UCL

CENTRE FOR EDUCATION POLICY & EQUALISING OPPORTUNITIES

Working Paper No. 23-09

Modelling evidence-based practice in initial teacher training: causal effects on teachers' skills, knowledge and self-efficacy

Sam Sims
Ambition Institute & UCL

Harry Fletcher-Wood
Ambition Institute

Thomas Godfrey-Faussett
University of Oxford

Peps Mccrea
Ambition Institute

Stefanie Meliss
Ambition Institute

Teacher education/training often incorporates observable examples of focal teaching practices – models. Yet, there is little causal evidence on the benefits of models or how best to design them. We used a classroom simulator experiment to test the effects of video models on trainee teachers' skills, knowledge, and self-efficacy in relation to using retrieval practice at the end of a primary school science unit. Results showed that models improved participants' skills, but not their knowledge or self-efficacy. Adding annotations to the models had no additional benefit. Incorporating models in initial teacher training can help new teachers make better use of evidence-based teaching practices.

VERSION: August 2023

Suggested citation: Sims, S., Fletcher-Wood, W., Godfrey-Faussett, T., Mccrea, P., Meliss, S. (2023). *Modelling evidence-based practice in initial teacher training: causal effects on teachers' skills, knowledge and self-efficacy* (CEPEO Working Paper No. 23-09). Centre for Education Policy and Equalising Opportunities, UCL.

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- We use a classroom simulator experiment to test the value of modelling in initial teacher education.
- Video models improve trainee teachers' skills in the use of evidence-based retrieval practice methods.
- However, models do not improve trainee teachers' knowledge or self-efficacy with respect to evidence-based retrieval practice methods.
- Adding annotations to the video models, highlighting and explaining the evidence-based practices, has no additional detectible benefit over a simple video model.
- Teacher educators should consider incorporating models in initial teacher education to help trainees develop evidence-based practice.

Why does this matter?

Teachers can understand the theory behind some evidence-based practice without knowing how to put it into practice in the classroom. Modelling can bridge this 'knowing-doing gap'.

Modelling evidence-based practice in initial teacher training: causal effects on teachers' skills, knowledge and self-efficacy

Sam Sims [a b]
Harry Fletcher-Wood [a]
Thomas Godfrey-Faussett [a c]
Peps Mccrea [a]
Stefanie Meliss [a d]

[a] *Ambition Institute*
[b] *UCL*
[c] *University of Oxford*
[d] *University of Reading*

Teacher education/training often incorporates observable examples of focal teaching practices – models. Yet, there is little causal evidence on the benefits of models or how best to design them. We used a classroom simulator experiment to test the effects of video models on trainee teachers' skills, knowledge, and self-efficacy in relation to using retrieval practice at the end of a primary school science unit. Results showed that models improved participants' skills, but not their knowledge or self-efficacy. Adding annotations to the models had no additional benefit. Incorporating models in initial teacher training can help new teachers make better use of evidence-based teaching practices.

Key words: teachers, professional development, models

Acknowledgements: Thanks to Jennifer Barker, Abigail Brown, Hilary Spencer, Marie Hamer, Paul Murphy, Raksha Pattni, Sarah Cottingham, Steve Farndon, Nick Pointer, Nick Rose, Susan Dutta, Anna Bartkiewicz, Gorana Henry, Tessa Willy, Jemima Rhys-Evans, Laura Senior, and the staff and pupils at Dixons Academies Trust.

Declaration of interests: All five of the authors work for organisations that provide teacher training in return for fees.

Funding: This work was supported by Ambition Institute.

Introduction

Policymakers and educators have long disputed whether initial (pre-service) teacher education is adequately preparing trainees for the classroom (Knight, 2021; Orchard & Winch, 2015; Zeichner, 2006). For example, Kagan (1992) argued that initial teacher education programmes placed too much emphasis on theoretical knowledge and were thereby failing to equip pre-service teachers the skills needed to manage their classrooms and provide effective instruction. Relatedly, Kennedy (1999) has consistently highlighted the 'problem of enactment', whereby knowledge of an idea or theory substantially underdetermines what teachers should do to put that theory to use in the classroom. Without receiving guidance on how to act on a theory "teachers can learn and espouse one idea, yet continue enacting a different idea, out of habit, without even noticing the contradiction" (Kennedy, 1999, p947). This problem has also been referred to as the knowing-doing gap (Knight et al., 2013).

How should teacher educators address this challenge? One frequently proposed solution is to incorporate modelling - observable examples of teaching practice - to illustrate the use of theory in practice. Indeed, a recent systematic review found that around two thirds of evaluated in-service professional development programmes incorporate some kind of modelling of teaching practice (Sims et al., 2022). Having said that, a recent representative poll of teachers in England found that only 21% of them reported that the PD they took part in 'Always' or 'Often' included modelling (Ofsted, 2023), suggesting that modelling may be less common outside of formal, manualised PD programmes. Modelling has also been incorporated in theories of teacher development. For example, representations of teaching practice (of which models are one important type) play a prominent role in Practice Based Teacher Education (PBTE), where they are theorised to help pre-service teachers attend to and notice important features of teaching practice (Grossman, 1992; Grossman et al., 2009; Hauser & Kavanagh, 2019; Kosko et al., 2021). Recent theories of effective (in-service) teacher PD also posit that modelling is causally active in helping teachers develop new teaching techniques (Sims et al., 2023).

As a result of its widespread use and theoretical prominence, modelling has now become the focus of a growing academic literature. For example, there are many illuminating case studies of the use of modelling in initial (pre-service) and continuing (in service) teacher education/training (Eick et al., 2003; Loughran, 1995; Loughran & Berry, 2005; Kluth & Straut, 2003; Saclarides & Munson, 2021). Libraries of video models play a prominent role in the extensively evaluated My Teaching Partner instructional coaching programme (Allen et al., 2011; Allen et al., 2015). Yet, causal evidence on the impact of modelling remains scarce. For example, a recent systematic review of teacher preparation practices does not appear to include any impact evaluations of modelling

(Mancenido, 2022). This reflects a general dearth of what Hill et al. (2021) refer to as *effectiveness research* in teacher education. Relatedly, the existing literature contains little evidence on which types of models are most effective. For example, models differ in terms of what they make visible to trainee teachers (Grossman et al., 2009) and how they make links to the underpinning theory (Rich & Hannafin, 2009). Understanding how to highlight theory and link it to practice within a model is therefore critical if research is to provide actionable insights for teacher educators responsible for designing and delivering professional development (Daniel & De Bruckyere, 2021; Hill et al., 2013).

In this paper, we address this gap in the literature using the pathbreaking classroom simulator experiment paradigm developed by Cohen, Wong, Krishnamachari, & Berlin (2020). This allows us to test the impact of different models of evidence-based practice by randomly allocating initial teacher trainees to three treatment arms: 1) restudying a summary of the evidence underpinning the evidence-based practice (*restudy*), 2) watching a video model of the evidence-based practice (*model*), 3) watching a video model of the evidence-based practice with the evidence integrated into the model *(model with theory).* This allows us to make two novel contributions to the literature. First, we provide the first causal test of the theory that modelling helps teachers develop skills in the use of evidence-based teaching practices. Second, we provide new causal evidence on how best to design video models so that teachers can make the links between the observable teaching techniques and the underpinning theory. Our findings are of direct relevance to teacher educators looking to support early-career teachers' development of evidence-based practice.

Theory and hypotheses

Representations make aspects of teaching visible to trainee teachers and can include worksheets, lesson plans, or videos (Grossman et al., 2009; Grossman et al., 2018). Where representations directly depict teaching, this is known as a model – an observable example of some focal teaching practice (Sims et al., 2022). Some models are 'live' in that they are delivered in person, for example when a coach demonstrates a teaching move to a coachee. Other models are 'symbolic' in that they are captured in an image. Some types of models, such as video or live modelling, can be annotated or talked over in a way that would be difficult with live classroom teaching. This is important because it can help to reveal the underlying principles at work, the purposes behind decisions, or elements which aren't visible in the model. Regardless of specific design choices, models generally serve to help trainee teachers develop a mental 'image' of the focal teaching practice (McDonald et al., 2013), which can then be used to guide trainees' practice.

Modelling and skills

Skills are improvable abilities to perform actions that bring about a socially desirable outcome (Green, 2011). Models are thought to support the development of teaching skills by providing a cognitively efficient guide to such action, in the sense that *a picture is worth a thousand words* (Noble, 1997). Cognitive scientists have long known that providing novices with worked examples helps them to learn procedural knowledge (Booth et al., 2015; Sweller, 2006). Procedural knowledge refers to memory of the series of steps or actions needed to accomplish a goal, and often underpins the actions that skilled individuals use to bring about some outcome (Rittle-Johnson, Schneider, & Star, 2015). Recent research on the 'human movement effect' suggests that worked examples can also help with learning skills, in that humans have considerable capacity for learning from watching moving images of people doing things (Höffler & Leutner, 2007; Sepp et al., 2019; Van Gog et al., 2009; Wulf et al., 2010).

We are not aware of any experimental study isolating the effects of modelling on teacher skills. However, empirical support for the importance of modelling is available from two other domains. First, psychologists have shown using highly stylised lab experiments that modelling helps with the acquisition of new skills (Richardson & Lee, 1999; Weeks & Anderson, 2000). Second, many experimental studies in the medical education literature have found that modelling helps trainees with the acquisition of new clinical (Cordovani & Cordovani, 2016) and surgical skills (Harris et al., 2018). These studies in the medical and surgical education literature often use exposure to written guidance as an active control condition (e.g., Custers et al., 1999). Based on the preceding theory and empirical evidence, we hypothesise that:

> H1: Exposure to a video model of some evidence-based teaching practice will improve pre-service teachers' skills in the use of that evidence-based practice, relative to rereading the evidence behind the practice (with no model)

As regards the design of models, careful observational studies have found that novice teachers often struggle to notice the important features of a representation of practice (van Es & Sherin, 2002; Sherin & van Es, 2005; Brunvand & Fishman, 2006). The relevant information contained within the model may therefore be lost in the "complex perceptual field" of a classroom scene (Goodwin, 1994, p. 606). Even if trainee teachers do notice the important features of some model, they may fail to understand how a particular approach brings about greater pupil learning (Rich & Hannafin, 2009). Theorists have therefore emphasised the importance of highlighting relevant features of the model and explicitly providing the underpinning knowledge about how some aspect of practice supports pupil learning (Goodwin, 1994; McGrew et al., 2018; Sherin & van Es, 2009). This is thought to help teachers better understand the links between their actions and pupil

learning, thus supporting skilful teaching. Empirical research using stylised lab experiments supports the notion that models which label relevant features and state the underpinning knowledge contribute to faster skill growth, relative to models that do not do this (Carroll & Bandura, 1990). However, we also note that results from analogous studies conducted in the domain of physical education are somewhat more mixed (Han et al., 2022). Based on the preceding theory and empirical evidence, we hypothesise that:

> H2: Exposure to a video model in which the important aspects of practice are highlighted and the underlying knowledge is stated will improve pre-service teachers' skills in the use of evidence-based practice, relative to exposure to the same model without highlighting the important aspects of practice or stating the underlying knowledge.

Modelling and knowledge

Modelling has traditionally been thought of as useful for helping observers acquire the skills represented in the model. However, researchers have become increasingly interested in whether modelling can also help the observer acquire knowledge. There is a long-running debate in the math education literature (Baroody, 2003) about whether pupils should be taught procedural knowledge (which often underpins skill) first, or whether they should be taught conceptual knowledge (underlying mathematical facts and principles) first. However, recent empirical work suggests that there is in fact a bi-directional relationship, in which procedural and conceptual knowledge are mutually supportive of each other (Rittle-Johnson & Schnieder, 2015). This suggests that integrating instruction on the two may benefit pupil learning of both. This is consistent with a large body of evidence from cognitive science showing that new knowledge is more likely to be retained if it relates to other existing knowledge (Van Kesteren et al., 2012; Van Kesteren et al., 2014).

More recently, researchers in the field of medical education have become directly interested in whether modelling helps support learning of new knowledge (Wood et al., 2007). In particular, they have begun testing whether integrating instruction on clinical procedural skills (how to treat a patient) with basic biochemistry knowledge leads to superior learning of the latter. As with the literature on math teaching, theorists argue that creating the connection between these two types of knowledge helps to secure both (Kulasegaram et al., 2013). Consistent with this, two experimental studies have now shown that integrating instruction on (clinical) skills in a video model with instruction on the underpinning (biochemistry) knowledge does indeed increase knowledge retention, relative to providing the instruction on the two separately (Cheung et al., 2019; Cheung et al., 2021). Reasoning by analogy with the math literature, and in line with the medical education literature, we hypothesise that:

H3: Exposure to a video model of some evidence-based teaching practice integrated with the underpinning knowledge will enhance pre-service teachers' knowledge, relative to just re-reading the underpinning knowledge.

Modelling and self-efficacy

Modelling is also thought to improve self-efficacy. Bandura (1977) defined perceived self-efficacy as personal judgements of one's capabilities to organise and execute action to attain designated goals. Teacher self-efficacy therefore refers to personal beliefs about one's abilities to help students learn (Woolfolk-Hoy, Hoy, & Davis, 2009). Bandura (1997) argued that self-efficacy beliefs can be developed through four different methods, one of which he called 'vicarious modelling' - observing somebody doing the action. Models appear to have a greater effect on self-efficacy when the observer perceives the modeler to be similar to them (Labone, 2004; Schunk & Hanson, 1985). This suggests that seeing somebody else do something prompts the observer to reason that *if you can do it, then I can do it too* (Johnson, 2010; Schunk & DiBenedetto, 2021). In short, when a pre-service teacher observes another teacher successfully using some practice, they are thought to positively update their beliefs about their own ability to use that teaching technique (Tschannen-Moran, Hoy, & Hoy, 1998).

Several qualitative studies have illuminated the links between modelling and pre-service teacher self-efficacy (Bautista, 2011; Bautista & Boone, 2015; Palmer, 2006; Palmer, 2011). Two experimental studies suggest that this reflects a genuine causal relationship between exposure to modelling (as opposed to instruction) and self-efficacy among pre-service teachers (Gorrell & Capron, 1990; Gorrell, 1993). Based on the preceding theory and empirical evidence, we hypothesise that:

H4: Exposure to a video model of some evidence-based teaching practice will increase pre-service teachers' self-efficacy in the use of that evidence-based practice, relative to re-reading the theory behind the evidence-based practice.

Current study

The aim of the current study is to test these hypotheses experimentally, by comparing different approaches to training early-career teachers. In particular, we set out to compare how the presence or absence of different types of models change teachers' skills, knowledge, and self-efficacy relating to evidence-based teaching practices. We wanted to focus our study on a well-researched, well-evidenced area of teaching practice. We therefore chose to focus on questioning for retrieval. Retrieval practice involves "prompting students to recall information from memory, rather than representing or restudying the information" (Perry et al., 2021, p. 69) and is known to improve

pupil learning of both factual and conceptual knowledge (for reviews, see Kornell & Vaughn, 2016; Yang et al., 2021). Questioning for retrieval involves teachers verbally posing questions to students for the purposes of retrieval practice. All participants in the study started by reading a written summary of the evidence around effective questioning for retrieval. We then randomly allocated participants to restudy the evidence summary on questioning for retrieval with no model (*restudy*), watch a video model of evidence-based questioning for retrieval (*model*), or watch a similar model with integrated text snippets explaining the rationale behind the teachers' actions (*model with theory*). This study was granted ethical approval by the UCL Institute of Education Research Ethics Committee.

Methods

Participants and design

We aimed to recruit at least 30 participants for each of the three arms in our experiment. This provided a comparable sample size to those in previous simulator experiments, which were able to detect effects across a range of outcome measures (Cohen et al., 2021; Hill et al., 2021). Individuals were eligible to participate in the experiment if they had enrolled on a primary (elementary) initial teacher training course in the 2022/23 academic year. Recruitment opened on 1st of October 2022 and closed on 23rd December 2022. We recruited participants by approaching initial teacher training providers and asking them to advertise the study to their trainees. Recruitment to the experiment was done on a rolling basis and participants were free to book a slot at a time that was convenient for them. The final group of participants (N = 89) should therefore be considered a convenience sample, with the representative participant in our study being a white, 29 year old female from Greater London. On completion of all data collection, participants were given an Amazon voucher in recognition of taking part.

We tested our hypotheses using an A/B/C test lab experiment, which are becoming increasingly common in this literature (Cohen & Wiseman, 2019; Cohen et al., 2020; Cohen et al., 2021). Unlike field experiments in education, which are often lacking in statistical power (Lortie-Forgues & Inglis, 2019; Spybrook et al., 2016), such lab experiments offer potentially better powered experimental tests of theoretically-derived hypotheses (Hill et al., 2021; Sims et al., 2023). We randomly allocated participants to the three experimental arms. To implement the randomisation, we generated a sequence of 150 random allocations using the Stata package RANDOMIZE (Kennedy & Mann, 2015). Participants were then randomised at the point of check-in. There was no way that participants could anticipate their treatment allocation when they booked their slot. Table 1 shows the balance of participant characteristics across the three arms. A joint (F) test of orthogonality between these characteristics and treatment allocation did not find any undue imbalance across

groups ($p = 0.72$). It may be noted that there are small numbers of participants in particular ethnicity cells in Table 1. However, any between-group differences in ethnicity are controlled for via the ethnicity covariates included in our models.

TABLE 1

Descriptive statistics for the three treatment arms

	Restudy	*Model*	*Model w/ theory*
Female (%)	74.2	89.6	86.2
Age (years)	29.7	28.5	29.6
Ethnicity (%)			
White	67.7	82.8	60.7
Minority ethnic	30.3	17.4	39.2
Region (%)			
East Mids / East	19.4	20.7	20.7
London / South East	29.1	37.9	34.5
North East / North West	29.1	24.1	20.7
West Midlands	19.4	17.2	20.7
Efficacy pre-test (Z score)	0.2	-0.2	0.01
Skill pre-test (Z score)	-0.1	0.07	0.03
No. of participants	31	29	29

Note. Teaching hours = estimated hours of accumulated teaching experience. Percentages may not sum to 100 within categories due to rounding. There were no participants from the South West region or Yorkshire and Humber region. East Mids = East Midlands. Some contiguous regions combined to avoid potential disclosure due to single observation cells. Some ethnic groups combined to avoid potential disclosure due to single observation cells SD = standard deviations. Model w/ theory = model with theory.

Procedure and stimuli

The experiment was conducted entirely online using Zoom video conferencing software. Four different experimenters took it in turns to facilitate the sessions. As previously mentioned, all participants began the experiment by reading the 'evidence-based instructional summary'. This document is central to our study, since it provides the basis for both our video models and the way in which we measure teacher skills within the simulator. The full document is available in Appendix A. For space reasons, we limit ourselves here to highlighting the five principles for questioning for retrieval contained in the summary:

1. When asking a question, teachers should make it clear that any student could be called upon to respond. This increases the benefits of questioning for retrieval by prompting more students in the class to search for and retrieve the correct answer from memory (Dallimore, Hertenstein, & Platt, 2013; Kalamar, 2018; MacSuga-Gage & Simonsen, 2015; Sumeracki & Castillo, 2022).
2. Teachers should give students three seconds or more between asking a question and calling on a student to answer. This gives all students a chance to retrieve the knowledge. If the answer is revealed faster than this, then it is more likely that some students will be restudying

the material, rather than retrieving it, which is known to be less effective than retrieval (Tobin, 1987; Yang et al., 2021).

3. If a student gets an answer incorrect, then teachers should frame this as a learning opportunity. This helps maintain students' motivation toward learning (Käfer et al., 2019; Metcalfe, 2017; Soncini, Matteucci, & Butera, 2021; Tulis, 2013).
4. If a student gives an incorrect response, teachers should inform the student that the answer is incorrect, as this focuses their attention on the correct answer. The benefits of incorrect retrieval are just as large as for correct retrieval, so long as teachers give the correct answer and then explain why this is correct by relating it to students' existing knowledge (Kornell, Klien, & Rawson, 2015; Metcalfe, 2017; Metcalfe & Huelser, 2020; Wong & Lim, 2019).
5. If a student is not able to give any answer to the question, teachers should proceed to give the student a partial hint. This maximises the extent to which students subsequently retain the target knowledge by allowing the student to retrieve the part of the answer not contained within the hint (Kornell & Vaughn, 2016; Vaughn & Kornell, 2019; Vaughn et al., 2022).

After reading the evidence-based instructional summary, all participants took part in a classroom simulator session task (McGarr, 2021) in which they were tasked with asking students a series of questions in a way that aligned with the evidence in the instructional summary. Participants were requested to ask the questions "in such a way that it encourages students to retrieve what they already know" and were asked to "use the information in the evidence-based summary to guide [their] practice". We used the Mursion simulator environment (Cohen et al., 2020; Ferguson & Sutphin, 2022) implemented within the online video conference call. Mursion is a mixed reality environment in which five primary/elementary school pupil avatars are controlled by a human simulator specialist and/or the underlying software (an image of the Mursion interface can be found in Appendix B). We provided the human simulation specialist with a script detailing how to respond to the teacher's questions. For example, the avatar pupils gave a correct response to the first and fourth question, a partially-correct response to the second and fifth question, and an 'I don't know' answer to the third and sixth question. This allowed us to ensure consistency across participants.

This simulator task was embedded in a wider 'scenario' that we designed for the purposes of the experiment. Participants entering the simulator were told that they had just finished teaching a year 4 (age 8-9) primary school science unit focused on the physics of sound. They were provided with a copy of the unit summary (see Appendix C), which was taken from a real primary school in England, and covers material from the English national curriculum. They were also provided with a set of six questions to ask the pupils, drawn from the unit summary, along with the desired answers to each question (Appendix C).

After the first attempt in the simulator, participants' experience diverged based on their treatment allocation. All participants were asked to "recap the evidence on questioning for retrieval" before "repeat[ing] the same teaching activity with the simulated group". Those randomly allocated to Arm 1 (*restudy*) were given 4.5 minutes to restudy the evidence-based instructional summary document, which all participants had already read prior to their first attempt in the simulator. This is a common approach to creating an active control group in the medical simulation literature, which has the benefit of equating the duration of training across the experimental arms (Cordovani & Cordovani, 2016; Custers et al., 1999; Harris et al., 2018). Those in Arm 2 (*model*) were shown a video in which a real primary school teacher asked five questions to a group of seven real primary school pupils. Some of these questions were met with correct responses, some with incorrect responses, and some with an 'I do not know' response. The teacher in the video consistently demonstrated all five of the evidence-based principles of questioning for retrieval set out above.

Those in Arm 3 (*model with theory)* were shown a very similar video, in which the footage shown to those in Arm 2 was interspersed with annotations containing some of the text from the evidence-based instructional summary. For example, after the teacher poses a question and waits three seconds before selecting a pupil to respond, the video cuts away to show the following text for five seconds: "By waiting three seconds after posing a question, the teacher gives all pupils sufficient time to attempt retrieval". Likewise, after the teacher receives an incorrect response from a pupil and frames this a learning opportunity, the video cuts away to show the following text for five seconds "By framing mistakes as an opportunity to learn, the teacher helps prevent pupils becoming demotivated." In line with the theory above, these text snippets were intended to highlight the relevant parts of the video model and make explicit the rationale for specific techniques demonstrated in the model. Five such statements were included in the Arm 3 video.

Both the Arm 2 and Arm 3 videos were 4.5 minutes long. Screenshots of the videos, and links to the full videos online, are available in Appendix D. Following this, all participants had a second attempt at the same simulator task.

Measures

We measured participants' skills in using questioning for retrieval in their first attempt in the simulator (pre-test) and in their second simulator attempt (post-test). We operationalised this measure using a novel coding framework applied to video clips of participants' teaching within the simulator. The video clips were first edited so that coders cold not tell from watching the video which treatment arm the participant was in. The coding tool was designed to capture the five principles of evidence-based questioning for retrieval set out above. For example, for principle 2 (wait time), for each of the six questions, we measured whether teachers left three seconds between

asking a question and asking a student to answer. Similarly, for principle 3 (framing incorrect answers as learning opportunities), there were two questions in the simulation in which the pupil gets the question wrong. In each case, we captured whether the participating teacher framed this error as a learning opportunity e.g., by saying that the class could now work together to get the answer right. In our coding framework we developed a rule for when to award credit for each of the five principles, a set of creditworthy examples, and a set of examples that were not creditworthy. We then refined this coding tool by piloting it on a number of pilot simulator sessions before the experiment began. The full coding tool is available in Appendix E.

Across the five metrics, the maximum score was 18 points, reflecting six opportunities to pose questions to all students, six opportunities to use wait time, two opportunities to frame errors as learning opportunities, two opportunities to give elaborative feedback, and two opportunities to give hints in response to 'I don't know' answers. Crucially, participants had to select the best responses based on how the pupils responded to the question they had asked. Cronbach's alpha across all the indicators was 0.84. We double-coded the first 18 simulator sessions (with raters blind to each other's scores) and calculated inter-rater agreement (Cohen's Kappa) to be 0.81. There were more opportunities to gain marks for some of our metrics (see Appendix E). For example, the wait time component of the outcome measure (maximum six marks) was worth more than the elaborative feedback component (maximum two marks). To give each of the five metrics equal weight, we standardised the five metrics separately, then summed them and standardised this total score.

We measured participants' knowledge using a six-item multiple-choice test. To ensure that participants in Arm 1 (*restudy*) and Arm 3 *(model with theory)* had equal exposure to the content, this test exclusively covered knowledge that was included in both the evidence-based instructional summary document and the video with integrated theory. We made two design choices intended to minimize the chances of participants guessing the correct answers. First, all question had four possible response options including plausible incorrect answers. Second, all questions followed a 'please select all correct answers' format, so that participants did not know how many correct answers there were for each question. There were a total of 11 correct responses across the six questions. We calculated a sum score capturing the total number of correct answers identified by participants, minus the total number of incorrect answers. The full test instrument is available in Appendix F. We collected this measure on just one occasion. We sent participants the test seven days after they took part in the simulator, and asked them to complete it immediately (late responses are addressed in the analysis below). We decided not to collect a pre-test measure because our piloting of the test showed clear ceiling effects when the test was administered immediately after participants had been exposed to the evidence-based instructional summary document, but no ceiling effects a

week later. We judged that a pre-test measure collected prior to exposure to the instructional summary would likely have shown floor effects because the material would likely be unfamiliar to many of our early-stage trainee participants. Collecting our post-test measure with a seven-day delay was necessary to assess knowledge retention.

We measured participants' self-efficacy in using questioning for retrieval immediately after their first attempt in the simulator (pre-test) and immediately after their second simulator attempt (post-test). We operationalised this measure using a heavily adapted version of the Teacher Self-Efficacy Questionnaire (Tschannen-Moran & Hoy, 2001). We asked participants to reflect on the simulator session they had just completed and used the stem 'how well do you feel you' applied to five questionnaire items, each of which corresponded to the five principles of evidence-based questioning for retrieval. For example, for principle 5, we asked 'how well do you feel you... provided hints when students were struggling to answer a question?' The full questionnaire is available in Appendix G. Responses were collected on a five- point scale ranging from 'Not at all well' to 'Extremely well'. Cronbach's alpha across the five items was 0.78. We calculated an overall score using confirmatory factor analysis. Descriptive statistics for the pre-test are in Table 1.

The overall design of the experiment, including stimuli, measures, and treatment arms, is summarised in Figure 1 below. Figure 2 provides a CONSORT diagram summarizing the flow of participants through the experiment. One participant from the *model with theory* arm declined to provide a post-test measure of self-efficacy when responding to our post-test questionnaire and therefore could not be used in our self-efficacy analyses. One further participant, also from the *model with theory* arm, declined to provide demographic information and therefore could not be included in our (pre-registered) regression analyses.

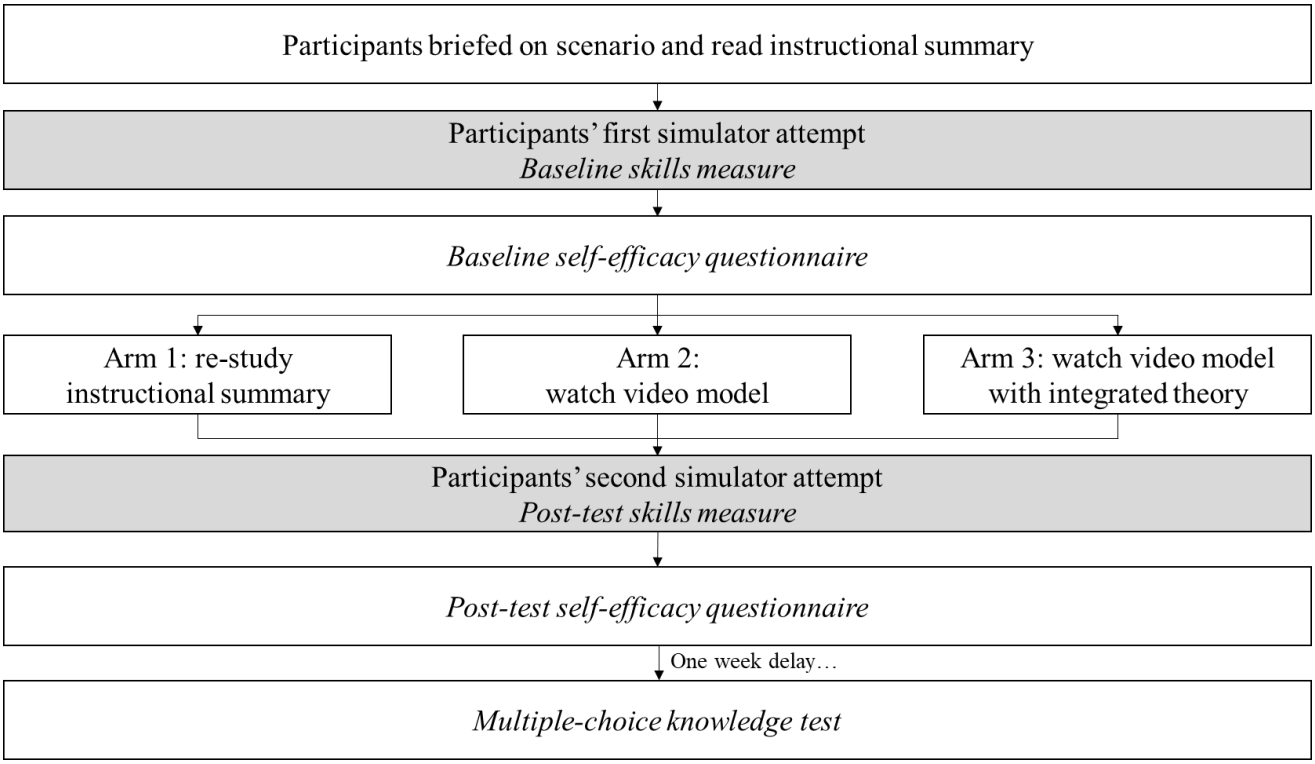


FIGURE 1. *Summary of the experimental design*

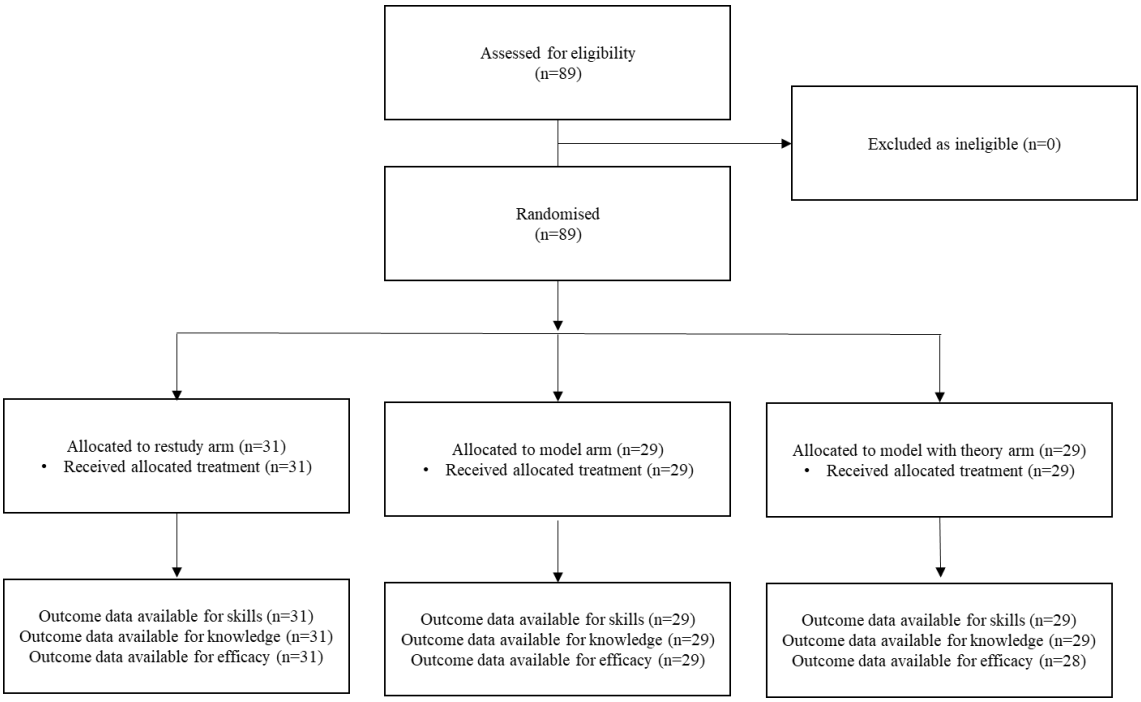


FIGURE 2. *CONSORT diagram*

Analysis

For each of our hypotheses, we begin with a simple graphical presentation of the results before proceeding to formal regression analyses. Multi-arm parallel group trials allow for many

possible pairwise comparisons, which may create problems with multiple hypothesis testing (Juszczak et al., 2019). We therefore aimed to run a parsimonious set of models and tests, focused on testing our study hypotheses. We pre-registered our analysis plan on the Registry of Efficacy and Effectiveness Studies (Registry ID: 14922.1v1 https://sreereg.icpsr.umich.edu/sreereg/subEntry/17401/pdf?section=all&action=download). We conducted a complete case analysis of our data. All analyses were conducted using Stata 17.

To test H1 and H4, we estimate the following model using ordinary least squares regression:

Model 1: $Y_i = \alpha + \beta_1 Model_i + \beta_2 Y_{i,t-1} + \beta'_3 \boldsymbol{X}_i + \varepsilon_i$

Where:

- i indexes individual participants in the experiment
- Y_i is the relevant post-test outcome measure, standardised to have a mean of zero and standard deviation of one
- $Model_i$ is a dummy-coded variable, which takes the value zero for individuals allocated to Arm 1 (*restudy*) or value one for individuals allocated to either Arm 2 (*model*) or Arm 3 (*model with theory*).
- $Y_{i,t-1}$ is our pre-test outcome measure
- $\boldsymbol{X_i}$ is a vector of covariates: female, age, ethnicity
- β_1 provides an estimate of the average effect of allocation to either Arm 2 (*model*) or Arm 3 (*model with theory*), relative to Arm 1 (*restudy*)
- ε_i is a mean zero random error term

Recent work in the econometrics literature has shown that, in experiments with more than two arms, regression coefficients for a given treatment arm may be contaminated by the effects of the other treatment arms (Goldsmith-Pinkham et al., 2022). This is potentially a problem in our trial. However, unbiased estimation of the causal effect across any two treatment arms can still be achieved by dropping participants in the third treatment arm and then running a model with a single treatment dummy variable (Goldsmith-Pinkham et al., 2022). To test H2, we therefore dropped the Arm 1 (*restudy*) participants from the sample and ran the following model:

Model 2: $Y_i = \alpha + \beta_1 Arm3_i + \beta_2 Y_{i,t-1} + \beta_3 X_i + \varepsilon_i$

Where:

- $Arm3_i$ is a dummy, which takes the value one for individuals allocated to Arm 3 (*model with theory*)
- β_1 now captures the effect of allocation to Arm 3 (*model with theory*), relative to Arm 2 (*model*)

Similarly, to test H3, we include the Arm 1 (*restudy*) and Arm 3 (*model with* theory) participants but drop the Arm 2 (*model*) participants from the sample, and then run Model 2. In this case, β_1 captures the effect of allocation to Arm 3 (*model with theory*), relative to Arm 1 (*restudy*).

Results

Hypothesis 1 and 2: teachers' skill in using questioning for retrieval

Our first hypothesis was that exposure to any video model would increase teachers' skills in using questioning for retrieval. The left hand panel of Figure 3 provides a simple graphical presentation of our results. The vertical axis shows the raw sum score on our skills measure, which has a minimum value of zero and a maximum value of 18. The horizontal axis shows the change from the pre-test (first simulator attempt) to the post-test (second simulator attempt). Participants allocated to the *restudy* condition (solid black line) made no measurable improvements in their use of questioning for retrieval between the two simulator attempts. By contrast, participants allocated to either of the two model conditions (dashed line) almost doubled their score (from 6.4 to 11.3) between the two simulator attempts.

Column 1 of Table 2 reports formal regression results. The outcome measure has been constructed to give equal weight to the five different components. It has also been standardised to have mean of zero and standard deviation of one, meaning that the OLS regression coefficients can be interpreted as Cohen's *d* effect sizes. The results show that exposure to the video model improved teachers' use of questioning for retrieval by 0.80 SD, relative to *restudy* (95% CI = 0.39, 1.20). This difference is statistically significant at conventional levels ($p < 0.001$). The model reported in column 2 of Table 2 includes a dummy-coded variable for three of the four experimenters who helped to conduct the experiment. This acts as a check whether the individual who conducted the particular experimental session influenced the outcomes. The coefficient on the *Any Model* is almost unchanged (0.79), as is the R^2, and none of the experimenter dummies are statistically significant at conventional levels. In Column 1 and Column 2 of Table 2, pre-test questioning for retrieval skills also predicted post-test questioning for retrieval skills, but the correlation was quite small (coefficients ranged from 0.29-0.30). This small coefficient likely reflects the fact that participants were in their first term as trainee teachers and the material was therefore new to them.

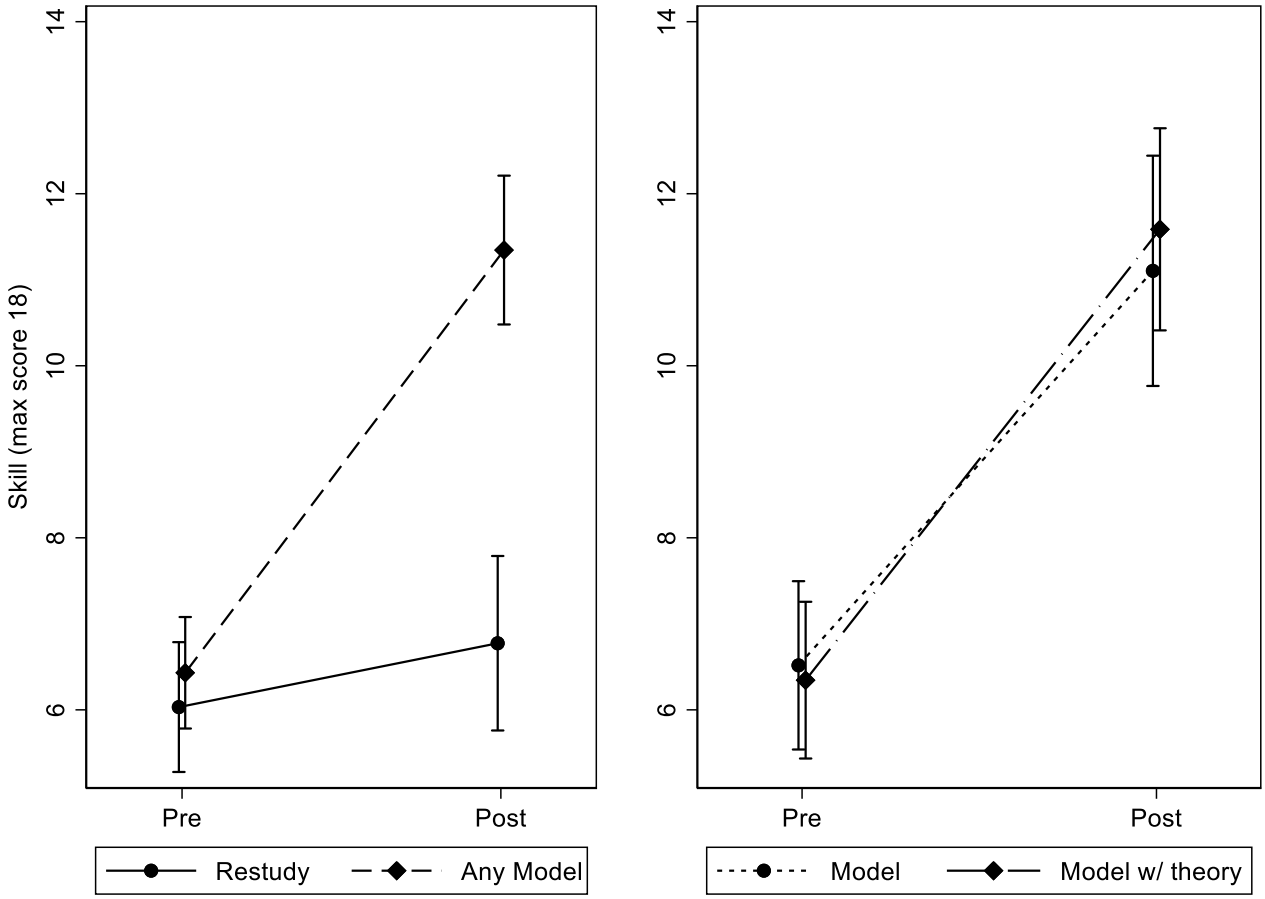


FIGURE 3. *Changes in teacher skills using questioning for retrieval, across treatment arms*

Note. N=89 (left panel) and 58 (right panel). Vertical error bars represent 95% confidence intervals. The measure of skills in using questioning for retrieval on the vertical axis is a raw sum score.

Our second hypothesis was that exposure to a video model incorporating the underlying theory would increase teachers' skills in using questioning for retrieval practice, relative to the simple video model. The right hand panel of Figure 3, which follows the same format as the left hand panel, provides a simple graphical presentation of our results. The vertical axis again shows the raw sum score. Participants allocated to the *model* condition and the *model with theory* condition show very similar improvement between their first and second simulator attempts. Indeed, there is no measurable difference between the two. Column 3 of Table 2 reports formal regression results, which confirm the absence of any statistically significant difference in improvement ($p = 0.477$).

TABLE 2
Modelling the results for teacher skill (Hypotheses 1 and 2)

	(1) Skill in using questioning for retrieval (z score)	(2) Skill in using questioning for retrieval (z score)	(3) Skill in using questioning for retrieval (z score)
Any model (ref: *restudy*)	0.797** (0.203)	0.791** (0.205)	
Model with theory (ref: *model*)			0.184 (0.256)
Pre-test skills	0.295** (0.101)	0.292** (0.105)	0.260* (0.122)
Age	0.006 (0.013)	0.007 (0.014)	-0.013 (0.018)
Female	0.141 (0.255)	0.127 (0.260)	-0.04 (0.379)
Ethnicity: Asian	-0.659 (0.575)	-0.532 (0.604)	-0.616 (0.642)
Ethnicity: Black	0.002 (0.642)	0.122 (0.657)	0.428 (0.738)
Ethnicity: Mixed	-0.705 (0.817)	-0.601 (0.834)	-0.662 (0.867)
Ethnicity: White	-0.368 (0.539)	-0.241 (0.566)	-0.387 (0.581)
Experimenter: 1		0.380 (0.381)	
Experimenter: 2		0.281 (0.346)	
Experimenter: 3		0.050 (0.310)	
Model	Model 1	Model 1~	Model 2
R^2	0.311	0.329	0.171
N	88	88	57

Note. Each column is a separate regression model. Standard errors shown in parentheses. * = $p < 0.05$. ** = $p < 0.01$. ~ Model 1 with the addition of experimenter fixed effect. N = number of participants included in the model. The outcome measure fives equal weight to each of the five components of questioning for retrieval and has been standardised to have a mean of zero and standard deviation of one.

Hypothesis 3 and 4: teacher knowledge and self-efficacy

Our third hypothesis was that exposure to the video with integrated theory would increase teachers' knowledge, relative to restudying the underlying theory. The left hand panel of Figure 3 provides a simple graphical presentation of our results. The vertical axis shows the net score on our knowledge measure, which has a maximum value of 11. Participants exposed to the *restudy* condition (leftmost plot) or *model with theory* condition (rightmost plot) displayed very similar levels of knowledge. Column 1 of Table 3 reports formal regression results. The knowledge outcome measure has again been standardised to have mean of zero and standard deviation of one, meaning that the OLS regression coefficients can be interpreted as Cohen's *d* effect sizes. The results confirm that there was no measurable difference in the levels of knowledge in the two groups ($p = 0.465$).

One potential concern with our delayed knowledge outcome measure is that there may be non-random differences in the delay between groups. The knowledge test was sent to each participant seven days after they participated in the simulator and participants were asked to respond immediately. The median delay in response was indeed seven days in the overall sample, the *restudy* group, and the *model with theory* group. However, the standard deviation in delay in the overall sample was five days. In column 2 of Table 3, we report a sensitivity test in which we include a variable capturing the number of days between participants participation in the simulator and completing the follow-up knowledge test. The coefficient of interest remains non-significant and the coefficient on the delay variable itself is also non-significant ($p = 0.506$).

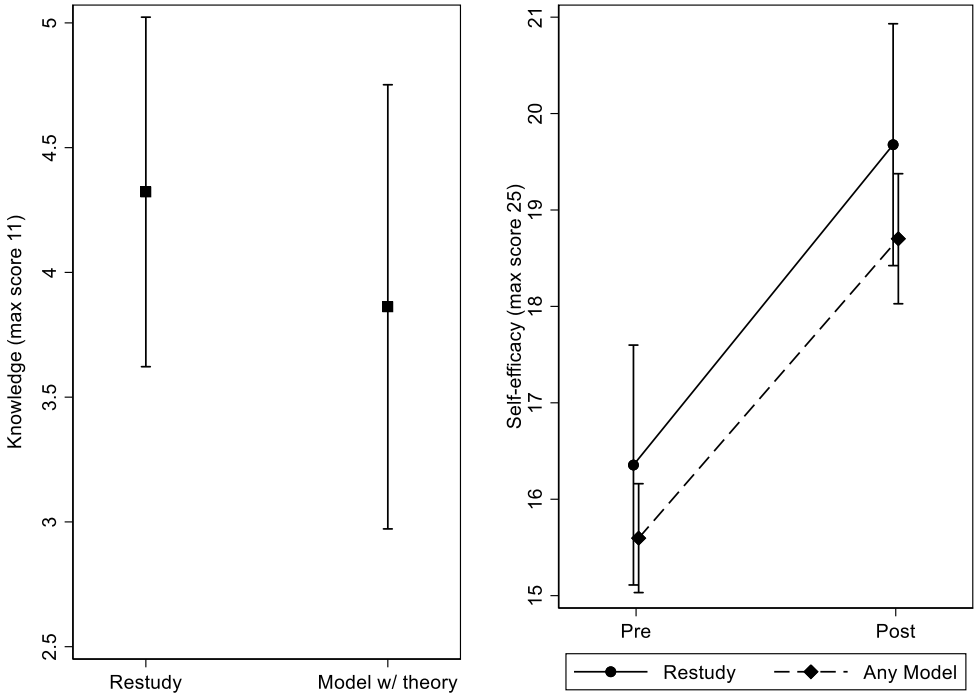


FIGURE 4. *Teacher knowledge and self-efficacy outcomes across treatment arms*

Note. N=60 (left panel) and 88 (right panel). Vertical error bars represent 95% confidence intervals.

Our fourth and final hypothesis was that exposure to any video model would increase teachers' self-efficacy in using questioning for retrieval practice. The right hand panel of Figure 4 provides a simple graphical presentation of our results. The vertical axis shows the raw sum score on our self-efficacy measure, which has a minimum value of zero and a maximum value of 25. The horizontal axis again shows the change from the pre-test (first simulator attempt) to the post-test (second simulator attempt). Participants exposed to either of the two modelling conditions (dashed black line) saw very similar improvements in their self-efficacy to those exposed to the *restudy* condition (solid black line). Column 3 of Table 3 reports formal regression results. The knowledge

outcome measure has again been standardised to have mean of zero and standard deviation of one, meaning that the OLS regression coefficients can be interpreted as effect sizes. The results confirm that there was no measurable difference in the rate at which the two groups improved their self-efficacy ($p = 0.640$).

TABLE 3
Modelling the results for teacher knowledge and self-efficacy outcomes (Hypotheses 3 and 4)

	(1) Knowledge of questioning for retrieval (z score)	(2) Knowledge of questioning for retrieval (z score)	(3) Self-efficacy in using questioning for retrieval (z score)
Model with theory (ref: *restudy*)	-0.191 (0.259)	-0.176 (0.262)	
Any Model (ref: *restudy*)			-0.080 (0.170)
Knowledge test delay (days)		-0.013 (0.021)	
Self-efficacy pre-test			0.704** (0.082)
Age	0.012 (0.017)	0.011 (0.17)	-0.009 (0.011)
Female	0.171 (0.329)	0.178 (0.331)	-0.039 (0.213)
Ethnicity: Asian	0.242 (0.743)	0.026 (0.075)	-0.087 (0.468)
Ethnicity: Black	0.350 (0.838)	0.298 (0.847)	0.356 (0.512)
Ethnicity: Mixed	1.841 (1.182)	1.765 (1.195)	-0.011 (0.662)
Ethnicity: White	0.712 (0.721)	0.657 (0.731)	-0.205 (0.434)
Model	Model 2	Model 2	Model 1
R^2	0.517	0.588	0.525
N	59	59	87

Note. Each column is a separate regression model. Standard errors shown in parentheses. * = $p < 0.05$. ** = $p < 0.01$. N = number of participants included in the model.

Discussion

Models are thought to play an important role in helping teachers notice and attend to important features of teaching practice (Grossman et al., 2009; Kosko et al., 2021). Proponents of models argue that this helps teachers develop a mental image of the focal teaching techniques, which in turn helps them to translate theory into classroom practice (McDonald et al., 2013). However, there is currently no empirical evidence on the causal effects of models on teacher skill development and there is consequently little consensus on whether or how models should be incorporated in teacher professional development. One third of evaluated PD programmes do not incorporate any

models (Sims et al., 2022) and the proportion of non-evaluated PD that do not include modelling is likely higher still (Ofsted, 2023). We set out to provide new evidence on the effects of different types of models on initial teacher trainees' development, in order to better inform teacher educators' design choices.

We found clear evidence that exposure to models improved teachers' skills in the use of evidence-based questioning for retrieval methods, relative to restudying a summary of relevant research. This is the first such causal evidence on the impact of modelling in teacher professional development and represents the primary contribution of this paper. This empirical finding also provides support for two schools of thought on teacher training. First, it supports PBTE theorists' argument that models should be incorporated in initial teacher training. Second, a recent systematic review suggested that modelling is an 'active ingredient' of effective teacher development (Sims et al., 2022). This research provides the first direct causal support for this hypothesis.

By contrast, we did not find that models which clearly labelled and explained the important features of the focal teaching practice resulted in a statistically significant improvement in teachers' skills in the use of questioning for retrieval, relative to a simple video model. When interpreting this finding, it should be kept in mind that all participants had already been exposed to an evidence-based guide that decomposed questioning for retrieval into five constituent parts. It is also interesting to consider our findings on our teacher knowledge and teacher skill outcomes together. Participants in our *model with theory* condition gained more skills (measured within the simulator), and no less knowledge (measured a week later), than participants in our *restudy* condition. This is despite the two conditions spending approximately the same amount of time exposed to the two stimuli.

We did not find that teachers exposed to video models improved their self-efficacy, relative to those who restudied a summary of relevant research. This is somewhat surprising, given that a large body of empirical research has found that modelling supports the development of pre-service teacher self-efficacy (Bautista, 2011; Bautista & Boone, 2015; Gorrell, 1993; Gorrell & Capron, 1990; Palmer, 2006; Palmer, 2011). One potential concern here is that our questionnaire instrument has not previously been shown to be sensitive to changes across a single training session. However, we did in fact detect a statistically significant increase in self-efficacy between the pre- and post-test measurements. Our null finding is instead driven by this increase being of equal magnitude in the modelling and non-modelling groups (Figure 4). This observation is particularly interesting when considered in conjunction with our results on teachers' skills in using questioning for retrieval. Participants exposed to our video models between two simulator attempts improved their skills in the simulator and judged their abilities to have improved accordingly. Participants who restudied the relevant research between two attempts in the simulator judged their abilities to have improved

despite not showing any measurable improvement in these skills. While we can only speculate as to the reasons for this, it may be the case that merely accumulating experience attempting to use questioning for retrieval increased participants' sense of self-efficacy. This intriguing finding should be investigated in further research.

Limitations

Our findings should, of course, be interpreted in light of the limitations of this study. Three in particular stand out. Foremost amongst these is that the research took place within a 'lab' (as opposed to field setting) implemented in a classroom simulator. This has important advantages in terms of statistical power, experimental control, and potential reproducibility (Cohen et al., 2021; Falk & Heckman, 2009). However, there are also important limitations in terms of reduced ecological validity, in particular around the low-stakes nature of the simulator sessions and participant motivation. Our lab-based findings are best interpreted as a test of theory, which can in turn inform the decisions made by teacher educators (Mook, 1983; Sims, 2023; Trafimow, 2022). A second limitation of our research relates to the outcome measures. Our measure of teacher skill is grounded firmly in the empirical literature on questioning for retrieval and showed high inter-rater reliability. However, it has not been previously validated. As more lab experiments are conducted in the domain of teacher education, researchers should prioritise the development and validation of appropriate outcome measures (Hill et al., 2021). A third limitation relates to the statistical precision of our estimates. The 95% confidence intervals of our estimates are quite wide, ranging from 0.33 to 0.51 across our models. While this does not prevent us from detecting a statistically significant effect for modelling (d = 0.8; 95% CI = 0.39, 1.20) it may have hampered our ability to detect a smaller effect, for example in our comparison between the two types of video models (d = 0.18; 95% CI = -0.33, 0.70). In mitigation, the novelty of simulator experiments in education make it hard to estimate power prior to a study and post-hoc power calculations are potentially misleading (Gelman, 2019). As further simulator studies are published, better effect size benchmarks will become available to guide study design.

Implications for teacher educators

Taking into account findings in the existing literature, we believe our results have implications for teacher educators. Crucially, the results from our theoretical tests align with the findings on the importance of modelling from a meta-analysis of evaluations of real-world teacher professional development programmes (Sims et al., 2022). Teacher educators should therefore consider incorporating models into professional development intended to improve teaching skills. Doing so is likely to help trainee teachers put the theory from their course into practice in their classrooms, thus bridging the 'knowing-doing gap' (Knight et al., 2013). Professional development

programmes might consider incorporating libraries of video models exemplifying good practice. There may also be a case for developing publicly available video libraries of video models of evidence-based teaching techniques that are available to all trainees.

Besides the development of recorded models, we see two broad ways in which teacher educators can incorporate models into their work. The first is to provide representations of practice outside of real classroom settings (Grossman, 2018). For example, this might occur during an off-site session or during a focused instructional coaching session. In such cases, trainees can be presented with models focused on specific aspects of teaching practice, isolated from a wider pedagogical sequence. Our results provide direct support for the benefits of this sort of modelling when it comes to developing teacher skills. With this type of modelling, our results suggest that it may not be necessary to label and explain specific aspects of the model, particularly if sufficient decomposition and theorisation of the target teaching practice has occurred prior to viewing the model.

The second way that teacher educators can integrating modelling into their work involves modelling larger lesson sequences in authentic settings, perhaps via co-teaching or lesson observations. Again, our results provide support for this sort of modelling, though clearly this setting is less similar to our experimental setup, so caution is required. In particular, the existing literature suggests that it may be necessary for teacher educators to retrospectively highlight certain aspects of their practice and then explain the rationale for this to the trainee (Eick et al., 2003; Kluth & Straut, 2003). Otherwise, teacher educators run the risk of trainees missing the most valuable aspects of the lesson, or misunderstanding the reasons for their value (Brunvand & Fishman, 2006; van Es & Sherin, 2002; Sherin & van Es, 2005; Rich & Hannafin, 2009). Taking this evidence into account, we do not think our results should be interpreted to mean that labelling and explaining is unnecessary when modelling is occurring as part of a larger authentic lesson sequence and trainees may not have been primed as to what to look for.

Whichever way teacher educators go about incorporating models in their work, they should keep in mind that teachers may need support to reintegrate the specific techniques depicted in these models into the flow of real-world pedagogical and curricular sequences. Only then can teachers realise the value of evidence-based teaching techniques in their own classrooms (Janssen et al., 2015).

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y. & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034–1037.

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y. & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475– 489.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191.

Bandura, A. 1997. *Self-efficacy: The Exercise of Control*. Freeman.

Baroody, A. J. (2003). *The development of adaptive expertise and flexibility: the integration of conceptual and procedural knowledge*. Erlbaum.

Bautista, N. U. (2011). Investigating the use of vicarious and mastery experiences in influencing early childhood education majors' self-efficacy beliefs. *Journal of Science Teacher Education, 22*(4), 333-349.

Bautista, N. U., & Boone, W. J. (2015). Exploring the impact of TeachME™ lab virtual classroom teaching simulation on early childhood education majors' self-efficacy beliefs. *Journal of Science Teacher Education, 26*(3), 237-262.

Booth, J. L., McGinn, K. M., Young, L. K., & Barbieri, C. (2015). Simple practice doesn't always make perfect: Evidence from the worked example effect. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 24-32.

Brunvand, S., & Fishman, B. (2006). Investigating the impact of the availability of scaffolds on preservice teacher noticing and learning from video. *Journal of Educational Technology Systems, 35*(2), 151-174.

Carroll, W. R., & Bandura, A. (1990). Representational guidance of action production in observational learning: A causal analysis. *Journal of Motor Behavior, 22*(1), 85-97.

Cheung, J. J., Kulasegaram, K. M., Woods, N. N., & Brydges, R. (2019). Why content and cognition matter: integrating conceptual knowledge to support simulation-based procedural skills transfer. *Journal of General Internal Medicine, 34*(6), 969-977.

Cheung, J. J., Kulasegaram, K. M., Woods, N. N., & Brydges, R. (2021). Making Concepts Material: A randomized trial exploring simulation as a medium to enhance cognitive integration and transfer of learning. *Simulation in Healthcare, 16*(6), 392-400.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis, 42*(2), 208-231.

Cohen, J., Krishnamachari, A., & Wong, V. C. (2021). *Experimental evidence on the robustness of coaching supports in teacher education* (EdWorkingPaper: 21-468). Annenberg Institute.

Cohen, J., & Wiseman, E. (2019). Approximating complex practice: Teacher simulation of text-based discussion. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Denver, CO.

Cordovani, L. & Cordovani, D. (2016). A literature review on observational learning for medical motor skills and anesthesia teaching. *Advances in Health Sciences Education, 21*(5), 1113–1121.

Custers, E. J., Regehr, G., McCulloch, W., Peniston, C., & Reznick, R. (1999). The effects of modeling on learning a simple surgical procedure: see one, do one or see many, do one?. *Advances in Health Sciences Education, 4*(2), 123-143.

Dallimore, E. J., Hertenstein, J. H., & Platt, M. B. (2013). Impact of cold-calling on student voluntary participation. *Journal of Management Education, 37*(3), 305-341.

Daniel, D. B., & De Bruyckere, P. (2021). Toward an ecological science of teaching. *Canadian Psychology, 62*(4), 361–366.

Eick, C. J., Ware, F. N., & Williams, P. G. (2003). Coteaching in a science methods course: A situated learning model of becoming a teacher. *Journal of Teacher Education, 54*(1), 74-85.
Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535-538.
Ferguson, S., & Sutphin, L. (2022). Analyzing the impact on teacher preparedness as a result of using Mursion as a risk-free microteaching experience for pre-service teachers. *Journal of Educational Technology Systems, 50*(4), 432-447.
Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery, 269*(1), e9-e10.
Goldsmith-Pinkham, P., Hull, P., & Kolesár, M. (2022). C*ontamination bias in linear regressions* (No. w30108). National Bureau of Economic Research.
Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*(3), 606-633.
Gorrell, J., & Capron, E. (1990). Cognitive modeling and self-efficacy: Effects on preservice teachers' learning of teaching strategies. *Journal of Teacher Education, 41*(5), 15-22.
Gorrell, J. (1993). Cognitive modeling and implicit rules: Effects on problem-solving performance. *The American Journal of Psychology, 106*(1), 51-65.
Green, F. (2011). *What is Skill?: An Inter-Disciplinary Synthesis*. London: Centre for Learning and Life Chances in Knowledge Economies and Societies.
Grossman, P. L. (1992). Why models matter: An alternate view on professional growth in teaching. *Review of Educational Research*, *62*(2), 171-179.
Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055-2100.
Grossman, P. (Ed.). (2018). *Teaching core practices in teacher education*. Harvard Education Press.
Han, Y., Syed Ali, S. K. B., & Ji, L. (2022). Use of observational learning to promote motor skill learning in physical education: a systematic review. *International Journal of Environmental Research and Public Health, 19*(16), 10109.
Hauser, M., & Kavanagh, S. S. (2019). Practice-based teacher education. *Oxford Research Encyclopedia of Education*.
Harris, D. J., Vine, S. J., Wilson, M. R., McGrath, J. S., LeBel, M. E. & Buckingham, G. (2018). Action observation for sensorimotor learning in surgery. *Journal of British Surgery, 105*(13), 1713–1720.
Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher, 42*(9), 476-487.
Hill, H. C., Mancenido, Z., & Loeb, S. (2021). Effectiveness research for teacher education. EdWorkingPaper No. 20-252. *Annenberg Institute for School Reform at Brown University*.
Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and instruction, 17*(6), 722-738.
Hoy, A. W., Hoy, W. K., & Davis, H. A. (2009). Teachers' self-efficacy beliefs. In: K. R. Wentzel, & A. Wigfield (Eds.) Handbook of motivation at school (pp. 641-668). Routledge.
Janssen, F., Grossman, P., & Westbroek, H. (2015). Facilitating decomposition and recomposition in practice-based teacher education: The power of modularity. *Teaching and Teacher Education, 51*, 137-146.
Johnson, D. (2010). Learning to teach: the influence of a university-school partnership project on pre-service elementary teachers' efficacy for literacy instruction. *Reading Horizons, 50*(1).
Juszczak, E., Altman, D. G., Hopewell, S., & Schulz, K. (2019). Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. *JAMA, 321*(16), 1610-1620.
Käfer, J., Kuger, S., Klieme, E., & Kunter, M. (2019). The significance of dealing with mistakes for student achievement and motivation: results of doubly latent multilevel analyses. *European Journal of Psychology of Education, 34*(4), 731-753.

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, *62*(2), 129-169.

Kalamar, K. (2018). Questioning techniques that increase student engagement during the mathematics lesson (Doctoral dissertation, Moravian College).

Kavanagh, S. S., Conrad, J., & Dagogo-Jack, S. (2020). From rote to reasoned: Examining the role of pedagogical reasoning in practice-based teacher education. *Teaching and Teacher Education, 89*, 102991.

Kennedy, C., & Mann, C. B. (2017). Randomize: Stata module to create random assignments for experimental trials, including blocking, balance checking, and automated rerandomization. Statistical Software Components S458028, Boston College Department of Economics.

Kennedy, M. M. (1999). The role of preservice teacher education. In Darling-Hammond, L. and Sykes, G. *Teaching as the Learning Profession: Handbook of Teaching and Policy* (pages 54-86). Jossey Bass.

Kennedy, M. M. (2016). How does professional development improve teaching?. *Review of Educational Research, 86*(4), 945-980.

Kluth, P., & Straut, D. (2003). Do as we say and as we do: Teaching and modeling collaborative practice in the university classroom. *Journal of Teacher Education, 54*(3), 228-240.

Knight, B., Turner, D., & Dekkers, J. (2013). The future of the practicum: Addressing the knowing-doing gap. In Lynch, D. E., & Yeigh, T. (Eds). Teacher education in Australia: Investigations into programming, practicum and partnership, 63-76.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283.

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation, 65*, 183-215.

Kosko, K. W., Ferdig, R. E., & Zolfaghari, M. (2021). Preservice teachers' professional noticing when viewing standard and 360 video. *Journal of Teacher Education, 72*(3), 284-297.

Knight, J. (2021). If it Ain't Broke, Handle with Care. Report by the Special Interest Group on Initial Teacher Training (ITT) of the All Party Parliamentary Group for the Teaching Profession. Retrieved from: Microsoft Word - 21.06.30 Report of the ITE SIG to the APPG on the Teaching Profession - PROOFED (JMC4) (wordpress.com)

Kulasegaram, K. M., Martimianakis, M. A., Mylopoulos, M., Whitehead, C. R., & Woods, N. N. (2013). Cognition before curriculum: rethinking the integration of basic science and clinical learning. *Academic Medicine*, *88*(10), 1578-1585.

Labone, E. (2004). Teacher efficacy: Maturing the construct through research in alternative paradigms. *Teaching and Teacher Education, 20*(4), 341-359.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?. *Educational Researcher, 48*(3), 158-166.

Loughran, J. (1995). Practising what I preach: Modelling reflective practice to student teachers. *Research in Science Education, 25*(4), 431-451.

Loughran, J., & Berry, A. (2005). Modelling by teacher educators. *Teaching and Teacher Education, 21*(2), 193-203.

MacSuga-Gage, A. S., & Simonsen, B. (2015). Examining the effects of teacher-directed opportunities to respond on student outcomes: A systematic review of the literature. *Education and Treatment of Children*, 38(2), 211-239.

Mancenido, Z. (2022). *Impact evaluations of teacher preparation practices: challenges and opportunities for more rigorous research* (EdWorkingPaper No. 22-534). Annenberg Institute for School Reform at Brown University.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education, 64*(5), 378-386.

McGarr, O. (2021). The use of virtual simulations in teacher education to develop pre-service teachers' behaviour and classroom management skills: implications for reflective practice. *Journal of Education for Teaching*, *47*(2), 274-286.
McGrew, S., Alston, C. L., & Fogo, B. (2018). Modeling as an example of representation. In: Grossman, P. (Ed.) *Teaching Core Practices in Teacher Education* (pp. 441-463). Harvard Educational Press.
Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465-489.
Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia*, *138*, 107296.
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379-387.
Noble, J. M. (1997). Observational learning: is a picture really worth a thousand words?. University of Colorado at Boulder.
Van Gog, T., Paas, F., Marcus, N., Ayres, P., & Sweller, J. (2009). The mirror neuron system and observational learning: Implications for the effectiveness of dynamic visualizations. *Educational Psychology Review*, *21*(1), 21-30.
Ofsted. (2023). Independent review of teachers' professional development in schools: phase 1 findings. Retrieved from: Independent review of teachers' professional development in schools: phase 1 findings - GOV.UK (www.gov.uk)
Orchard, J., & Winch, C. (2015). What training do teachers need?: Why theory is necessary to good teaching. *Impact*, *2015*(22), 1-43.
Palmer, D. H. (2006). Sources of self-efficacy in a science methods course for primary teacher education students. *Research in Science Education*, *36*(4), 337-353.
Palmer, D. H. (2011). Sources of efficacy information in an inservice program for elementary teachers. *Science Education*, *95*, 577–600.
Perry, T., Lea, R., Jørgensen, C. R., Cordingley, P., Shapiro, K., Youdell, D., ... & Pomareda, C. (2021). *Cognitive science in the classroom*. Education Endowment Foundation.
Rich, P. J., & Hannafin, M. (2009). Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection. *Journal of Teacher Education*, *60*(1), 52-67.
Richardson, J. R., & Lee, T. D. (1999). The effects of proactive and retroactive demonstrations on learning signed letters. *Acta Psychologica*, *101*(1), 79-90.
Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In: Kadosh, R. C., & Dowker, A. (Eds.). Oxford handbook of numerical cognition, 1118-1134.
Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, *27*(4), 587-597.
Saclarides, E. S., & Munson, J. (2021). Exploring the foci and depth of coach-teacher interactions during modeled lessons. *Teaching and Teacher Education*, *105*, 103418.
Sepp, S., Howard, S. J., Tindall-Ford, S., Agostinho, S., & Paas, F. (2019). Cognitive load theory and human movement: Towards an integrated model of working memory. *Educational Psychology Review*, *31*(2), 293-317.
Schunk, D. H., & DiBenedetto, M. K. (2021). Self-efficacy and human motivation. *Advances in Motivation Science*, *8*, 153-179.
Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*(3), 313.
Sherin, M. G., & van Es, E. A. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education*, *13*(3), 475-491.
Sherin, M., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20-37.
Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., ... & Anders, J. (2022). *Effective teacher professional development: new theory and a meta-*

analytic test (EdWorkingPaper No. 22-507). Annenberg Institute for School Reform at Brown University.

Sims, S., Anders, J., Inglis, M., Lortie-Forgues, H., Styles, B., & Weidmann, B. (2023). *Experimental education research: rethinking why, how and when to use random assignment* (No. 23-07). UCL Centre for Education Policy and Equalising Opportunities.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education, 39*(3), 255-267.

Soncini, A., Matteucci, M. C., & Butera, F. (2021). Error handling in the classroom: an experimental study of teachers' strategies to foster positive error climate. *European Journal of Psychology of Education*, *36*(3), 719-738.

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science, 8*(2), 282–293.

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, *16*(2), 165–169.

Tobin, K. (1987). The role of wait time in higher cognitive level learning. Review of *Educational Research, 57*(1), 69-95.

Trafimow, D. (2022). A new way to think about internal and external validity. *Perspectives on Psychological Science*, 17456916221136117.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202-248.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783-805.

Tulis, M. (2013). Error management behavior in classrooms: teachers' responses to student mistakes. *Teaching and Teacher Education, 33*, 56–68.

van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10*(4), 571-596.

Van Kesteren, M. T., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211-219.

Van Kesteren, M. T., Rijpkema, M., Ruiter, D. J., Morris, R. G., & Fernández, G. (2014). Building on prior knowledge: schema-dependent encoding processes relate to academic performance. *Journal of Cognitive Neuroscience*, *26*(10), 2250-2261.

Vaughn, K. E., Fitzgerald, G., Hood, D., Migneault, K., & Krummen, K. (2022). The effect of hint strength on the benefits of retrieval practice. *Applied Cognitive Psychology*, *36*(2), 468-476.

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications*, *4*(1), 35

Weeks, D. L., & Anderson, L. P. (2000). The interaction of observational learning with overt practice: effects on motor skill learning. *Acta Psychologica, 104*(2), 259-271.

Wong, S. S. H., & Lim, S. W. H. (2019). Prevention–permission–promotion: A review of approaches to errors in learning. *Educational Psychologist, 54*(1), 1-19.

Whitcomb, J. A. (2012). Learning and pedagogy in initial teacher preparation. In: Reynolds, W., & Miller, G. (Eds.) *Handbook of Psychology*, *Educational Psychology* (pp. 441-463). Wiley.

Wulf, G., Shea, C., & Lewthwaite, R. (2010). Motor skill learning and performance: a review of influential factors. *Medical Education*, *44*(1), 75-84.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399.

Zeichner, K. (2006). Reflections of a university-based teacher educator on the future of college-and university-based teacher education. *Journal of Teacher Education*, *57*(3), 326-340.

Appendix A: Evidence-based instructional summary

Verbal questioning for retrieval: an evidence-based instructional summary

In this experiment, we're looking at ways to help teachers boost student learning using verbal questioning for retrieval. This instructional summary describes the evidence around how and why this kind of questioning can help. You can find references to the supporting evidence, and a summary diagram, on the final page. We'll be asking you to use these teaching techniques in the simulation once you've finished reading this document.

What is questioning for retrieval?

Retrieval means recall

Retrieval practice refers to "any activity that requires students to recall information from memory rather than representing or restudying the information".[i] **Verbal questioning for retrieval** involves teachers **verbally posing questions to students about content they learned in previous lessons**. When the teacher asks the questions, pupils are prompted to search their memory for the answer, and then retrieve that answer.

Why should teachers use questioning for retrieval?

Retrieval boosts pupil learning – more than restudy

Research shows that **when pupils retrieve knowledge from memory it improves their subsequent retention and transfer of that knowledge**.[ii] This includes research based in primary schools. Retrieval is effective in helping pupils remember both factual knowledge (e.g., Paris is the capital of France) and conceptual knowledge (e.g., a capital city is the location of the seat of government in a country). Crucially, retrieval has been shown to be more effective for improving retention than getting pupils to spend the same amount of time restudying the same material.[iii] Teachers using questioning for retrieval is therefore likely to be a good use of limited classroom time.

Retrieval reminds, highlights, and consolidates memory

Research suggests three main ways in which retrieval helps pupils retain knowledge.[iv] First, by **re-exposing pupils to the material** following the initial learning episode. Second, by posing questions to pupils, retrieval prompts them to **pay attention** to the new knowledge. Third, by prompting pupils to **search for and reactive related prior knowledge**, the memory becomes better consolidated.

How should teachers use questioning for retrieval in the classroom?

Increase impact by… Asking the whole class

Teachers can increase the impact of questioning for retrieval by increasing the number of pupils that attempt to retrieve the prior learning. Here are three ways that teachers can do this:

First, teachers should **ask questions to the entire class**, rather than asking specific pupils. This increases the number of pupils that attempt to retrieve prior learning because it prompts all pupils in the class to search for and retrieve the correct answer from their memory. Research has shown that pupils tend to learn more when their teachers ask a greater number of such whole-class questions.[v]

Waiting three seconds

Second, after a teacher poses a question, they should **wait for at least three seconds before allowing somebody to give the answer**. This ensures that all pupils have sufficient time to attempt independent retrieval, thereby increasing student participation.[vi] Waiting three seconds after posing a question gives all pupils a chance to attempt independent retrieval.[vii] By contrast, if the answer is revealed too quickly, then some pupils will be restudying the material, rather than retrieving it. This is undesirable since restudying is known to be less effective than retrieval.[viii]

Nominating a respondent

Third, teachers should **select pupils to give a response to the question** without regard to who has their hand up. Over time, this increases the number of pupils that attempt to retrieve prior learning because all pupils know that they may be called upon to give an answer. When pupils know that anyone could be asked, pupils tend to show greater engagement and retain more of the content as a result.[ix]

What should teachers do if pupils do not give the correct answer?

Answering wrongly still helps

Calling on pupils who do not have their hand up increases the chances that a pupil will not give the correct answer. This is not a problem, since **even incorrect retrieval improves learning**, as long as it is accompanied by corrective feedback.[x] However, it does raise the question of how teachers should respond to an incorrect answer.

Correct wrong answers

If the pupil gives an incorrect response, teachers should clearly but **gently inform the student that the answer is incorrect and then provide the correct answer**. When a pupil realises that the answer is incorrect, this focuses their attention on the correct answer, which improves subsequent retention.[xi] When teachers give the correct answer, they should explain why this is correct by relating it to pupils' existing knowledge. When teachers elaborate on the correct answer like this, retention is improved.[xii]

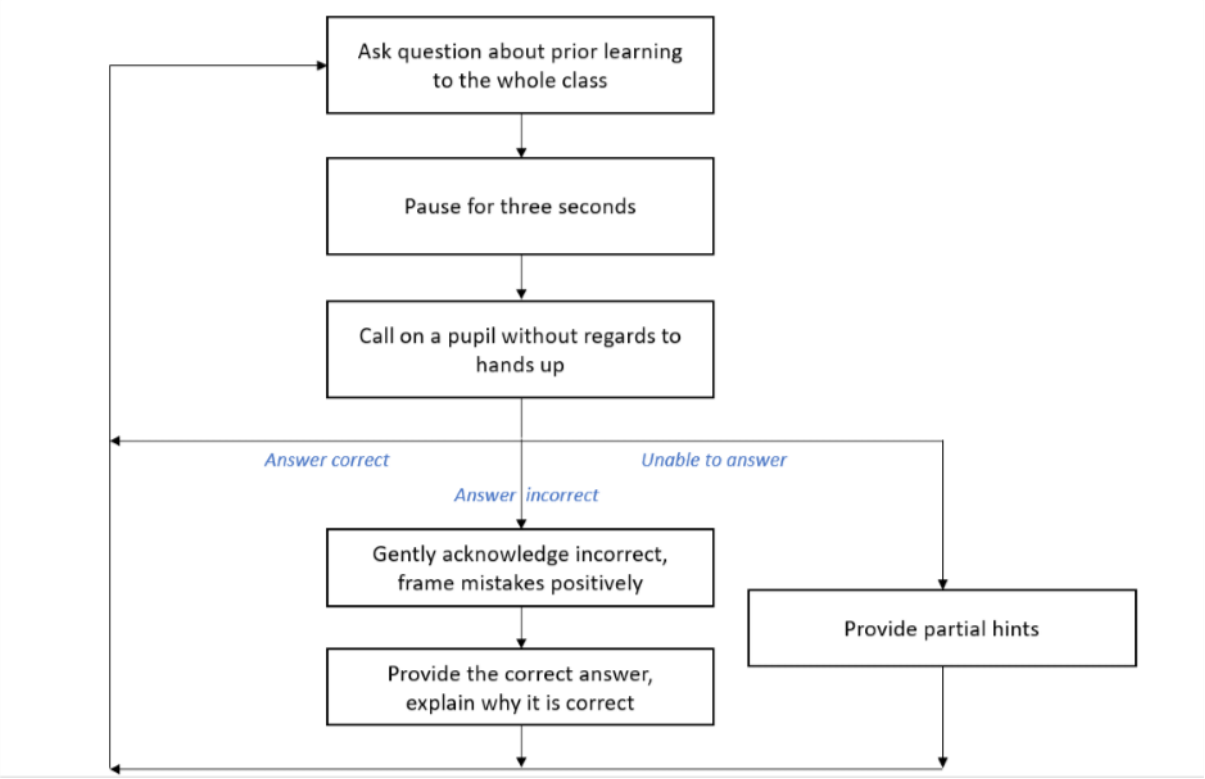
Frame mistakes positively

Pupils may also feel disappointed or embarrassed when they realise they have given an incorrect answer. Teachers should therefore **find a gentle way of letting pupils know that an answer is incorrect**.[xiii] When teachers frame mistakes positively, as something that we can learn from, this results in improved pupil motivation toward learning.[xiv]

Give hints when pupils don't know

Some pupils may not be able to provide an answer to the question at all, stating only that they do not know. In such cases, teachers should proceed to **give the pupil a partial hint**. Research shows that giving progressively more complete hints maximises the extent to which pupils subsequently retain the target knowledge.[xv] This is because it allows pupils to still conduct retrieval for the parts of the answer that are not contained within the hint.

Steps to effective verbal retrieval practice


Ask question about prior learning to the whole class
Pause for three seconds
Call on a pupil without regards to hands up
Answer correct
Unable to answer
Answer incorrect
Gently acknowledge incorrect, frame mistakes positively
Provide the correct answer, explain why it is correct
Provide partial hints

Appendix B: Image of the Mursion simulator interface

FIGURE A1. *Screenshot of the Mursion interface from a pilot session*

Appendix C: Sound unit summary and questions

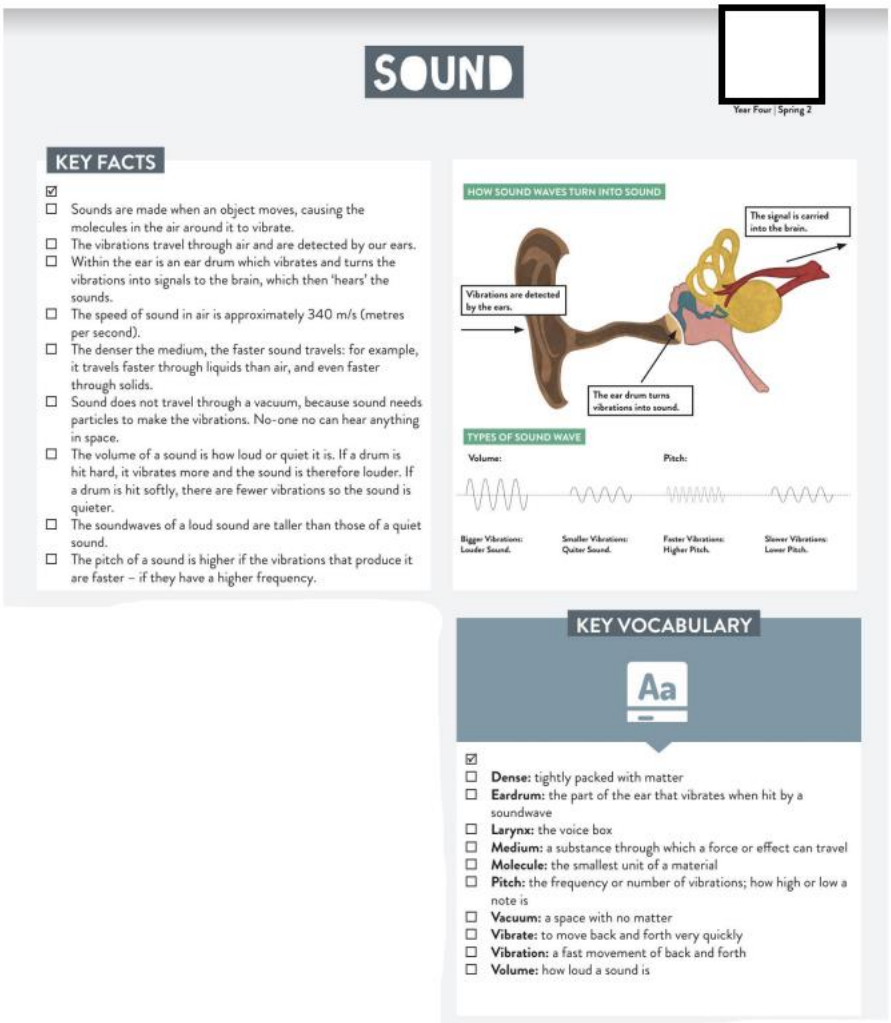


FIGURE B1. *Unit summary graphic shown to participants before simulator session*

	Question	Correct answer
Q1	What is a vibration?	A fast back and forth movement
Q2	What is a medium?	A substance through which a force or effect can travel
Q3	What is the name for the part of our ear that vibrates when it gets hit by a soundwave?	Ear drum
Q4	If a sound is high-pitched, what will the vibrations look like?	Fast vibrations
Q5	What causes a sound to be made?	A moving object, which causes the air/medium around it to vibrate.
Q6	Can sound travel through a vacuum?	No (because there is no air to vibrate)

TABLE B1. *Questions to be asked in the simulator*

Appendix D: Video models

FIGURE D1. *Screenshot of the video, as seen by participants in the Model arm and the Model with Theory arm*

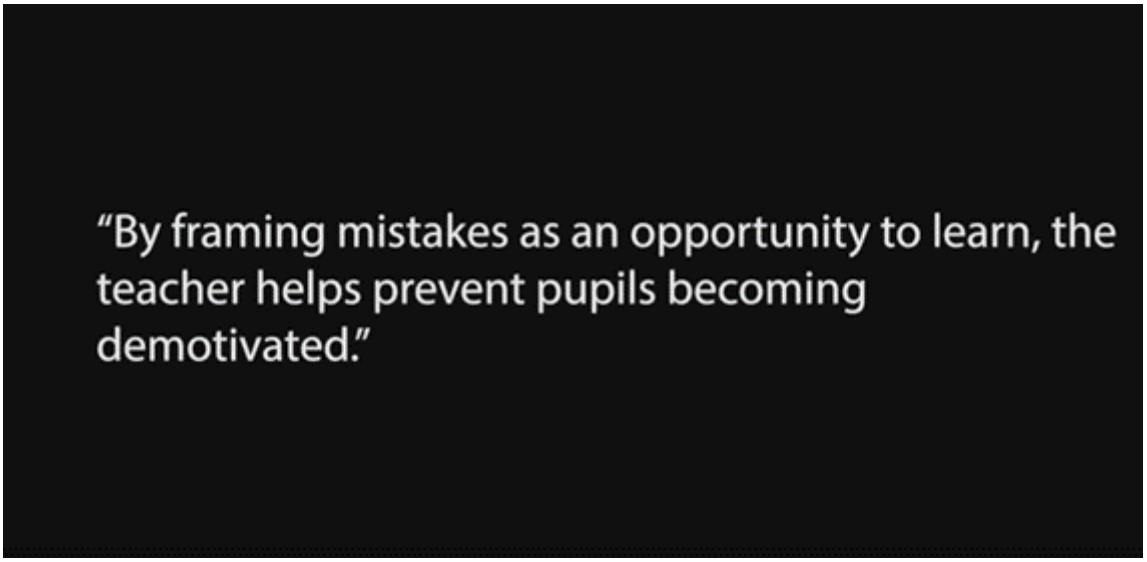


FIGURE D2. *Screenshot of the video in which the footage was interspersed with some of the text from the evidence-based instructional summary, as seen in the Model with Theory arm.*

The full video for the model arm can be found here:
https://estream.dixonsat.com/GetMP4.ashx?ppID=2&file=5043_4m~m4FRphbn.mp4&source=8&bb=0&bt=0&po=0&pi=0&ds=267.04&so=4&st=0&tf=0&cs=OfoMHiJgiJuxulEhcMf9xnK3gGLu63rAFDR3HYMfAP9ZDE72WQNL10A_rO1B_BIsrtIoj03siARVac~f9mACdA

The full video for the model with theory arm can be found here:
https://estream.dixonsat.com/GetMP4.ashx?ppID=2&file=5045_4o~osjSKry7.mp4&source=8&bb=0&bt=0&po=0&pi=0&ds=267.84&so=4&st=0&tf=0&cs=BRBU7KbiNdzCtX4wcWusIZTqP6zbX2du8~cFsQDM7E~nds1cnwT4mpT7uP727152RbbD8idqVTJCoat1rq2mDw

Appendix E: Coding tool

	Rule	**Examples**	**Non-examples**
% questions where all pupils know they could be called upon (maximum 6 points)	If the teacher issues verbal covering statement (across the set of all upcoming questions) that they will be cold calling, then award six marks. Could use 'cold call' language OR equivalent statement (implied/explicit) that student could be asked to respond regardless of hands up. However, if the teacher subsequently violates this for a given question (e.g., nominate-ask) then subtract the mark for that specific question. If no covering statement, then assess on a question-by-question basis, awarding one point per question.	Covering statement: "<u>anyone</u> could be picked for <u>every</u> question", "we're going to go through some questions now, and I will be cold calling" "I want everyone to think about this, and I will be choosing who answers" Question-by-question: "I'm going to cold call you" "Hands down, I will choose somebody to answer" "I might ask anybody to answer this"	Covering statement: "I'm going to ask each of you questions" (doesn't necessarily apply to all questions) "Fred, what is X?" "What is X, Fred?" "Hands up please"
% questions >3 secs wait time (maximum 6 points)	Credit if there are three seconds of time (measured using timer shown on bottom of video) between teacher first asking the question and teacher first nominating somebody to answer. STILL credit if not a cold call Do NOT award if they ask pupils to discuss in pairs within three seconds of asking question. Wait-time is unknown in this case. Do NOT award if hint is given prior to 3-seconds-after question is asked.	Question… nominate Question… other question-related speech… nominate Fred have a think about X … 3 seconds.., what do you think?	Fred…. 3 seconds… question Doesn't count if doing 'talk-partners' or 'turn-and-talk'

% incorrect answers framed as a learning opportunity (maximum 2 points)	If they say this is a positive opportunity and this is EXPLAINED with reference to its potential for subsequent learning	"Don't worry because this is an opportunity to learn" "Don't' worry, we can now work together to get his right"	"Don't worry." / "That's OK." "Don't worry, this is a good thing." "The right answer is…" "Well done for trying."
% incorrect answers, corrected w/ elaboration (maximum 2 points)	The teacher needs to both 1) state or recognise the correct answer and SUBSEQUENTLY 2) relate correct answer to other knowledge (information from the rest of the unit, not included in focal answer or question) OR teacher can state other knowledge IF they immediately state correct answer as a consequence of that knowledge. Do not award if teacher ONLY provides additional information as a hint (i.e. before stating/affirming the correct answer) Do not award if elaboration only comes from pupil, not the teacher.	"Actually, a medium is the thing in between what makes the sound and our ear, like the air" "That's not right, what makes a sound is something that moves and causes the medium around it to vibrate, like our voice box making the air vibrate"	"No. The correct answer is X " with no associated elaboration "Vibrations are the sound, but what makes those vibrations? [Student then responds correctly].
% non-answers given a hint (maximum 2 points)	Award a mark if teacher hears a student say 'I don't know' and then gives same pupil some additional related information, or a separate related question aimed to help infer or support reasoning about correct answer. Hint has to be correct, but can be weak. Do NOT credit if hint is merely restating the same question.	"It's a type of musical instrument" "What would vibrate in a vacuum?" "Do you remember when we learned about what a vacuum is?"	Rephrasing the question with no additional information. Encouragement with no information. Providing all the information.

Appendix F: Test instrument

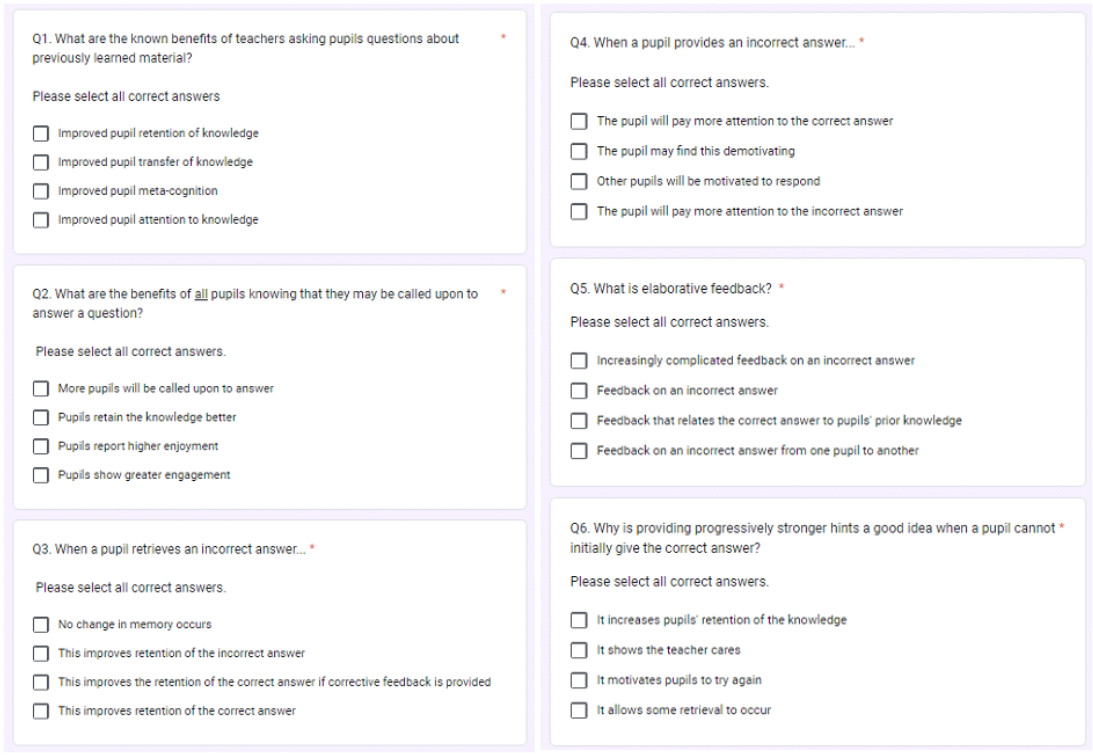
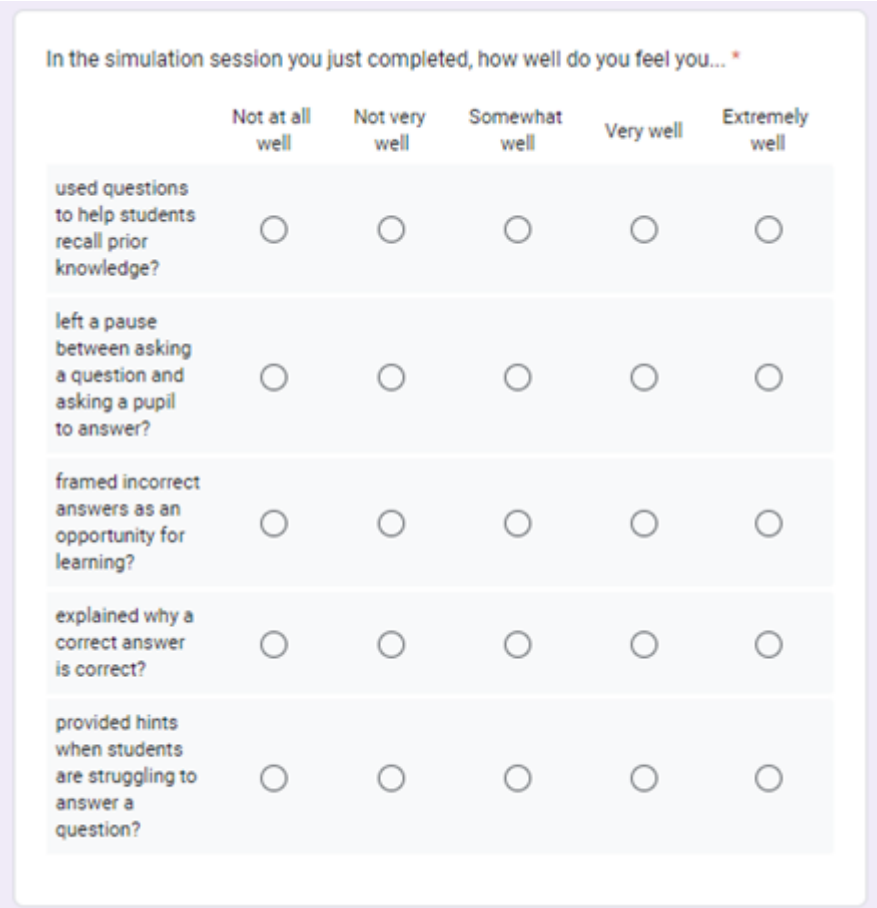


FIGURE G1. *Screenshot of the test instrument*

Appendix G: Efficacy questionnaire



In the simulation session you just completed, how well do you feel you... *

	Not at all well	Not very well	Somewhat well	Very well	Extremely well
used questions to help students recall prior knowledge?	○	○	○	○	○
left a pause between asking a question and asking a pupil to answer?	○	○	○	○	○
framed incorrect answers as an opportunity for learning?	○	○	○	○	○
explained why a correct answer is correct?	○	○	○	○	○
provided hints when students are struggling to answer a question?	○	○	○	○	○

FIGURE F1. *Screenshot of the adapted efficacy questionnaire instrument*