. **Acceptability:** whether the recording tool worked for teachers in real classrooms.
2. **Safety:** how well our data governance and safeguards worked in practice, including automated redaction of personal data and the parent/pupil research opt-out.
3. **Feasibility:** how reliably teacher education experts could identify specific teaching practices from transcripts using our framework, and whether hearing or watching the lesson changed what they identified.

We kept training and calibration to a minimum, accepting lower agreement scores in return for seeing how experts' judgements vary when applying the framework. The resulting variation could reflect ambiguity in what was being judged, imprecise definitions, or real differences in judgement (Aroyo and Welty, 2015; Plank, 2022).

This report sets out what we learned under each of these three objectives, what we have already changed as a result, and what we plan to do next.

3. Data and methods

This study was approved by the NIoT's Ethics Committee (approval number: NIOT-26-2). Detailed methods are available in the supplementary materials (S1 to S9).

The Lesson Recording tool

The tool was co-developed with ECTs and mentors on the NIoT's ECT programme[3]. It is integrated directly into the NIoT's online learning platform for ECTs (PRISM). The tool allows ECTs to securely record their lessons from any device with an internet connection, microphone, and camera (Figure 1). ECTs can record lessons for their own reflection and, if they choose, share them with their mentor.

Mentors can watch their ECT's recordings and write private notes, which they can use to support feedback conversations (Figure 1). They can also record their own lessons to model effective practice, which can be watched and discussed with their ECT in their regular mentor meetings.

[3] In England, new teachers complete a two-year statutory induction as early career teachers (ECTs). During induction, they teach a reduced timetable, follow a training programme based on the ITTECF, and have a designated mentor who provides regular one-to-one mentoring. The NIoT's ECT programme is one such training programme.

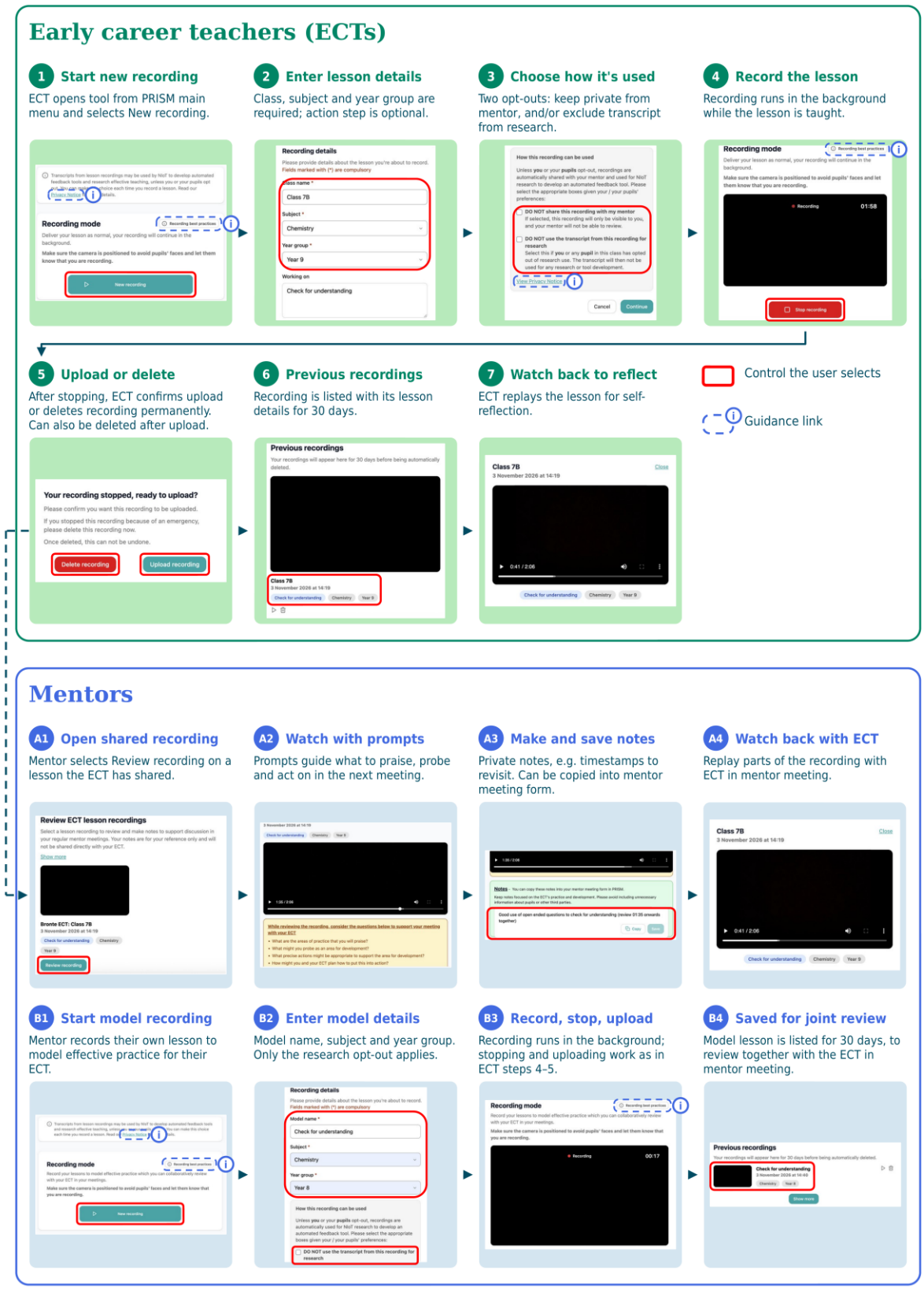


Figure 1. User flows for the Lesson Recording tool for early career teachers (top) and mentors (bottom). ECTs can record their lessons for reflection and choose whether to share each recording with their mentor and whether to allow research use of its transcript (step 3). Mentors

review shared recordings (A1–A4) and can record their own lessons to model effective practice (B1–B4). Red boxes mark controls the user selects; dashed blue boxes mark guidance links.

For the tool to support research, and in future to provide automated feedback at scale, lesson recordings need to be converted into text. Each recording was therefore transcribed automatically. Names and other direct identifiers were removed using automated methods before any transcript was shared with labellers (supplementary materials, S3).

Schools and teachers

We invited schools from the NIoT's founding multi-academy trusts (MATs) and 'Associate Colleges'[4], including schools whose 'super users' had helped design the recording tool. Five schools signed up, and four went on to record lessons between April and June 2026, involving four ECT-mentor pairs (supplementary materials, S1 and S2).

Each ECT and mentor was asked to record up to eight lessons, and each mentor was asked to review at least one of their ECT's recordings. We shared step-by-step guidance, including video demonstrations of the tool for ECTs and mentors. They were invited to share feedback by email throughout the pilot.

Data governance and safeguards

We want to build a research infrastructure that is both scalable and safe. We are committed to 'security by design'—building in safeguards from the outset to protect teachers and pupils.

To that end, we worked with legal experts, data protection specialists, and officials from the Department for Education (DfE) and the Office of the Children's Commissioner for England to produce a comprehensive data governance package. This included a Legitimate Interests Assessment (LIA), a Data Protection Impact Assessment (DPIA), and a data-sharing agreement (DSA) with participating schools. In parallel, we developed materials for school leaders, teachers, parents, and pupils, including information sheets, privacy notices and FAQs. The governance package and supporting materials were reviewed and critiqued by the NIoT's Ethics Committee, our Data Protection Officer (DPO) and the project's Strategic Advisory Board.

This work established distinct lawful bases for the project's two purposes: supporting ECTs' development rests on 'public task', while the research element—including

[4] Associate Colleges are local partner organisations, working with a wide variety of communities across the country, with established expertise in teacher professional development. They work with the NIoT to deliver and contextualise programme training for participants in their local area; typically they work with successful groups of schools within their local areas.

developing and evaluating the automated feedback tool—rests on 'legitimate interests'. Although we did not rely on consent as our legal basis for data processing, schools, teachers, parents and pupils still had different levels of agreement and choice. For the pilot, schools signed the DSA and consented to research taking place in their setting. Individual teachers were also asked for their consent, while parents received a notification explaining that their child's teacher might record some lessons for the pilot, and that they could contact the school if they did not want their child's class to take part in the research. Pupils also received an age-appropriate information sheet. The secondary version explained that pupils could ask their parent or carer to contact the school if they did not want their class to take part.

One of the pilot's goals was to test how the parental opt-out worked in practice: whether schools found the process workable and how many parents opted out. This information helped us understand how far the opt-out might affect research integrity in the main phase. During the pilot, a single opt-out meant that no recordings could take place in that class. This was more conservative than the approach planned for the main phase, because the pilot research used audio and video recordings, whereas main-phase data collection will use transcripts only. In the main phase, an opt-out will apply only to research use of transcripts and will not prevent teachers recording lessons to support ECT development and mentoring.

Further safeguards protected teacher and pupil data at each stage. All data were stored securely within NIoT systems, and access to the tool was protected by two-factor authentication. All recordings were deleted at the end of June 2026. In the main phase, recordings will be automatically deleted after 30 days. Teachers could delete recordings at any time and for any reason, including immediately after recording and before upload, for example if something sensitive, such as a confidential conversation, was accidentally captured. Transcripts used for research had names and other direct identifiers automatically removed. Teachers were also asked to tell pupils when they were recording a lesson and to position the device to avoid recording pupils' faces.

Teaching-practice framework

We developed a framework of 29 teaching practices grouped into four domains (Table 1), starting from the ITTECF and selecting practices that we considered to be both theoretically important and observable in a lesson transcript (supplementary materials, S4). Not all of them were intended as markers of good practice. Most came directly from ITTECF statements and are broadly regarded as such (e.g. *breaking down an explanation*), but others describe activities, such as *direct instruction* or *setting an independent task*.

Each practice was given a short definition written in terms of what the teacher says (for example, "teacher explicitly references a previous lesson and connects it to the current content"). To support consistent labelling, we also developed a codebook which set out the longer definitions, decision rules, examples and non-examples. The short practice definitions are given in Table S1, and an example codebook entry in Table S2.

Table 1. The 29 practices in the pilot framework, by domain

Domain	Practices
Managing behaviour and high expectations	Setting expectations in advance; correction or sanction; positive reinforcement (of behaviour); courteous or warm language; encouraging effort; waiting for silence
Explanation and modelling	Signposting learning; linking to prior learning; direct instruction; anticipating mistakes; narrating thinking; making the abstract concrete; breaking down an explanation; abandoning an explanation
Feedback	Feedback on written work; feedback on an individual's spoken answer; feedback after a whole-class check; setting a task to act on feedback; self or peer assessment; re-teaching; sharing pupil work
Questioning	Scaffolding; thinking time; peer discussion; questions about what is being taught now; questions about earlier content; discovery questions; elaboration questions; setting an independent task

Expert labelling of transcripts

Twenty-eight experienced teacher educators labelled the transcripts (seven labellers in each domain). For each domain, two labellers labelled every transcript independently so that we could calculate agreement. For 16 lessons, labellers labelled the transcript again while listening to the audio (six lessons), watching the video with audio (six lessons), or both in turn (four lessons; supplementary materials, S6 and S7). We call each time a labeller worked through a transcript for one domain a 'pass'. Two independent passes by different labellers on the same recording and domain, using the same input (transcript only, with audio or with video), form a 'pair'.

Labellers were asked to label each transcript before hearing or seeing the lesson, since doing so could change what they noticed in the text. Some labellers worked out of order (e.g. labelling the transcript while listening to the audio before labelling the transcript alone), so the main agreement analysis uses only the 107 of 130 transcript pairs where both labellers labelled the transcript first. We repeated the

analysis with all 130 pairs to check that this did not change the overall pattern of results (supplementary materials, S8 and S11).

All labellers were NIoT staff and experienced teachers. They had taught for a median of 18 years (range 4 to 35). All had observed at least 20 lessons in their career, and 18 had observed more than 200. Labellers were treated as experts applying shared definitions with professional judgement. Disagreements between them were especially valuable at this stage, since they showed where definitions needed sharpening and which practices might need to be dropped or replaced.

We labelled 34 recordings from primary, secondary and special school settings, covering several subjects (Table 2).

Table 2. Labelled recordings by subject and setting

Subject	Primary	Secondary	Special school	Total
Science	1	9	0	10
English	8	0	1	9
Modern Foreign Languages	0	7	0	7
Maths	4	0	0	4
Other subjects (each fewer than 3)	4	0	0	4
Total	17	16	1	34

Labellers were asked to read and label the transcripts in a secure platform (Label Studio; see example in Figure 2), focusing only on the teaching practices in their allocated domain. Their task was to identify whether a practice was present in a specific part of the transcript, not to judge how well it was done or to form a view of the lesson as a whole. Accordingly, they highlighted each section of the transcript where a codebook definition applied and assigned the matching practice, labelling only observable evidence rather than assumed intent. As well as labelling teaching practices, all labellers were also asked to label any sections which they thought contained transcription errors or personal data.

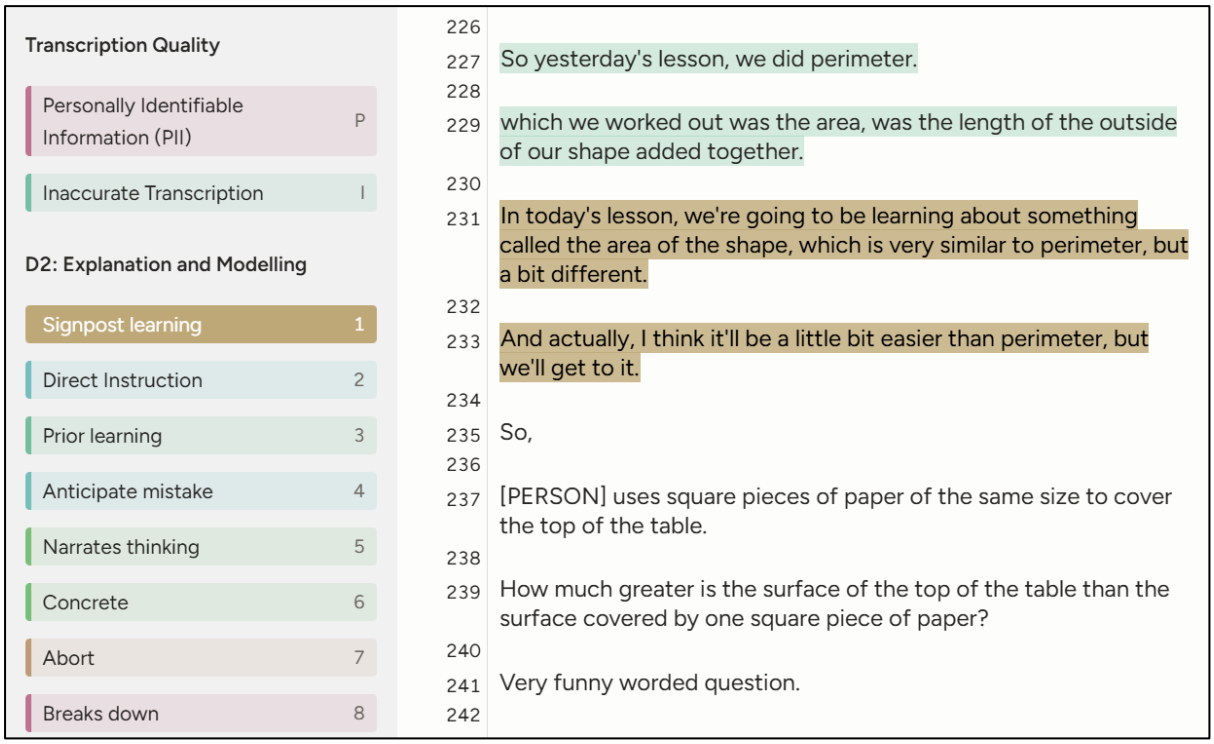


Figure 2. Labelling a transcript in Label Studio. Labellers highlighted each part of the transcript (right) where a practice from their domain applied, and chose the matching practice from the list (left). This example shows labels from the explanation and modelling domain. Labels for transcription errors and personal information were available in every domain.

Measuring agreement

We measured agreement in several ways, since each measure points to a different kind of change the framework may need, and because where and how labellers disagree is as informative as how often they agree (James, 2026). To put agreement in context, we first counted how often each practice was labelled.

Did labellers both notice the practice?

For each lesson, we asked whether both labellers had marked each practice at least once. We refer to this as 'presence agreement': when at least one labeller marked a practice in a lesson, how often did the other labeller mark it in that lesson too? We report this as a percentage of those lessons. Presence agreement tells us whether labellers recognised that a practice was happening somewhere in the lesson, even if they did not mark the same moments as evidence of it. Lessons where neither labeller marked a practice are also a kind of agreement, but we leave them out, because including them inflates the agreement score when the practice is rare.

Did labellers mark the same moments?

Our main agreement measure was the F1 score (Hripcsak and Rothschild, 2005), which looks at individual moments within each lesson. F1 tells us whether labellers share the same interpretation of a practice, and so highlight the same parts of the

transcript as evidence of it. F1 ranges from 0 (no agreement) to 1 (complete agreement).

We counted two labels as matching when they overlapped, even partly, and named the same practice. Because partial overlaps count, we also measured how much text matching labels shared, from 0% (none) to 100% (identical), and report the median for each practice (supplementary materials, S9.2). A low median overlap means labellers identified the same moment but drew its boundaries differently, for example one marking a single sentence and the other a longer passage around it.

Unlike other agreement measures (e.g. Cohen's kappa[5]), F1 has no accepted thresholds for what counts as "good" agreement in a task like this (James, 2026), so we focus on comparing practices with each other rather than judging each score against a fixed standard. We also calculated 95% confidence intervals to show how precise each estimate is (supplementary materials, S9.3).

Did labellers identify different practices for the same moments?

We counted how often two labellers marked the same moment, but with different practices, since this points to practices whose definitions overlapped (supplementary materials, S9.2 and S9.4).

Audio and video comparisons

For the 16 lessons labelled again with audio or video, we compared each labeller's labels with and without audio or video, to see how far hearing or seeing the lesson changed what they marked. We also compared agreement between labellers—with and without audio or video—on the same recordings.

Labeller workshops

After labelling, we held a workshop for the labellers in each domain to gather reflections on the task and framework, and to explore in depth where and why they agreed or disagreed on two or three 'spotlight' practices, to inform codebook revision for the main phase. For each spotlight practice, labellers discussed real examples of agreement and disagreement from the pilot until they reached consensus on whether each should count.

[5] Cohen's kappa is a commonly used statistic for measuring inter-rater reliability (sometimes known as 'inter-annotator' agreement) for qualitative or categorical data.

4. Findings

Acceptability: did the recording tool work for teachers in real classrooms?

Most lessons uploaded successfully. Seven teachers attempted to upload 60 recordings, and 40 (67%) were successful (supplementary materials, S2). Most of the failed uploads were longer lessons (50 minutes or more) recorded in the Safari browser, and we have since made uploads of longer lessons more reliable across browsers. Once lessons were uploaded, ECTs and mentors reported that they found the tool easy to use and the guidance clear. Teachers recorded and reviewed lessons using the guidance provided, without additional support.

Their feedback also highlighted the benefits of the tool for mentoring. For example, one ECT said: "My mentor used one of those videos for my mentor feedback this week – it's already paying off!" The same ECT also valued their mentor seeing lessons they would not usually observe: "It's also been nice having my mentor be able to see some of my lessons that they usually wouldn't be able to."

Teachers also suggested ways the tool could better support lesson review and mentoring, and we have already made some improvements in response. For example, we have added primary subjects to the subject list and changed the way videos load so that it is easier to move around a recording during playback.

Other suggestions are beyond the current scope but will inform future development. For example, mentors asked to see lesson transcripts alongside recordings, to help them locate relevant parts of a lesson and to support scripting during mentoring. Transcripts are currently used only for research and are not shown to teachers, but we plan to make them available, at the latest when automated feedback is added. They also said that being able to tag or pin specific points in a recording, such as areas of strength or questions for discussion, would help make mentoring meetings both high-quality and efficient.

Safety: did our processes protect teachers and pupils?

Parent and pupil research opt-out mechanism

We found that the opt-out mechanism was workable. Four of the five schools that signed the consent form and data-sharing agreement managed the process and went on to record lessons. However, it introduced an additional administrative barrier for schools. The extra steps slowed set-up for some schools, meant some schools declined participation altogether, and delayed data collection. The fifth school did not share the parent and pupil communications or collect opt-outs, so it did not go on to record any lessons. At the same time, opt-outs themselves were rare. Across the participating schools, only one parent opted out. This suggests that

the main cost of the opt-out lies in the administration it asks of schools and teachers, not in parents withdrawing their children's lessons from research.

For the main phase, we have tried to reduce this burden. Induction tutors will enable the tool through a guided sign-up process in PRISM, supported by FAQ-style guidance and a video walk-through. A parental opt-out in the main phase will prevent transcripts from that class being used for research, although lessons can still be recorded for professional development. The school will inform the relevant ECT or mentor of any opt-outs, who will select the research opt-out option in PRISM when starting a recording in that class (Figure 3).

Even so, we recognise that the opt-out will still affect research integrity: the sample will be smaller, and may underrepresent schools with less capacity to manage the extra steps. It is important to note that the active opt-in required from schools and its associated administration will also be an earlier barrier. Schools with concerns about recording lessons may decide not to take part at all.

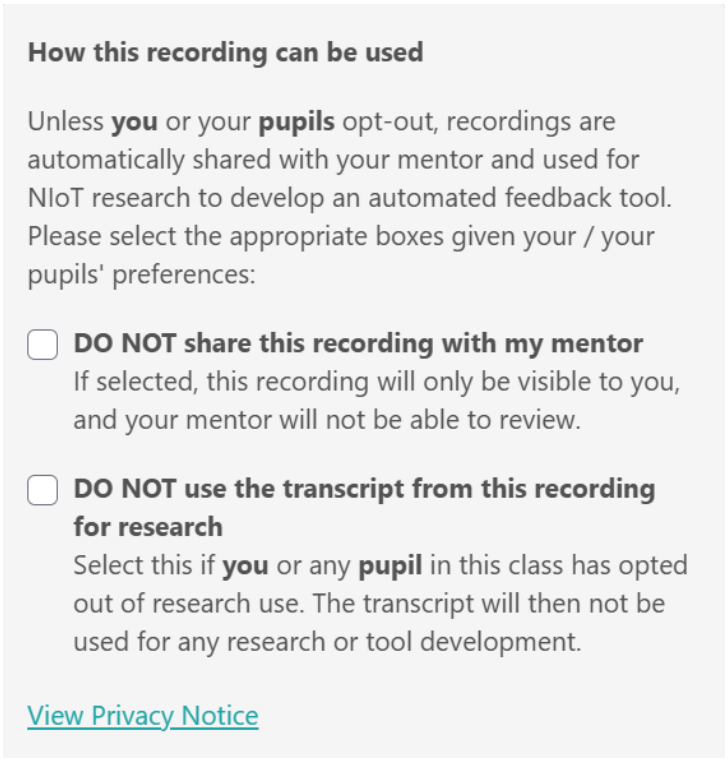


Figure 3. The research opt-out in PRISM. When starting a recording in a class where a parent has contacted the school to opt their child out of research—whether at the parent's or pupil's request—the teacher selects the option not to use the transcript for research. The tool refers to this as a pupil opt-out for simplicity, because the opt-out applies to the pupil's class.

Keeping pupils off camera

To protect children's privacy, teachers were asked to position the device so that pupils' faces were not captured. The pilot showed that this safeguard was largely

effective but could not rule out pupil capture completely. Teachers and labellers both reported pupils sometimes coming into view, for example when walking into the classroom.

Accordingly, the main phase will not rely on camera positioning alone, but will include safeguards that limit who can see recordings and for how long. Research will use transcripts only, so no researcher will ever see pupil images. The recordings themselves will only be accessible to the ECT and their mentor, and will be deleted automatically after 30 days. As a result, any pupil inadvertently captured on video will be seen by a small number of people who already teach or support that class, and only for a limited time to support ECT development and mentoring.

Removing personal data from transcripts

We aimed to anonymise transcripts automatically, using existing language-processing tools to remove direct identifiers such as names and email addresses (supplementary materials, S3). To test how well this worked on the pilot transcripts, we used two checks. First, labellers flagged any personal information they found while labelling (see Figure 2). Second, we carried out a structured re-identification risk assessment on a random subset of transcripts, using the Information Commissioner's Office 'motivated intruder' test. This test asks whether someone determined to identify a person could do so from the redacted transcript.

Both checks showed that the remaining risk came mainly from pupils' first names that the automated process missed. Labellers flagged at least one instance of residual personal data in 31 of 34 transcripts (supplementary materials, S10.3). The risk assessment confirmed that the automated process removed the majority of identifiers, but found that some pupils' first names had been missed in six of the eight transcripts assessed. These were often unusual names, or names that are also everyday words. One transcript also included a school's name. Without these names, the transcripts did not contain enough context to identify anyone.

In response, we have added three safeguards for the main phase. First, manual quality assurance will continue: labellers will mark any remaining personal information, which will then be automatically redacted each week. Second, transcripts will be checked against a list of all school names, which catches school names more reliably than general organisation detection. Third, main-phase data will undergo a further re-identification risk assessment before it is classed as "anonymised", and before it is used to develop automated feedback tools or shared more widely.

Feasibility: could experts label teaching practices reliably?

Some practices appeared many times in every lesson, while others were rarely seen at all (Figure 4; supplementary materials, Table S4). The two most common were *questions about what is being taught now* and *courteous or warm language.* On average, each labeller marked each of these around 29 times per lesson, and *direct instruction* around 20 times. At the other extreme, *sharing pupil work* was marked only five times across all lessons and labellers combined, and five further practices were marked less than once per lesson on average: *abandoning an explanation, thinking time, self or peer assessment, re-teaching,* and *setting a task to act on feedback*. This matters for the agreement results that follow, because the rarest practices offer few examples to compare, so agreement on them rests on less evidence.

These counts also depend partly on how labellers split their highlighting into separate labels. When we merged back-to-back labels for the same practice into a single label, counts fell by around a tenth (Figure 4), indicating that labellers sometimes split one continuous episode into several labels. For the main phase, we will therefore include more specific instructions on when to label a continuous episode once and when to split it, with worked examples from the pilot.

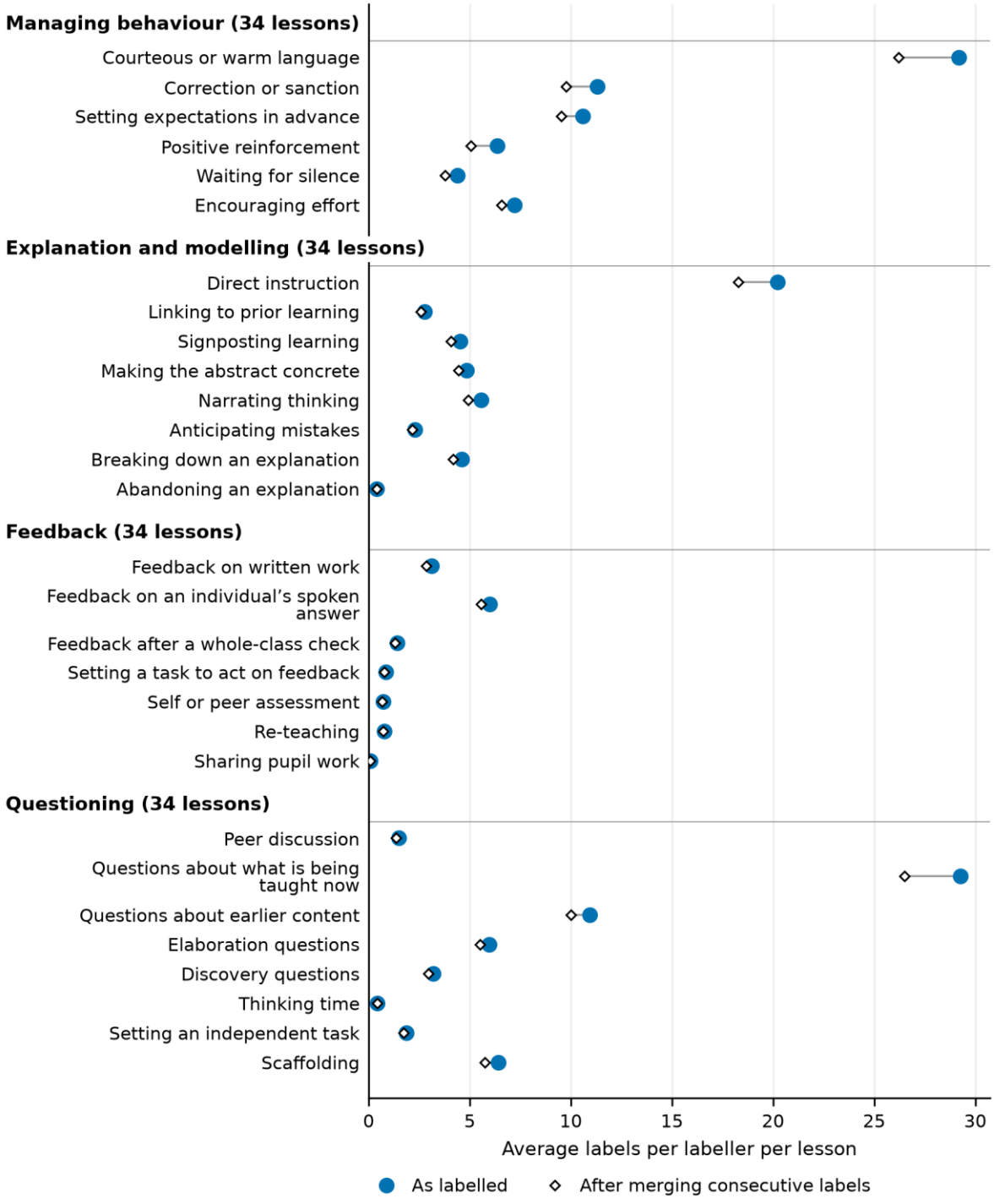


Figure 4. How often each teaching practice was labelled. Average number of practice labels per labeller per lesson, as labelled (filled circle) and after joining consecutive labels for the same practice (open diamond). Where the two coincide, joining labels made no difference. Results based on all 266 transcript passes, including those completed after the labeller had heard or seen the lesson.

Agreement varied between practices, from 0 to 0.56 (Figure 5; supplementary materials, S10.2). It was highest for two practices in the managing behaviour domain: *courteous or warm language* (0.56) and *correction or sanction* (0.51). For example, both labellers marked "[PERSON], turn that off, please" as a correction. *Peer discussion* came next (0.47), though this score is less certain, since it rests on far fewer labels, followed by *direct instruction* (0.44) and *questions about what is being taught now* (0.43). Within domains, practices that teachers tend to announce explicitly, such as *linking to prior learning* (0.38) or *signposting learning* (0.37), agreed better than those that call for interpretation, such as *narrating thinking* (0.23), *scaffolding* (0.11) or *breaking down an explanation* (0.06).

Low agreement showed two contrasting patterns. Some practices were labelled often but rarely in the same place: *scaffolding*, for example, was marked by at least one labeller in 25 of 26 lessons but scored only 0.11. Since these practices were common, this suggests a lack of consensus on which moments count as evidence for them. Others were simply too rare to judge. For example, *sharing pupil work* scored 0 and had only five labels across all main-analysis pairs.

The feedback domain had the lowest agreement overall, with every practice scoring below 0.3. The highest were *feedback on written work* (0.27), on an *individual's spoken answer* (0.22) and after a *whole-class check* (0.17). However, the wide confidence intervals for *written work* and *whole-class check* show that these scores depended on which transcripts were included in the analysis. Labellers also often disagreed about whether feedback happened at all. For all but one feedback practice, more lessons were marked by one labeller than by both (left-hand side of Figure 5).

Outside feedback, however, labellers usually agreed that a practice occurred, even when they marked different moments as evidence for it. For the most common practices, if one labeller marked the practice at some point in a lesson, the other almost always did too. Presence agreement was 93% to 100% for *courteous or warm language*, *correction or sanction*, *setting expectations in advance*, *direct instruction*, and *questions about what is being taught now* (Figure 5). Yet agreement on which sections to mark was only 0.35 to 0.56 for the same practices, suggesting that each labeller often marked sections the other did not.

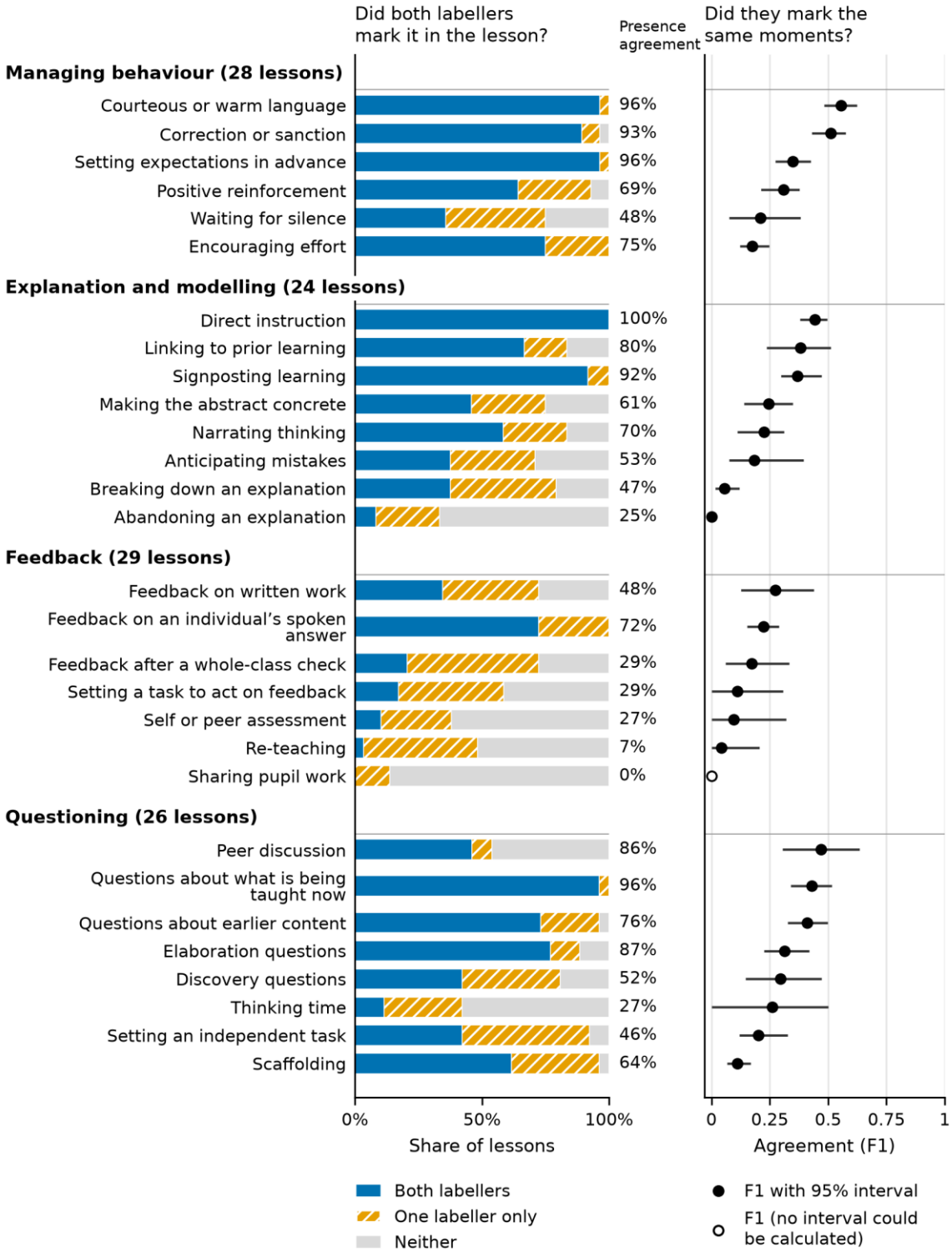


Figure 5. Agreement between labellers for each teaching practice. Left: share of lessons in which both labellers (blue), one labeller (orange) or neither (grey) marked the practice, with presence agreement (the percentage of lessons where both labellers marked the practice at some point, out of all the lessons where the practice was marked). Right: agreement on which moments were marked (micro-averaged F1, from 0 to 1), with 95% confidence intervals. Results based only on transcripts labelled before either labeller had heard or seen the lesson (107 pairs, 24 to 29 lessons per domain).

When labellers did agree on a moment, their labels mostly covered the same text. For 18 of the 27 practices with agreeing labels, the median overlap between agreeing labels was 80% or more. Exceptions included *setting an independent task* (45%), *encouraging effort* (54%) and *linking to prior learning* (58%; supplementary materials, S10.2).

Labellers also often chose different practices for the same moment (Figure 6). This happened most often in questioning. For example, where one labeller chose *questions about what is being taught now*, the other chose *questions about earlier content* 138 times. Labellers rarely split in this way on feedback practices, partly because those practices were rarely labelled. Although the larger counts in Figure 6 partly reflect how often these practices occur, each frequent pair calls for a decision when we revise the codebook: either sharpen the boundary and make clear when both labels should apply, or, where the distinction cannot be drawn reliably, merge the two practices or drop one.

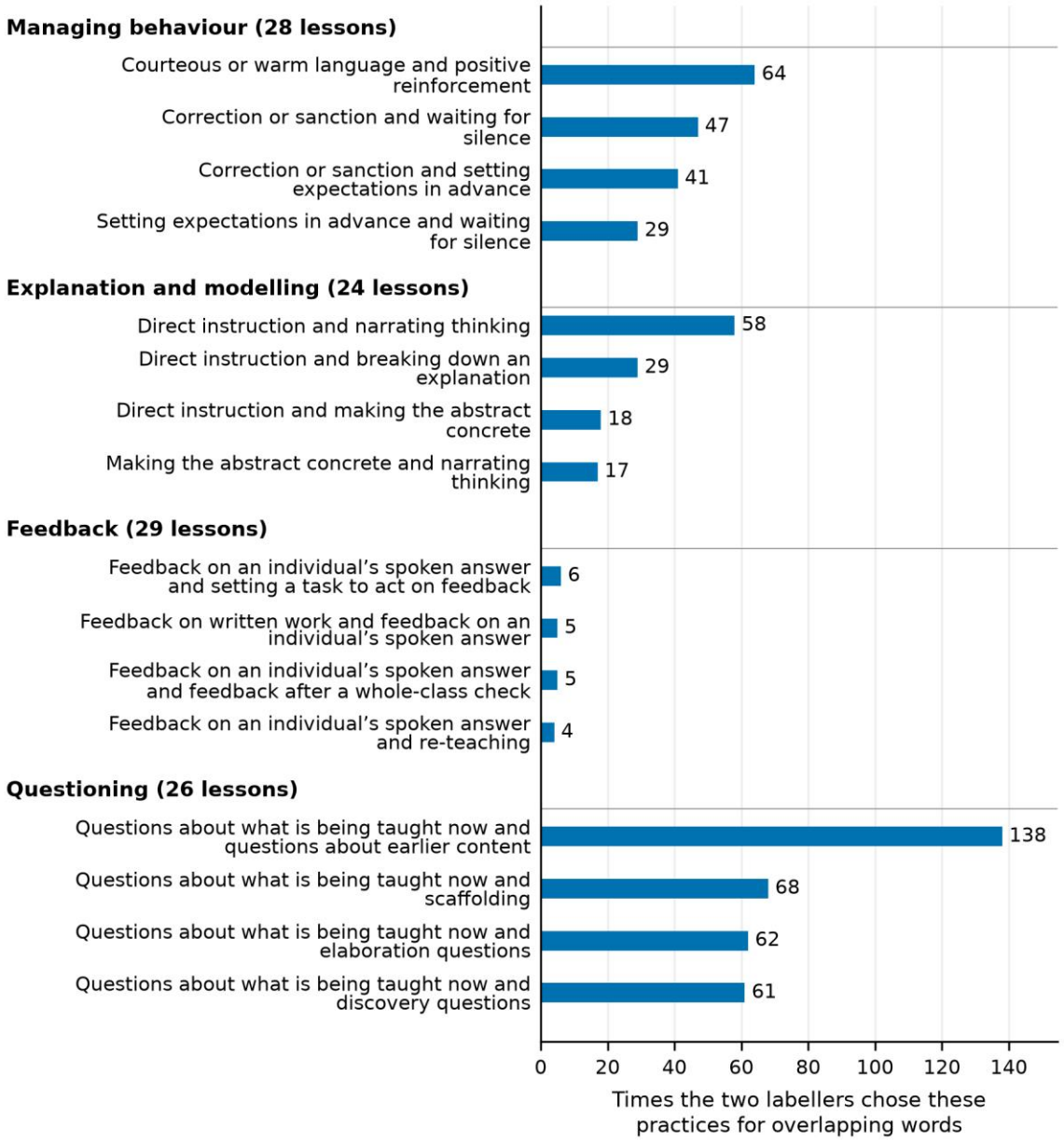


Figure 6. Practices labellers chose between for the same moment. The four pairs of practices in each domain that labellers most often split on: times one labeller marked one practice and the other labeller the other practice, for overlapping words. Main analysis only (107 pairs). Note: larger counts partly reflect how often the practices occur.

The patterns described in this section held up under further checks. Agreement scores changed little when we joined labels that had been split into pieces, removed obvious labelling slips, or included transcripts labelled after hearing or seeing the lesson. Some scores did shift when one labeller—or for rarer practices, one lesson—was left out, so they partly reflect this particular group of labellers and lessons. Full results for these checks are in the supplementary materials (S11, Table S9).

Labeller workshops

The workshops held with each domain's labellers helped to explain some of these patterns and began the revision of the codebook. The spotlight discussions suggested that, for some practices, consensus should be achievable with further calibration, which was limited in the pilot. For others, however, even extended discussion did not bring labellers into agreement, suggesting that consensus may be out of reach.

Labellers felt, for example, that *direct instruction* was under-labelled precisely because it happens so often, yet it was easy to reach consensus when looking at one example at a time. Discussion of this practice also addressed an open question: whether a teacher repeating, elaborating on or consolidating a pupil's answer should count as *direct instruction*. The consensus was that it should, as long as the teacher states knowledge rather than simply repeating a one-word answer.

Other practices proved harder. For instance, labellers found it difficult to separate behavioural expectations from procedural ones. They also held differing views of what 'counts' as metacognition and felt that whether a teacher had broken down an explanation was subjective and difficult to judge at the level of individual utterances. *Abandoning an explanation* raised a different kind of concern. Labellers felt that feedback on this practice could land badly with ECTs, when it often reflects a teacher responding sensibly to confused or disengaged pupils.

Reflecting on the task overall, labellers found it difficult but valuable. Compared with the synthetic training examples, real classroom data were noticeably 'messier', containing transcription errors and many more ambiguous cases. Moreover, labellers agreed that applying a framework consistently was difficult, even for the same person over time. This within-person variation is well documented. Trained classroom observers' judgements drift as they learn an instrument and as fatigue sets in (Casabianca et al., 2015), and annotators who re-label the same material often label around a quarter of it differently (Abercrombie et al., 2025).

Even so, labellers found the work interesting and were excited about its potential. Since labellers were experienced teachers and teacher educators, this endorsement is a valuable early sense check. As one put it: "We are on to something really great here through this work." Labelling also prompted immediate insights, for example on the value of different types of questions: "when people ask elaboration or good discovery questions it's gold dust." Labellers also saw potential for this work to make feedback less vague, by grounding it in specific moments from an ECT's lessons.

What are the limitations of transcripts?

A key question for the pilot was whether transcripts alone capture enough of a lesson to identify teaching practices. We explored this in three ways. First, we

looked at the quality of the transcripts themselves. Second, we drew on labellers' feedback about working from transcripts. Third, labellers labelled 16 lessons a second time with the audio and/or video, after first labelling from the transcript only. This allowed us to see how far each labeller changed their own labels when given more information, and whether two labellers agreed more closely with each other. These comparisons are exploratory: the sample was small, and because the audio or video pass always came second, any changes may reflect a second look and not the audio or video itself.

Starting with quality, every transcript contained errors. Labellers flagged a median of 18 inaccurate sections per lesson (range 6 to 41; S9.5 and S10.3), and about a third of these were flagged by more than one labeller. For example, "a whiteboard pen" was transcribed as "one the white of 10", and "date and" as "dating". As expected, labellers spotted more errors when listening to or watching the recording than when reading the transcript alone.

Labeller feedback highlighted that modern foreign language lessons were particularly affected, because automated transcription assumes speakers are using English. One option for future development is to use 'continuous language detection', which allows transcription to follow changes between languages. However, the service we currently use cannot detect switches within a sentence, which are likely to be common in these lessons. We will therefore evaluate how well the automated methods perform across subjects, including modern foreign languages, before the tool is rolled out.

Speakers were also not reliably identified. In principle, the transcript should label each change of speaker ('Person 1', 'Person 2'). In practice, it typically detected only two speakers per lesson, and 14 of the 34 lessons were transcribed as a single speaker. As a result, labellers, and eventually the automated tool, may incorrectly identify a pupil's contribution as teacher speech. This highlights the importance of keeping humans in the loop. In the trial, teachers will see the recording alongside AI insights and can therefore set aside errors like this. Even so, later phases will need to test whether the rate of such errors is acceptable to users, given that perceived accuracy is a key barrier to teachers using automated feedback (Jacobs et al., 2022; Demszky et al., 2025; Malamut et al., 2026).

Related to this, pupil talk was often missing. A teacher's question might be clear, for example, but the transcript would then jump straight to teacher feedback such as "that's right", with nothing in between. We anticipated this from prior work (e.g. Demszky et al., 2025) and took it into account when developing the framework. For example, we did not include revoicing, or restating what a pupil said, a focus of earlier work (Suresh et al., 2021), since it cannot be identified without knowing what

the pupil said. This remains a limitation of an approach that prioritises accessible, scalable recording on devices teachers already have.

Despite these problems, labeller feedback suggested it was still largely possible to identify practices. For example, one labeller in the questioning domain said "the transcript is all that's needed for the questioning domain. I got 98% of it from it rather than audio or video". This is consistent with recent research suggesting that AI models can also identify teaching practices from imperfect transcripts, provided key words are captured accurately (Wang and Chen, 2024). Labellers did, however, describe the difficulty of not making assumptions beyond the text in front of them, and some felt it was "easier to watch", since that is how they usually observe lessons.

When labellers could also hear the lesson, they reported that they generally found the teacher audible despite background noise. Recordings often also captured pupils talking and moving, and sometimes teaching assistants; speech was harder to distinguish during independent practice or when several people spoke at once. Labellers also reported that audio resolved many ambiguities in the transcript, while video helped with practices that are partly non-verbal. A frequent source of confusion was in questioning, where it was often unclear from the transcript whether the teacher or a pupil was speaking, and audio usually settled this. For most practices, labellers felt video added little once audio was available. The exceptions were mostly in managing behaviour. With video, labellers could see a *correction or sanction* given without words, such as writing a pupil's name on the board, or a teacher standing still while *waiting for silence*. Video also helped labellers judge the warmth behind *encouraging effort*, such as how a teacher says "come on now".

Yet agreement between labellers was about the same with or without audio or video. On the same recordings, agreement between the two labellers in the audio or video passes was within 0.05 of their agreement in the transcript passes in every domain. Because of missing data, however, these comparisons rest on only two to six recordings per domain, so they should be interpreted cautiously (Table S8).

Labellers did, however, change many of their own labels on a second pass. When they labelled the same transcript again with audio or video, their second pass matched their first with an F1 of between 0.43 (explanation and modelling, video) and 0.60 (questioning, audio). In every domain, between 43% and 64% of the labels in the later pass had no match in the earlier one (Table S7). The design makes it hard to say how much of this change came from the audio or video. Because the second pass always came after the first, some changes may simply reflect labelling the same lesson twice, as anyone's judgement shifts over time.

5. Discussion

The pilot showed that the core components of ATP worked in practice, and identified where they needed to change before the main phase of data collection and labelling. Teachers were able to use the recording tool in real classrooms, and our safeguards protected teachers and pupils, with the pilot identifying where transcript anonymisation needed strengthening before wider rollout. Expert labellers were able to identify teaching practices from transcripts, even with transcription errors. Equally, the pilot showed limits in how consistently some practices can be labelled and in what transcripts capture. Below, we discuss what the agreement results suggest automated methods should try to identify; how to handle disagreement; what transcripts can and cannot capture; and the trade-offs involved in doing this research safely.

Starting with agreement, the practices labellers agreed on most shared common features. They mostly occurred often, were defined by what the teacher said, and had clear verbal markers. Examples include *courteous or warm language*, *correction or sanction*, *direct instruction*, and *questions about what is being taught now*. Even so, agreement on these practices was likely lower than it would be with more training and calibration. The workshops support this. When looking at one example at a time, labellers quickly reached consensus on what counted as *direct instruction*.

Practices with low agreement fell into four groups. Some had definitions that labellers understood differently, such as what 'counts' as *encouraging effort* or *narrating thinking*. Some depended on information a transcript does not often contain, such as *waiting for silence.* Some were too rare to judge, such as *abandoning an explanation* or *sharing pupil work*. And for some pairs of practices, labellers found the boundary hard to agree, such as *questions about what is being taught now* and about *earlier content*, or *courteous or warm language* and *positive reinforcement*. Three of these groups echo the main sources of disagreement identified in prior annotation research: genuine ambiguity in what is being judged, information missing from what labellers can see, and definitions that are not precise enough (Plank, 2022; James, 2026).

These results reflect a long-standing finding that identifying teaching practices and evaluating lessons involve judgement, and so vary between people, including between experts (e.g. Kane and Staiger, 2012). The pilot highlights that a shared framework alone is not enough for consistent labelling between experts. Because training was deliberately minimal, our results show what a framework achieves on its own, and leave open how far agreement can rise with calibration. This mirrors wider questions in teaching about professional judgement and shared language: teachers

and teacher educators routinely use terms such as scaffolding or metacognition, yet, as the workshops showed, can mean quite different things by them. Labelling, and the discussions it prompted, brought these differences to the surface. This suggests that labelling may have value as professional development in its own right, consistent with recent findings (Ashraf et al., 2026).

The same subjectivity is present whenever feedback is given and received, which has direct implications for what automated feedback should try to do. From the outset, and based on input from our Expert Working Group and Strategic Advisory Board, ATP has aimed to support mentors' judgement, not replace it. This is why the framework focuses on whether a practice occurred, not how well it was done. Automated feedback can then point mentors to relevant moments, while judgements about what worked, and how the teacher could improve, stay with the mentor. ECTs' and mentors' feedback in the pilot supports this approach. They asked for access to transcripts and a way to tag lessons, so that they could find relevant moments quickly in mentor meetings and relate feedback to specific moments rather than to the whole lesson. These requests come from a small number of teachers, but they match what we anticipate our automated feedback tool will try to do. Recent work with instructional coaches illustrates this in practice, with coaches using automated feedback mainly as a source of evidence, while choosing which moments to discuss, interpreting them with teachers, and judging when the tool was wrong (Malamut et al., 2026).

What the pilot adds is that merely identifying whether some practices occurred involves considerable judgement. Automated methods may therefore need to work with simpler or broader categories than we first planned, while still helping mentors quickly see where to look. For example, rather than asking labellers to decide whether a question checks understanding of what is being taught now or of earlier content—a distinction they often split on—it may be more accurate and useful to identify questions that check for understanding more broadly. Similarly, for practices that call for more interpretation, such as *narrating thinking*, identifying a broader category such as modelling could still show mentors where these practices are likely to occur. We will test changes like this in the next round of labelling.

For the practices we do try to identify, how to respond to disagreement is a central question. Where it reflects definitions that are not yet precise enough, more extensive training and several rounds of calibration should reduce it. We also plan to develop a gold standard: a set of transcripts with expert-agreed labels, against which labellers can be trained and checked. Where disagreement reflects genuine differences in professional judgement, close calibration could hide differences that matter. Recent work on other subjective annotation tasks suggests ways to take account of such disagreement, such as training on the spread of labellers'

judgements rather than a single agreed label (Uma et al., 2021), learning from each labeller separately (Rodríguez-Barroso et al., 2024), or predicting how much labellers are likely to disagree about a given example (Wan et al., 2023). For ATP, this could mean feedback that indicates not only where a practice may have occurred, but how confident or contested that judgement is.

Whatever approach we take, automated methods can only identify what is in the transcripts they work from. Transcripts missed important signals, such as tone of voice, non-verbal behaviour and much of what pupils say, which may help explain why agreement was lowest for the feedback domain; coaches using automated feedback have reported the same limitation (Malamut et al., 2026). Yet, even with omissions, transcripts carried enough evidence for labellers to identify most practices, especially in questioning, and prior work suggests that transcription errors need not prevent models from detecting practices either (Jensen et al., 2020; Wang and Chen, 2024).

Although audio and video helped resolve ambiguities in the transcript, they did not, by themselves, improve agreement between labellers. This highlights the difference between reliability and validity. Agreement shows reliability, the consistency with which labellers apply a framework, but not validity, whether their labels capture what actually happened (Artstein and Poesio, 2008; James, 2026). Hearing or seeing the lesson may have made labels more valid, for example by clarifying who was speaking or revealing a correction given without words, even where labellers did not converge. More broadly, high agreement can arise from oversimplified definitions, while low agreement can reflect genuine ambiguity rather than error (Plank, 2022). Even a gold standard cannot settle this on its own, since it reflects the judgements of the experts who agreed it. Validity will need to be built up through further evidence, such as close examination of ambiguous cases, comparisons with other sources and, in later phases of TIDE, whether practices are linked to pupils' outcomes.

Relying on transcripts nonetheless carries two risks. The first is accuracy. Perceived accuracy is a key barrier to teachers using automated feedback (Jacobs et al., 2022; Demszky et al., 2025; Malamut et al., 2026), so transcription errors, such as pupils' answers attributed to the teacher, could undermine trust in the feedback even where they do not prevent practices being identified. The second, and more fundamental, is focus. If automated feedback draws only on what transcripts capture, mentors may come to focus on what can be measured rather than on the full range of what matters in teaching (Campbell, 1979). How mentors are supported to use automated feedback, as one source of evidence alongside their own observation and knowledge of the ECT, will therefore need careful thought, and the

trial's examination of unintended consequences will help show whether these risks arise in practice.

Some of these transcript limitations stem from a deliberate trade-off. We set up the recording tool to be accessible to as many teachers as possible, using devices they already have, at the cost of lower audio quality and transcript accuracy. Our safeguards involve a similar trade-off: they protect individual teachers and pupils but come at a cost to schools. Requiring schools to opt in limits which schools take part, and managing the research opt-out adds to their workload. Together, these reduce the number of schools in which the tool is available and may make those that take part less representative, and some teachers may avoid using the tool rather than navigate the opt-out process.

Even so, building a safe system is essential to earning trust. Prior work highlights concerns about privacy as a barrier to teachers engaging with tools like this (Demszky et al., 2025). Schools have told us the safeguards are well thought through and the governance package comprehensive, and both our advisory groups have responded positively. Getting this right is integral to being trusted to tackle the questions this work sets out to answer. Balancing these protections against the burden on schools remains a priority, and we will continue to work with schools and others, such as our ethics committee and advisory groups, to find a workable model for the trial in 2027–28.

6. Limitations

The pilot was small, and so our sample of teachers and schools is unlikely to be nationally representative of schools in England: they chose to take part after being invited, and two had helped to design the recording tool. We also did not gather feedback systematically from all participating teachers and so the positive comments reported above might not be representative of their experiences with the tool as a whole. The 34 labelled recordings also came from only seven teachers and were unevenly spread across subjects and year groups. Teachers also chose which lessons to record. If they were more likely to record lessons they expected to go well, the recordings may not represent their usual practice. This will need to be considered when interpreting later analyses linking observed practice to teacher value-added.

How the labelling was set up also limits what the results can show. Labellers were trained before the labelling platform and pilot data were ready, so agreement would likely have been higher, and departures from the intended workflow fewer, had they practised on real transcripts in the platform. Labellers were also paired unevenly across recordings, so some agreement scores depended on which labellers were

included. Because audio and video passes always came after the transcript pass, changes may reflect a second look rather than the audio or video itself, and these comparisons rest on very few recordings (supplementary materials, S10.4).

Without a gold standard, we could measure whether labellers agreed, but not whether they were right. Two labellers can agree and both be mistaken: in the workshops, there were cases where the original labellers had agreed on a label that the group, after discussion, judged to be incorrect. For the main phase, we plan to use a gold standard, a set of transcripts whose labels multiple experts have agreed through discussion, to train and calibrate labellers.

7. Conclusion and next steps

The pilot showed that ATP's core components worked and gave us a clear agenda for the main phase. The recording tool worked in real classrooms to support both ECT development and research. Our safeguards protected teachers and pupils, but the administrative burden they add may limit school uptake, which needs further consideration before the trial in 2027–28. Transcripts contain useful signals of teaching practice, but the framework needs refinement and simplification so that labellers can apply it more consistently.

Our immediate priority will be revising the framework and codebook. Drawing on agreement scores, workshop discussions, and pilot examples, we will decide which practices to keep, revise, merge, or drop, and will add examples and non-examples from real classroom data. In doing so, we will ask what kind of disagreement each practice shows. Where it reflects definitions that are not yet precise enough, we can sharpen them. Where it reflects genuine differences in expert judgement, there may be no single right answer, and this shapes what an automated tool should try to identify. A second round of labelling on a small sample will then show whether the revisions have improved agreement. Several choices remain open, including who will label the main dataset.

We will use the labelled data to develop and test the automated feedback approaches. Before the feedback tool is rolled out, we will evaluate how well it performs across subjects, including modern foreign languages, and whether the error rate is acceptable to teachers and mentors. In parallel, we will prepare for the 2027–28 trial. We will work with schools on a model that balances protections for teachers and pupils against the burden they place on schools. Building on how the recording tool was first developed, we will also do more co-design with ECTs and mentors, both to improve the tool in response to their feedback and to shape how automated feedback is presented to them.

8. References

Abercrombie, G., Dinkar, T., Curry, A. C., Rieser, V., & Hovy, D. (2025, November). Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP* (pp. 63–74).

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24.

Artstein, R., & Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Ashraf, S., Malamut, J., Kumar, V., Pan, G., Nam, H., Tan, M., Langlois, L., Santos-Deonizio, L. C., Higgins, H. S., & Demszky, D. (2026, July). EduCoder: An open-source annotation system for education transcript data. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) (pp. 597–604).

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311–337.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Coe, R., Rauch, C. J., Kime, S., & Singleton, D. (2020). Great teaching toolkit. *Evidence Review. Published by Evidence Based Education in partnership with Cambridge Assessment International Education*.

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 483–505.

Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2025). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education*, *227*, 105183.

Demszky, D., Wang, R., Geraghty, S., & Yu, C. (2023). Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform. EdWorkingPaper No. 23-891. *Annenberg Institute for School Reform at Brown University*.

Department for Education. (2024). *Initial teacher training and early career framework.* https://assets.publishing.service.gov.uk/media/661d24ac08c3be25cfbd3e61/Initial_Teacher_Training_and_Early_Career_Framework.pdf

Goodrich, J. (2024). *Responsive Coaching: Evidence-informed instructional coaching that works for every teacher in your school*. Hachette Learning.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298.

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, 103631.

James, J. (2026). Counting on consensus: Selecting the right inter-annotator agreement metric for NLP annotation and evaluation. *arXiv preprint arXiv:2603.06865*.

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020, April). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464.

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614.

Malamut, J., Demszky, D., Bywater, C., Reinhart, M., & Hill, H. (2026). Facilitating evidence-based instructional coaching with automated feedback on teacher discourse. *Journal of Mathematics Teacher Education*. Advance online publication. https://doi.org/10.1007/s10857-026-09789-4

Plank, B. (2022, December). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10671–10682).

Rodríguez-Barroso, N., Cámara, E. M., Collados, J. C., Luzón, M. V., & Herrera, F. (2024). Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, *12*, 630–648.

Segura-Bedmar, I., Martínez, P., & Herrero-Zazo, M. (2013, June). SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics* (*SEM), *Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 341–350).

Suresh, A., Jacobs, J., Clevenger, C., Lai, V., Tan, C., Martin, J. H., & Sumner, T. (2021, June). Using AI to promote equitable classroom discussions: The TalkMoves application. In *International Conference on Artificial Intelligence in Education* (pp. 344–348). Cham: Springer International Publishing.

Tan, M., & Demszky, D. (2026). Do As I Say: What Teachers' Language Reveals About Classroom Management Practices. *Educational Researcher*, *55*(6), 349–361.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, *72*, 1385–1470.

Wan, R., Kim, J., & Kang, D. (2023, June). Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12, pp. 14523–14530).

Wang, D., & Chen, G. (2024). Are perfect transcripts necessary when we analyze classroom dialogue using AIoT? *Internet of Things*, *25*, 101105.

9. Supplementary materials

S1. School and teacher recruitment

We recruited schools by two routes. First, we invited the schools whose staff had helped design the recording tool as 'super users': five ECTs and four mentors from four schools, originally recruited by email, newsletters and other channels. Second, we invited other schools from the NIoT's founding MATs and Associate Colleges, again by email, newsletters and other channels, and three expressed interest.

Of the seven schools that expressed interest (four 'super user' schools and three others), five signed the data-sharing agreement and four went on to record lessons: two 'super user' schools and two others. The fifth school signed the agreement but did not send out the parent and pupil information or collect opt-outs, so did not record. The other two schools did not complete sign-up. The four recording schools comprised one primary, two secondary and one special school, with four ECT-mentor pairs (eight teachers).

S2. Recordings

Seven of the eight teachers recorded lessons between April and June 2026. In total, they attempted 60 recordings, of which 40 (67%) uploaded successfully. Most failed uploads were lessons of 50 minutes or more recorded in the Safari browser. Successful uploads lasted a median of 42 minutes (range 14 to 69), based on the 37 with a recorded duration.

Of the 40 successful recordings, 34 were used for labelling: 17 primary, 16 secondary and one from the special school. They were chosen according to which recordings were available at the time of labelling, while aiming for a balance across phase, year group and subject (Table 2). Labelled recordings lasted a median of 45 minutes (range 25 to 69), based on the 31 with a recorded duration.

S3. Transcription and redaction

Recordings were transcribed automatically with Microsoft Azure Speech, which labels each change of speaker ('Person 1', 'Person 2' and so on). Personal data were then removed in two layers: Azure's conversational personal-information redaction and spaCy, an open-source language-processing library, to catch names that Azure missed. Removed text was replaced with the placeholder '[PERSON]', whatever type of information it was.

The entity types removed included: personal details (names, addresses, ages, dates of birth, email addresses and phone numbers); identification numbers (e.g. driving licence, passport); financial details (e.g. bank account); digital identifiers (e.g. IP

addresses); and the names of organisations, such as companies, agencies and institutions. After the pilot, we added a lookup list of school names as an additional layer, since school names were not always recognised as organisations.

S4. Developing the teaching-practice framework

We developed the framework in stages, starting from the Initial Teacher Training and Early Career Framework (ITTECF; DfE, 2024). Throughout, the aim was to select practices that are both theoretically important and observable in a lesson transcript.

We began by mapping the elements of the Great Teaching Toolkit (Coe et al., 2020) and the Simple Model of Teaching (Goodrich, 2024) onto the ITTECF's "Learn how to… by…" statements, and vice versa, and added findings from literature reviews on the practices most strongly linked to pupil outcomes. Two experienced practitioners then rated how likely each ITTECF statement was to be observable from a transcript, and we used their combined ratings to prioritise practices for further consideration. Since many ITTECF statements combine several practices, we next split them into discrete, observable practices: for example, a statement about creating an environment where mistakes, effort and perseverance are part of the daily routine became three practices. We also reviewed previous natural language processing (NLP) work in classrooms, to see what had been measured successfully before and how.

Drawing on all three sources (the ITTECF statements, the pupil attainment literature and prior NLP work), we built an inventory of candidate practices. At this stage, we also revisited practices initially rated as unlikely to be observable, to check that nothing had been dismissed prematurely. From this inventory, we drafted a framework proposal and stress-tested it in a series of workshops with experienced teacher educators. These discussions refined the final set of 29 practices for the pilot, which we grouped into four domains.

Table S1. The 29 practices in the pilot framework, with their short definitions and the ITTECF statements they draw on.

Domain	Practice	Definition	Relevant ITTECF statement(s)
Managing behaviour and high expectations	Setting expectations in advance	Teacher states or directs expected behaviour in advance, not reactively.	Teach and rigorously maintain clear behavioural expectations (e.g. for contributions, volume level and concentration).
	Correction or sanction	Teacher reacts to non-compliance with a correction, warning, or sanction.	Use early and least-intrusive interventions as an initial response to low level disruption. Establishing a supportive and inclusive environment with a predictable system of reward and sanction in the classroom. Apply rules, sanctions and rewards consistently in line with school policy.
	Positive reinforcement	Teacher gives praise, recognition or reward for compliant behaviour already shown.	Teach and rigorously maintain clear behavioural expectations (e.g. for contributions, volume level and concentration). Establishing a supportive and inclusive environment with a predictable system of reward and sanction in the classroom. Apply rules, sanctions and rewards consistently in line with school policy.
	Courteous or warm language	Teacher uses polite, respectful or warm wording towards pupils.	Create a culture of inclusion, respect and trust in the classroom that supports all pupils to succeed (e.g. modelling the courteous behaviours expected of pupils).
	Encouraging effort	Teacher encourages pupils to attempt difficult work, persist, or frame mistakes as learning.	Acknowledge and praise pupil effort, emphasising progress being made. Use intentional and consistent language that promotes challenge and aspiration. Create a positive environment where making mistakes and learning from them, and the need for effort and perseverance, are part of the daily routine.
	Waiting for silence	Teacher withholds speech until pupils are silent, specifically before the teacher or a pupil speaks.	Teach and rigorously maintain clear behavioural expectations (e.g. for contributions, volume level and concentration).

Explanation and modelling	Signposting learning	Teacher explicitly states what pupils will learn, practise or be able to do, and/or why it matters.	No direct ITTECF statement; added following workshop discussions
	Direct instruction	Teacher tells pupils a rule, fact, definition, procedure, or explanation.	Using modelling, explanations and scaffolds, acknowledging that novices need more structure early in a domain.
	Linking to prior learning	Teacher explicitly references a previous lesson/topic and connects it to the current content.	Draw explicit links between new content and the core concepts and principles in the subject. Link what pupils already know to what is being taught (e.g. explaining how new content builds on existing knowledge).
	Anticipating mistakes	Teacher explicitly flags a specific likely error or misconception and warns about it.	Identify possible misconceptions and plan how to prevent these forming. Expose potential pitfalls and explain how to avoid them.
	Narrating thinking	Teacher states out loud what they are thinking, noticing, or deciding while working through a problem.	Narrate thought processes when modelling to make explicit how experts think (e.g. asking questions aloud that pupils should consider when working independently and drawing pupils' attention to links with prior knowledge).
	Making the abstract concrete	Teacher translates an abstract idea into something more tangible, such as an analogy, metaphor, example or diagram.	Use concrete representations of abstract ideas (e.g. using analogies, metaphors, manipulatives for counting, examples and non-examples). Interleave concrete and abstract examples, slowly withdrawing concrete examples and drawing attention to the underlying structure of problems.
	Abandoning an explanation	Teacher explicitly signals the explanation is being abandoned and restarted or corrected.	No direct ITTECF statement; added following workshop discussions
	Breaking down an explanation	Teacher explicitly segments an explanation into at least two distinct steps with clear step markers.	Break complex material into smaller steps (e.g. using partially completed examples to focus pupils on the specific steps).

Feedback	Feedback on written work	Teacher gives specific feedback on pupils' written or practical work.	Make use of formative assessment. Using verbal feedback during lessons in place of written feedback after lessons where possible. Monitoring pupil work during lessons, including checking for misconceptions. Focus on specific actions for pupils and provide time for pupils to respond to feedback.
	Feedback on an individual's spoken answer	Teacher gives specific feedback on an individual pupil's spoken answer.	Make use of formative assessment. Using verbal feedback during lessons in place of written feedback after lessons where possible. Focus on specific actions for pupils and provide time for pupils to respond to feedback. Understand that written marking is only one form of feedback.
	Feedback after a whole-class check	Teacher gives specific feedback following a simultaneous whole-class response (e.g. mini whiteboards, a poll, a hinge question or a choral response).	Make use of formative assessment. Using verbal feedback during lessons in place of written feedback after lessons where possible. Focus on specific actions for pupils and provide time for pupils to respond to feedback. Understand that written marking is only one form of feedback. Working with colleagues to identify efficient approaches to marking and alternative approaches to providing feedback (e.g. using whole class feedback or well supported peer- and self-assessment).
	Setting a task to act on feedback	Teacher instructs pupils to act on feedback: correct, improve, edit or redo work.	Focus on specific actions for pupils and provide time for pupils to respond to feedback.
	Self or peer assessment	Teacher instructs pupils to evaluate their own or a peer's work against given criteria.	Scaffold self-assessment by sharing model work and highlighting key details. Working with colleagues to identify efficient approaches to marking and alternative approaches to providing

			feedback (e.g. using whole class feedback or well supported peer- and self-assessment).
	Re-teaching	Teacher re-explains previously taught material because pupil work or answers show understanding is not secure.	Plan regular review and practice of key ideas and concepts over time (e.g. through structured talk activities). Encouraging pupils to share emerging understanding and points of confusion so that misconceptions can be addressed.
	Sharing pupil work	Teacher displays or reads out a pupil's written or practical work to the class.	Scaffold self-assessment by sharing model work and highlighting key details.
Questioning	Scaffolding	Teacher adds structure that helps pupils produce an answer: stems, prompts, narrowed questions.	Provide scaffolds for pupil talk to increase the focus and rigour of dialogue. Reframe questions to provide greater scaffolding or greater stretch.
	Thinking time	Teacher explicitly signals a pause for silent individual thinking before responses are taken.	Provide appropriate wait time between question and response where more developed responses are required.
	Peer discussion	Teacher instructs pupils to discuss a question in pairs or groups before answers are taken publicly.	Plan regular review and practice of key ideas and concepts over time (e.g. through structured talk activities).
	Questions about what is being taught now	Teacher asks a question or sets a short task to check understanding of content taught so far this lesson.	Structure tasks and questions to enable the identification of knowledge gaps and misconceptions (e.g. by using common misconceptions within multiple-choice questions). Include a range of types of questions in class discussions to extend and challenge pupils (e.g. modelling new vocabulary or asking pupils to justify answers). Elaborate on pupil contributions to support pupils' oral language skills and knowledge development.
	Questions about earlier content	Teacher asks a question or sets a short task to check recall of content from a previous lesson or topic.	Use assessments to check for prior knowledge and pre-existing misconceptions. Use retrieval and spaced practice to build automatic recall and application of key

			knowledge. Balance input of new content with revisiting prior learning so that pupils master important concepts. Take into account pupils' prior knowledge when planning how much new information to introduce. Start expositions at the point of current pupil understanding.
	Discovery questions	Teacher asks pupils to generate a concept, rule, or pattern for themselves that has not yet been taught.	Include a range of types of questions in class discussions to extend and challenge pupils (e.g. modelling new vocabulary or asking pupils to justify answers). Elaborate on pupil contributions to support pupils' oral language skills and knowledge development.
	Elaboration questions	Teacher prompts a pupil to extend, justify, clarify, or add detail to an answer already given.	Include a range of types of questions in class discussions to extend and challenge pupils (e.g. modelling new vocabulary or asking pupils to justify answers). Elaborate on pupil contributions to support pupils' oral language skills and knowledge development. Prompt pupils to elaborate when responding to questions to check that correct answers stem from secure understanding. Teach, model, and require high quality oral language, sometimes known as oracy, recognising that spoken language underpins the development of reading and writing (e.g. where appropriate, develop pupils' responses to questions into full sentences).
	Setting an independent task	Teacher sets a sustained task for pupils to complete without ongoing whole-class teacher explanation.	Balancing exposition, repetition, practice and retrieval of critical knowledge and skills.

S5. The codebook

The codebook gave each practice a definition, with further guidance on what the definition includes and excludes, alongside examples and non-examples. It also stated which practices could be applied to the same words: for example, *courteous or warm language* could be combined with any other behaviour practice, but *setting*

expectations in advance and *correction or sanction* could not. As well as guidance for each domain, it set out general principles:

- Code observable evidence, not assumed intent.
- Labels can overlap unless the codebook says otherwise.
- If no label applies, don't force one.
- Highlight the smallest meaningful span that contains the evidence, while ensuring any important context is captured.
- Do not highlight whole paragraphs unless the whole segment is needed.
- Do not avoid labelling a span because you think another domain also applies.

The codebook used in the pilot is available on request. Table S2 shows an example codebook entry for *setting expectations in advance*, with its definition, examples and non-examples.

Table S2. Example codebook entry. The entry for *setting expectations in advance*.

Definition includes	Behaviours that secure attention, establish silence, ensure readiness, or maintain an orderly classroom. Both fuller statements and brief cues ("Pens down", "Eyes this way"). Setting expected behaviour in advance or neutrally, not reactively in response to non-compliance already occurring.
Can co-occur with	*Courteous or warm language*, when polite wording is used to direct behaviour ("Please put your boards down"). Cannot co-occur with *correction or sanction* on the same utterance. If the wording is reactive, code *correction or sanction* only.
Examples	"Everyone should be silent while working." "Pens down. Eyes this way." "When someone is speaking, everyone else listens." "Please put your boards down." (+ *courteous or warm language*)
Definition does not include	Procedural instructions ("Open your books to page 3"). Praise for behaviour already shown. Correcting non-compliant behaviour.
Non-examples	"Open your books to page 3." (Procedural instruction, no label). "I can see everyone has their listening ears on. Well done." (Praise for behaviour already shown, code *positive reinforcement*). "I've already asked you once to put your pens down." (Reactive, code *correction or sanction*).

S6. Labeller training

Labellers were trained in a single day in May 2026, covering the task, the framework, the codebook for each domain and the secure handling of sensitive classroom data. Training stressed that a label should be used whenever a practice

occurred, regardless of how well it was done. Labellers were asked not to discuss their decisions with each other, so that agreement would reflect independent judgements, and the codebook was fixed for the whole labelling period. The aim was not to resolve disagreements, as in essay marking and moderation, but to use them to identify where the codebook needed work. Since the labelling platform was not ready on the training day, labellers practised on synthetic examples rather than real classroom transcripts. Labellers signed a non-disclosure agreement, worked only on NIoT devices, and had a safeguarding escalation protocol to follow.

S7. Labelling design

There were 28 labellers, seven per domain. Each labeller focused only on labelling practices from their allocated domain. Transcripts were allocated in proportion to the number of days each labeller had available for labelling, then balanced within domains. Every recording was labelled independently by two people in each domain, so each transcript was labelled by eight people in total. So that neither could see the other's labels, each labeller worked on their own copy of the transcript. Pairings varied across recordings: some pairs of labellers labelled the same recordings up to four times, while others were never paired. These differences in pairing reflected labellers' availability.

Each recording was assigned to one of four input conditions (Table S3). In every condition, labellers were asked to label the transcript first (transcript pass). In the audio and video conditions, they then labelled the transcript again while listening (audio pass) or watching and listening (video pass). In total, the design required 432 passes: 272 on transcripts alone (34 recordings, four domains, two labellers), plus 80 audio passes and 80 video passes. Agreement was measured by comparing two labellers' passes on the same recording and domain, which we call a 'pair'. The design gave 136 transcript pairs (four domains x 34 recordings), plus 40 audio and 40 video pairs (four domains x 10 recordings).

Table S3. Input conditions (number of recordings).

Condition	What labellers did	Total
T	Transcript only	18
TA	Transcript, then transcript with audio-only	6
TV	Transcript, then transcript with video (and audio)	6
TAV	Transcript, then with audio, then with video (and audio)	4

S8. Missing data and design deviations

Not every planned pass could be used. Of the 432 planned passes, 406 were available for analysis. Passes were missing or unusable for three main reasons. First,

recordings arrived later than planned, meaning transcripts were released in batches and labellers had less time to complete them, so not all passes were completed. Second, some labellers completed their audio or video pass on the same copy of the transcript as their transcript pass, so the later pass overwrote the earlier one. Third, some labellers worked out of order, hearing or seeing the lesson before labelling the transcript alone. These out-of-order transcript passes were kept but analysed separately (see below).

These deviations affected the analysis in two ways. First, they reduced the number of transcript pairs. Of the 136 planned, 130 had both passes available. However, 27 transcript passes had been completed after the labeller had heard or seen the lesson. Excluding the 23 pairs that included at least one of these out-of-order passes left 107. The main analysis therefore uses these 107 pairs, and all 130 are used as a sensitivity check (S11). Second, they reduced the number of audio and video pairs: only 30 of the 40 planned audio pairs and 31 of the 40 planned video pairs were usable. Because comparisons of agreement *between* labellers needed a usable transcript pair for the same recording and domain (S9.6), only two to six recordings per domain were available for this analysis.

S9. Supplementary analysis

S9.1. Counts and label length

Label counts used every usable transcript pass (266), including the 27 completed after the labeller had heard or seen the lesson, since these still describe what was marked in the transcript. For each practice, we calculated the mean number of labels per labeller per lesson and the percentage of lessons in which at least one labeller used it. Counts are given both as labelled and after merging consecutive labels for the same practice, since merged counts are closer to separate 'episodes' of a practice. Labels for the same practice were merged when they overlapped or touched, or when only spaces, a speaker label, or a single character separated them. Table S4 also gives rates per 1,000 transcript words and per hour of recording, since lessons varied in length. We also measured how long labels were, in words. A word was any run of non-space characters containing a letter or digit, leaving out speaker labels. A "[PERSON]" placeholder counted as one word.

S9.2. Agreement

We measured agreement for each practice with the F1 score, which runs from 0 (no agreement) to 1 (complete agreement). F1 is commonly used for labelling spans of text, since there is no fixed set of 'negative' cases (Hripcsak and Rothschild, 2005). The score reflects the share of all labels, from both labellers, that matched a label from the other. For example, if one labeller made 10 labels and the other 6, and 4 of each matched, F1 is 0.5 (8 out of 16).

Two labels agreed when they overlapped by any amount and named the same practice, following the type-matching approach from SemEval-2013 (Segura-Bedmar

et al., 2013), as implemented in the nervaluate Python package (version 1.2.1). Each label could be matched with at most one of the other labeller's labels: where more than one label for the same practice overlapped it, we used the one with the closest boundaries. We added up matched labels and total labels across all lessons before calculating F1, giving a single score per practice (micro-averaged F1).

Where labels agreed, we measured how much text they shared: the spoken text covered by both labels as a percentage of the spoken text covered by either, from 0% (none) to 100% (identical), leaving out spaces, line breaks and speaker labels. We report the median for each practice ('median overlap' in Table S5), which shows how closely labellers agreed on where a practice starts and ends when they agreed it was there.

To make sure our results did not depend on how labels were split or matched, we also calculated agreement on the text itself ('F1 on text' in Table S5). For each practice, we measured how much spoken text each labeller covered and how much both covered, and applied the same F1 formula: twice the text covered by both labellers, divided by the total text covered by each. Text was counted in characters, leaving out spaces, line breaks and speaker labels, and text covered by more than one label from the same labeller was counted only once.

S9.3. Confidence intervals

Confidence intervals for F1 scores were calculated by bootstrapping. For each practice, we resampled lessons with replacement 1,000 times, keeping both labellers' labels for each lesson together, and recalculated micro-averaged F1 for each resample. We then calculated a 95% bias-corrected and accelerated (BCa) interval from the 1,000 scores, a method that adjusts for skew in their distribution. An interval could not be calculated where any resample contained no labels for the practice, which happened only for *sharing pupil work*. For *abandoning an explanation*, labellers never agreed and every resample gave an F1 of 0, so the interval is shown as 0.00 to 0.00.

S9.4. Confusion between practices

Within each domain, we looked at places where the two labellers marked overlapping text with different practices, and neither labeller had also applied the other's practice to that text. We counted each such pair of overlapping labels once, using the main-analysis comparisons. This shows which practices labellers found hard to tell apart, and where definitions and non-examples need sharpening (Figure 6).

S9.5. Transcript quality and speakers

Each transcript was labelled by up to eight labellers (two per domain), so several could flag the same problem. Overlapping flags were therefore counted once, as a

single flagged section, and sections flagged by two or more labellers were treated as confirmed. Flags from audio and video passes were counted separately, since hearing the lesson may have revealed more errors. We also counted the speakers the transcription detected in each lesson, and how many times the speaker changed.

S9.6. Audio and video comparisons

Because only 16 recordings were labelled again with audio, video or both, these comparisons are descriptive only, with no confidence intervals or inferential tests. For each labeller, recording and domain, we compared the transcript pass—where it had been completed before the labeller heard or saw the lesson—with the same person's later audio or video pass. We calculated agreement between the two passes (F1), the change in the number of labels, and how many labels were added (in the later pass, with no match in the transcript pass) or dropped (the reverse), by domain (Table S7).

For each domain, we also compared agreement between the two labellers in the audio/video passes with their agreement in the transcript passes of the same recordings (Table S8). Only recordings with a usable pair in both passes, and with both transcript passes completed before audio or video, were included. Because numbers per practice were generally very small, all practices in a domain were pooled into a single score.

S10. Supplementary results

S10.1. How often each practice was labelled

Table S4 gives counts and label lengths for each practice. Of 12,374 labels in transcript passes, 11,145 remained after merging. Labels were short in managing behaviour (median 4 words) and questioning (median 8), and longer in explanation and modelling (median 17) and feedback (median 16).

Table S4. How often each practice was labelled (all 266 usable transcript passes)

Practice	Mean labels per labeller per lesson	Same, merged	Mean labels per 1,000 words	Mean labels per hour	% lessons where used	Words per label: median (range)
Managing behaviour and high expectations						
Setting expectations in advance	10.6	9.5	2.5	13.5	97%	8 (1 to 81)
Correction or sanction	11.3	9.8	2.7	14.4	97%	7 (1 to 66)
Positive reinforcement	6.4	5.1	1.5	8.9	88%	4 (1 to 54)
Courteous or warm language	29.2	26.2	6.9	35.6	100%	2 (1 to 91)
Encouraging effort	7.2	6.6	1.7	9.9	97%	6 (1 to 143)
Waiting for silence	4.4	3.8	1.0	5.8	76%	7 (1 to 34)
Explanation and modelling						
Signposting learning	4.5	4.1	1.1	5.6	97%	15 (1 to 84)
Direct instruction	20.2	18.3	4.8	24.8	100%	17 (1 to 274)
Linking to prior learning	2.8	2.6	0.7	3.3	85%	14 (1 to 66)
Anticipating mistakes	2.3	2.2	0.5	2.5	76%	13 (1 to 72)
Narrating thinking	5.6	4.9	1.3	6.7	79%	21 (1 to 223)
Making the abstract concrete	4.8	4.4	1.1	6.4	79%	17 (1 to 175)
Breaking down an explanation	4.6	4.2	1.1	4.2	85%	17 (1 to 223)
Abandoning an explanation	0.4	0.4	0.1	0.6	38%	9 (4 to 41)
Feedback						
Feedback on written work	3.1	2.8	0.7	4.1	74%	18 (1 to 255)
Feedback on an individual's spoken answer	6.0	5.6	1.4	8.0	100%	13 (1 to 197)
Feedback after a whole-class check	1.4	1.3	0.3	1.8	68%	19 (1 to 124)
Setting a task to act on feedback	0.8	0.8	0.2	1.1	56%	16.5 (2 to 197)
Self or peer assessment	0.7	0.7	0.2	0.8	35%	16.5 (1 to 85)
Re-teaching	0.8	0.7	0.2	1.0	47%	24 (2 to 255)
Sharing pupil work	0.1	0.1	<0.1	0.1	12%	21 (15 to 24)
Questioning						
Scaffolding	6.4	5.8	1.5	7.9	97%	10 (1 to 70)
Thinking time	0.4	0.4	0.1	0.6	41%	9 (1 to 31)
Peer discussion	1.5	1.4	0.4	1.8	53%	13 (2 to 85)
Questions about what is being taught now	29.3	26.5	6.9	37.5	100%	8 (1 to 95)
Questions about earlier content	10.9	10.0	2.6	12.1	97%	8 (1 to 87)
Discovery questions	3.2	2.9	0.8	4.4	79%	10 (1 to 55)
Elaboration questions	6.0	5.5	1.4	7.4	88%	7 (1 to 36)
Setting an independent task	1.9	1.7	0.4	2.4	91%	14 (1 to 538)

S10.2. Agreement

Table S5 shows the agreement results for the main analysis (107 pairs). Labels is the total number of labels from both labellers across all labelled lessons. Presence agreement is the percentage of lessons in which both labellers used the practice, out of those in which at least one did. Median overlap is the median share of text shared by matching labels, and F1 on text is agreement calculated on the text covered rather than on matched labels (S9.2). A dash means there were no matching labels, or an interval could not be calculated (S9.3).

Table S5. Agreement between labellers by practice (main analysis, 107 pairs)

Practice	Lessons	Labels	F1	95% interval	Presence agreement	Median overlap	F1 on text
Managing behaviour and high expectations							
Setting expectations in advance	28	644	0.35	0.27 to 0.42	96%	95%	0.32
Correction or sanction	28	603	0.51	0.43 to 0.57	93%	100%	0.51
Positive reinforcement	28	337	0.31	0.21 to 0.38	69%	69%	0.40
Courteous or warm language	28	1726	0.56	0.48 to 0.62	96%	93%	0.39
Encouraging effort	28	343	0.17	0.12 to 0.25	75%	54%	0.17
Waiting for silence	28	200	0.21	0.08 to 0.38	48%	100%	0.19
Explanation and modelling							
Signposting learning	24	244	0.37	0.30 to 0.47	92%	98%	0.44
Direct instruction	24	921	0.44	0.38 to 0.50	100%	82%	0.50
Linking to prior learning	24	147	0.38	0.24 to 0.51	80%	58%	0.31
Anticipating mistakes	24	109	0.18	0.07 to 0.39	53%	90%	0.20
Narrating thinking	24	285	0.23	0.11 to 0.31	70%	77%	0.26
Making the abstract concrete	24	252	0.25	0.14 to 0.35	61%	79%	0.31
Breaking down an explanation	24	254	0.06	0.02 to 0.12	47%	79%	0.08
Abandoning an explanation	24	13	0.00	0.00 to 0.00	25%	–	0.00
Feedback							
Feedback on written work	29	190	0.27	0.13 to 0.44	48%	64%	0.36
Feedback on an individual's spoken answer	29	343	0.22	0.15 to 0.29	72%	71%	0.25
Feedback after a whole-class check	29	81	0.17	0.06 to 0.33	29%	83%	0.21
Setting a task to act on feedback	29	54	0.11	0.00 to 0.31	29%	100%	0.08
Self or peer assessment	29	42	0.10	0.00 to 0.32	27%	91%	0.08
Re-teaching	29	46	0.04	0.00 to 0.20	7%	100%	0.04
Sharing pupil work	29	5	0.00	–	0%	–	0.00
Questioning							
Scaffolding	26	325	0.11	0.07 to 0.17	64%	100%	0.12
Thinking time	26	23	0.26	0.00 to 0.50	27%	100%	0.17
Peer discussion	26	81	0.47	0.30 to 0.64	86%	92%	0.50
Questions about what is being taught now	26	1550	0.43	0.34 to 0.52	96%	100%	0.41
Questions about earlier content	26	532	0.41	0.33 to 0.50	76%	100%	0.38
Discovery questions	26	189	0.30	0.14 to 0.47	52%	89%	0.34
Elaboration questions	26	326	0.31	0.23 to 0.42	87%	100%	0.30
Setting an independent task	26	99	0.20	0.12 to 0.33	46%	45%	0.11

S10.3. Transcript quality and speakers

Labellers flagged inaccurate transcription in every lesson, with a median of 18 flagged sections (range 6 to 41), and personal information in 31 of the 34 lessons (Table S6). Across transcript passes, they flagged 684 inaccurate sections, of which 225 were confirmed by two or more labellers, and 206 sections of personal information, of which 115 were confirmed. Hearing or watching the lesson revealed more, and different, errors: 733 of the 926 inaccurate sections flagged in audio or video passes were not flagged in that lesson's transcript passes. The transcription detected a median of two speakers per lesson, and in 14 lessons a single speaker label covered the whole transcript.

Table S6. Transcript quality and speaker identification per lesson (34 lessons; flags from transcript passes)

Measure	Median	Range
Words in transcript	3,912	1,726 to 8,159
Recording length, minutes (31 lessons)	45	25 to 69
Speakers detected	2	1 to 5
Speaker turns	23	1 to 222
Inaccurate sections flagged	18	6 to 41
per 1,000 words	4.6	0.8 to 12.9
confirmed by two or more labellers	5	1 to 19
Personal-information sections flagged	4.5	0 to 14
per 1,000 words	1.5	0.0 to 5.6
confirmed by two or more labellers	2	0 to 10

S10.4. Audio and video comparisons

When labellers labelled a lesson again, their later pass agreed with their own transcript pass with an F1 of between 0.43 (explanation and modelling, video) and 0.60 (questioning, audio) (Table S7). The number of labels they used changed by between −10% and +48%, depending on domain and input. Agreement between labellers changed by no more than 0.05 in any domain with audio or video, although these comparisons rest on two to six recordings per domain (Table S8).

Table S7. Same labeller: transcript pass compared with their later audio or video pass (all practices in a domain pooled)

Domain	Pass	Within-labeller comparisons	Labels: transcript	Labels: later pass	Change	Added	Dropped	F1 between passes
Managing behaviour and high expectations	Audio	11	817	851	+4%	442	408	0.49
	Video	12	935	1,085	+16%	612	462	0.47
Explanation and modelling	Audio	15	622	660	+6%	374	336	0.45
	Video	12	567	838	+48%	538	267	0.43
Feedback	Audio	12	202	216	+7%	120	106	0.46
	Video	15	254	228	-10%	108	134	0.50
Questioning	Audio	11	480	547	+14%	238	171	0.60
	Video	13	791	783	-1%	338	346	0.56

Table S8. Between labellers: agreement in audio or video passes and in transcript passes of the same recordings (all practices in a domain pooled)

Domain	Pass	Recordings	F1: transcript	F1: audio or video	Difference
Managing behaviour and high expectations	Audio	5	0.36	0.41	+0.04
	Video	4	0.32	0.37	+0.05
Explanation and modelling	Audio	5	0.35	0.34	-0.01
	Video	2	0.30	0.25	-0.05
Feedback	Audio	3	0.08	0.04	-0.04
	Video	6	0.19	0.20	+0.01
Questioning	Audio	3	0.41	0.36	-0.05
	Video	5	0.35	0.37	+0.02

S11. Sensitivity checks

This section shows how far the main agreement results depend on choices made in the analysis. Each check recalculated F1 for every practice in the same way as the main analysis, changing one thing at a time.

First, we included all 130 pairs, including transcripts labelled after either labeller had heard or seen the lesson (S8). Second, we merged consecutive labels for the same practice before matching, as we did for the practice counts (S9.1). Third, we removed labels that we deemed likely to be mechanical slips rather than purposeful judgements, such as a practice label on a single full stop, or on a pupil's name that the labeller marked as personal information at the same time. Specifically, we removed labels that covered no spoken words; labels that covered only a redacted name, or a short stretch the same labeller had marked as personal information; and labels shorter than a minimum length set for each practice. This removed 64 of the 9,964 labels (0.6%) in the main analysis. Finally, to see how far scores depended on particular labellers or lessons, we left out each labeller in turn, and then each lesson in turn, and report the range of scores.

Table S9 sets the main F1 scores alongside each check. Of the 29 practices, 23 changed by 0.05 or less with all 130 transcript pairs included, no score changed by more than 0.04 after merging, and none by more than 0.01 without labelling slips. The largest changes with all 130 pairs were for *abandoning an explanation* (0.00 to 0.21), *correction or sanction* (0.51 to 0.43), *narrating thinking* (0.23 to 0.30) and *breaking down an explanation* (0.06 to 0.13).

Table S9. Agreement (micro-averaged F1) under each sensitivity check

Practice	Main (107 pairs)	All 130 pairs	Merged	Slips removed	One labeller left out at a time (range)	One lesson left out at a time (range)
Managing behaviour and high expectations						
Setting expectations in advance	0.35	0.34	0.37	0.35	0.32 to 0.39	0.33 to 0.36
Correction or sanction	0.51	0.43	0.52	0.51	0.40 to 0.54	0.49 to 0.52
Positive reinforcement	0.31	0.26	0.27	0.31	0.27 to 0.34	0.29 to 0.33
Courteous or warm language	0.56	0.54	0.55	0.56	0.50 to 0.58	0.54 to 0.57
Encouraging effort	0.17	0.15	0.18	0.17	0.11 to 0.21	0.16 to 0.18
Waiting for silence	0.21	0.23	0.23	0.21	0.12 to 0.35	0.16 to 0.29
Explanation and modelling						
Signposting learning	0.37	0.38	0.40	0.38	0.31 to 0.47	0.34 to 0.39
Direct instruction	0.44	0.44	0.47	0.44	0.41 to 0.49	0.43 to 0.46
Linking to prior learning	0.38	0.38	0.37	0.38	0.30 to 0.47	0.35 to 0.43
Anticipating mistakes	0.18	0.23	0.19	0.19	0.10 to 0.26	0.12 to 0.21
Narrating thinking	0.23	0.30	0.24	0.23	0.17 to 0.28	0.19 to 0.24
Making the abstract concrete	0.25	0.28	0.26	0.25	0.21 to 0.28	0.21 to 0.27
Breaking down an explanation	0.06	0.13	0.06	0.06	0.03 to 0.14	0.04 to 0.07
Abandoning an explanation	0.00	0.21	0.00	0.00	0.00 to 0.00	0.00 to 0.00
Feedback						
Feedback on written work	0.27	0.25	0.29	0.28	0.11 to 0.35	0.16 to 0.30
Feedback on an individual's spoken answer	0.22	0.20	0.23	0.23	0.14 to 0.25	0.20 to 0.23
Feedback after a whole-class check	0.17	0.17	0.19	0.17	0.04 to 0.21	0.13 to 0.20
Setting a task to act on feedback	0.11	0.11	0.12	0.11	0.00 to 0.15	0.00 to 0.13
Self or peer assessment	0.10	0.08	0.10	0.10	0.06 to 0.13	0.05 to 0.14
Re-teaching	0.04	0.04	0.05	0.04	0.00 to 0.07	0.00 to 0.05
Sharing pupil work	0.00	0.00	0.00	0.00	0.00 to 0.00	0.00 to 0.00
Questioning						
Scaffolding	0.11	0.11	0.12	0.11	0.07 to 0.13	0.10 to 0.12
Thinking time	0.26	0.24	0.26	0.26	0.12 to 0.50	0.19 to 0.35
Peer discussion	0.47	0.41	0.49	0.47	0.33 to 0.59	0.42 to 0.51
Questions about what is being taught now	0.43	0.40	0.45	0.43	0.37 to 0.46	0.40 to 0.45
Questions about earlier content	0.41	0.35	0.43	0.41	0.29 to 0.46	0.38 to 0.42
Discovery questions	0.30	0.28	0.29	0.30	0.22 to 0.39	0.23 to 0.34
Elaboration questions	0.31	0.31	0.33	0.32	0.27 to 0.37	0.28 to 0.33
Setting an independent task	0.20	0.18	0.22	0.20	0.16 to 0.23	0.17 to 0.22