Instructional Science (2026) 54:32
https://doi.org/10.1007/s11251-026-09779-2

ORIGINAL RESEARCH

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

Sam Sims[1,2] · **Harry Fletcher-Wood**[1] · **Thomas Godfrey-Faussett**[1,3] · **Peps Mccrea**[1] · **Stefanie Meliss**[1,4]

Received: 18 June 2024 / Accepted: 22 January 2026

Abstract
Teacher training often incorporates observable examples of focal teaching practices – models. Yet there is currently little empirical evidence on the effects of modelling. We tested the effects of video models on trainees' skills, knowledge, and self-efficacy in relation to using an evidence-based teaching technique: retrieval practice. We recruited 89 first-year trainee teachers, gave them a document containing evidence-based guidance on how to use retrieval practice and then collected pre-test data on how well they were able to do this in a classroom simulator scenario. Participants were then randomised them to one of three groups: an active control group in which they restudied the document (no model), a video model of effective practice, or a similar video model annotated with the underpinning theory. We then collected post-test data in a second simulator exercise. Exposure to video models improved participants' use of retrieval practice methods relative to no model. However, adding the annotation to the models did not yield additional benefits. Models did not improve teachers' knowledge or self-efficacy. Findings support the theory that incorporating models in initial teacher training can help new teachers make use of evidence-based teaching practices.

Keywords Teachers · Professional development · Modelling · Simulation-based training

Introduction

Research is increasingly providing educators with evidence-based approaches to teaching and learning (Weinstein et al., 2018). However, there is a long running debate about whether initial (pre-service) teacher education is designed in a way that helps teachers put this evi-

✉ Sam Sims
sam.sims@ambition.org.uk

1 Ambition Institute, 156 Caledonian Rd, London N1 9UU, UK

2 UCL, London, UK

3 University of Oxford, London, UK

4 University of Reading, Berkshire, England

Published online: 03 March 2026

dence to use in 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 with the specific teaching skills needed to put these ideas into practice. Relatedly, Kennedy (1999) has argued that knowledge of an idea or theory substantially underdetermines what teachers should do to put that theory to use in the classroom. Knight (2013) refers to this succinctly as the 'knowing-doing gap'.

How should teacher educators address this challenge? Multiple solutions have been suggested in the literature. For example, researchers have advocated for using case-based methods (Darling-Hammond & Hammerness, 2002), observations of more experienced teachers' lessons (Jenkins, 2014), co-teaching with a more experienced teacher (Eick et al., 2003), simulation exercises in which teachers observe each other rehearsing (De Coninck et al., 2019; Frei-Landau et al., 2022), or watching video exemplars of specific teaching practices (Allen et al., 2011, 2015). One thing that these approaches typically have in common is that they incorporate models - observable examples of teaching practice. Models can be 'live' in that they are delivered in person (e.g., through co-teaching) or 'symbolic' in that they are captured in an image (e.g., in a video library). Around two thirds of evaluated in-service professional development programmes incorporate modelling (Sims et al., 2025) and around one fifth of teachers in England report that their recent professional development 'always' or 'often' involved modelling (Ofsted, 2023).

In line with this, 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; Kluth & Straut, 2003; Loughran, 1995; Loughran & Berry, 2005; Saclarides & Munson, 2021). This literature has illuminated the role and significance of modelling in allowing teacher educators to help new teachers develop a mental image of some particular practice and help draw new teachers' attention to the most important aspects of this practice (McDonald et al., 2013). This in turn opens up opportunities to connect those practices to the underlying pedagogical reasoning (Gibbons & Cobb, 2017; Saclarides & Munson, 2021), either through discussion or annotation of the model (McFadden et al., 2014).

Yet evidence on the effects of modelling in teacher education remains scarce. For example, a recent systematic review of teacher preparation practices does not appear to include any impact evaluations of modelling (Mancenido, 2024). 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. Importantly, 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). For example, some models can be annotated or talked over, which can help to reveal the principles underlying different pedagogical moves or the reasoning behind teachers' decisions. 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 (Daniel & De Bruyckere, 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 and colleagues (2020). This generally involves asking a participant to take part in two simulator exercises, while randomly allocating them to receive differing input between the two exercises. We use this to test the

impact of 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).* In doing so, we provide the first experimental test of the theory that modelling helps teachers develop skills in using evidence-based teaching practices. In addition, we provide new evidence on the value of annotating video models to make explicit 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.

Literature review and hypotheses

Modelling plays an important part in multiple theories in psychology and teacher development. Representations of teaching practices (of which models are one type) play a prominent role in Practice Based Teacher Education, where they are theorised to help pre-service teachers notice important features of teaching practice (Grossman, 1992; Grossman et al., 2009; Hauser & Kavanagh, 2019; Kosko et al., 2021). Bandura's social learning and social cognitive theory emphasises that observation of others is core to the way in which people learn new skills, helps to build a sense of self-efficacy in the learner and helps accelerates learning compared to personal trial and error (Bandura & Walters, 1977). Thus, teacher educators have recommended that teachers have much to learn from observing more experienced colleagues (Myers, 1978). Cognitive theorists have also emphasised learning from models, arguing that they more efficiently encode information than verbal descriptions of practices (Renkl, 2014; Sepp et al., 2019) and can therefore assist teachers in comprehending and, ultimately, emulating new practices (Sims et al., 2025). This research adopts a pluralistic theoretical approach, drawing on the practice-based, social learning and cognitive theories of modelling to develop a number of hypotheses (see below).

Modelling and skills

Skills are improvable abilities to perform actions that bring about a socially desirable outcome (Green, 2011). 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 et al., 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). Models provide a cognitively efficient way of communicating practice in the sense that *a picture is worth a thousand words* (Noble, 1997). This cognitive theory aligns with the observation of teacher educators that models help trainee teachers develop a mental 'image' of the focal teaching practice (McDonald et al., 2013), which can then be used to guide their practice.

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). While these studies are not drawn from the field of teacher education, we think the medical education field is similar enough to lend empirical support for our first hypothesis:

> 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 (Brunvand & Fishman, 2006; Sherin & van Es, 2005; van Es & Sherin, 2002). 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 - including those working in Practice Based Teacher Education - 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 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; Hoogerheide & Sepp, 2024). 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 & Schneider, 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., 2010; 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 (Woods 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, 2021). Reasoning by analogy with the math literature, and in line with the empirical evidence from 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

Theorists working in the social learning / social cognitive tradition have argued that 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 (Hoy et al., 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; Hoogerheide et al., 2016; Schunk & Hanson, 1985; Warner & French, 2020). 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 et al., 1998).

Qualitative studies have carefully illuminated the links between modelling and pre-service teacher self-efficacy (Palmer, 2006, 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, 1993; Gorrell & Capron, 1990). 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 decided to focus on video (rather than live) models on the basis that they are commonly used in teacher development (e.g., in the My Teaching Partner and Steplab programmes), have received considerable interest in the teacher education literature (Derry et al., 2014), and allowed us to use the exact same model for all participants in each arm of the experiment, thus maintaining experimental control. We decided to focus on models of an expert teacher demonstrating evidence-based practice, rather than video recording of the participants own practice. This was appropriate given the lack of experience of the new teachers participating in our study and was theoretically necessary to test our fourth hypothesis, relating to self-efficacy.

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). A large body of research shows that retrieval practice improves pupil learning of both factual and conceptual knowledge (for reviews of this evidence, see Kornell & Vaughn, 2016; Yang et al., 2021). This makes it highly appropriate content for initial teacher training. Indeed, all new teachers in England are now required to learn about retrieval practice (DfE, 2024). Questioning for retrieval – the focus of our study - 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*). The first arm should be considered as an active control. This study was granted ethical approval by the UCL Institute of Education Research Ethics Committee.

Methods

Participants and design

We aimed to recruit 30 participants for each of the three arms in our experiment and ultimately recruited 89 participants. Assuming an R^2 of 0.75, this gives us power of 0.8 to detect an effect of 0.4 or higher for any pairwise contrast between our three experimental arms. Our sample size is comparable to those in previous simulator experiments, which were

able to detect effects on a range of outcomes. Individuals were eligible to participate in the experiment if they had enrolled on a primary (elementary) initial teacher training course in England in the 2022/23 academic year. These are pre-service teachers who have not yet graduated from initial teacher training. Recruitment opened on 1st of October 2022 and closed on 23rd December 2022. We recruited participants by approaching multiple 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 drawn from multiple initial teacher training providers. Table 1 provides a breakdown of participants demographic characteristics 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 a classroom simulator experiment, which are becoming increasingly common in this literature (Cohen et al., 2020, 2024; Cohen & Wiseman, 2019). 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 because of the more proximal and theoretically-aligned outcomes measures that they tend to employ (Hill et al., 2021; Sims et al., 2023). The simulated environment also allows researchers to exercise tight experimental control by building the simulation scenarios in ways that standardise e.g., the number of opportunities that arise for using retrieval practice. We used the Mursion simulator environment (Cohen et al., 2020; Ferguson & Sutphin, 2022) implemented within an 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 Figure S5). The Mursion environment was appropriate for our study because it allowed us to control how the avatar pupils responded to the research participants. For example, when asked questions, pupils could retrieve either correct or incorrect information, thus requiring different responses from the participating teachers.

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
Self-perceived performance 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. 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

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, 2017). 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 ($p = 0.720$). 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.

Procedure and stimuli

The experiment was conducted entirely online using Zoom video conferencing software. This allowed the participants to take part remotely, allowed for video capture of the sessions, and allowed for easy communication between the participants and researchers. 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 Figures S1-4 in the Supplementary Materials. 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 et al., 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 et al., 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 et al., 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 et al., 2022; Vaughn & Kornell, 2019).

Participants were given as long as they needed to read the document from start to finish. 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 use the term 'simulator task' to describe what the participants were asked to do within the simulator environment. 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 Figure S6), 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 (Table S1). We provided the human simulation specialist with a script detailing how to respond to the teacher's questions (Table S1 – see Supplementary Materials). 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. Participants continued in the simulator until they had asked all six questions.

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 min 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. In line with the theory and evidence underpinning hypotheses 1, this video contained moving images of people, which efficiently communicate how the evidence-based practices can be used, thus supporting participants to develop a mental image, which can then be used to guide their practice (Höffler & Leutner, 2007; McDonald et al., 2013; Noble, 1997; Sepp et al., 2019; van Gog et al., 2009; Wulf et al., 2010). In line with theory and evidence underpinning hypothesis 4, the video models provide an opportunity to watch another early career teacher successfully using the teaching techniques, which should improve efficacy (Bandura, 1977; Labone, 2004; Hoogerheide et al., 2016; Schunk & Hanson, 1985; Schunk & DiBenedetto, 2021; Warner & French, 2020).

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.' Five such statements were included in the Arm 3 video.

In line with the theory and evidence behind hypotheses 2 above, these text snippets were included to highlight and thus draw attention to the most important aspects of the model (Goodwin, 1994, p. 606), as well as making explicit the rationale for specific techniques demonstrated in the model (Goodwin, 1994; McGrew et al., 2018; Sherin & van Es, 2009; Rich & Hannafin, 2009). In line with the theory and evidence underpinning hypothesis 3, displaying the procedural knowledge (video) alongside the declarative knowledge (written statements) in these models was intended to connect the two types of knowledge and thus increase the chances of that knowledge being retained (Cheung et al., 2019, 2021; van Kesteren et al., 2010; Kesteren et al., 2014; Kulasegaram et al., 2013; Rittle-Johnson & Schneider, 2015).

Both the Arm 2 and Arm 3 videos were 4.5 min long. Screenshots of the videos, and links to the full videos online, are available in Figures S7-8. Following this, all participants had a second attempt at the exact same simulator task. All participants completed the procedure described in this section in 30 min or less.

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 edited before being sent to the coders so that the coders could not tell which treatment arm the participant had been allocated to. 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 determined whether the teacher framed errors as learning opportunities. Our coding framework includes rules for awarding credit based on five principles, supported by examples of creditworthy and non-creditworthy responses. This framework was refined through pilot simulator sessions before the main experiment (see Table S2). 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 α

across all the indicators was 0.84. Further descriptives and psychometric information can be found in Table S3B. 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 Table S4). 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 was 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 the supplementary materials and descriptive statistics, and psychometric information is available in Table S3C. 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). Collecting our post-test measure with a seven-day delay was necessary to assess knowledge retention. We decided not to collect a pre-test measure of our teacher knowledge outcome. We judged that a pre-test measure collected prior to participants' exposure to the instructional summary would likely have shown floor effects because the material would likely be entirely unfamiliar to many of our early-stage trainee participants. When we piloted the study, we tried collecting a knowledge measure immediately after exposing participants to the instructional summary but before we applied the randomised treatment. However, we found clear ceiling effects, with many participants getting the maximum score. Removing the ceiling effects by making the test harder was challenging, given the limited amount of content in the written summary.

We captured participants' self-efficacy using an adapted version of the Teacher Self-Efficacy Questionnaire (TSEQ; Tschannen-Moran & Hoy, 2001). We used the TSEQ as the basis for our measure on the grounds that it has been extensively validated and is widely used in this literature. Following the second simulator attempt, we asked participants to reflect on the session they just had using the same stem 'Following that second simulation session, how well do you feel you *can*…'. For example, for principle 5, we asked 'how well do you feel you *can*… *provide* hints when students are struggling to answer a question?' The post-test measure hence measures participants judgement of their capabilities to execute the task (self-efficacy). Responses were collected on a five-point Likert scale ranging from 1 = 'Not at all well' to 5 = 'Extremely well'. Cronbach's α across the five items was 0.78 (see Supplementary Table S3A for further information). We calculated an overall score using confirmatory factor analysis.

Following the first simulator attempt, we asked participants to reflect on the simulator session they had just completed and used the stem 'In the simulation session you just

completed, how well do you feel you'.[1] This stem was again 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?' As such, the pre-test measure captures participants' subjective judgement on how well they executed the task. Since this captures a slightly different construct to self-efficacy, we refer to this pre-test questionnaire as our self-perceived performance pre-test. Descriptive statistics for the self-perceived performance pre-test are included in Table 1.

The overall design of the experiment, including stimuli, measures, and treatment arms, is summarised in Fig. 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.

Analysis

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.

[1] While the specific questions were framed as being related to first simulator attempt, the pre-test self-efficacy measure was embedded in a larger questionnaire instrument, which also captured background information about participants. When we asked the participants to respond to the overall questionnaire instrument, we framed it as relating to their teaching generally.

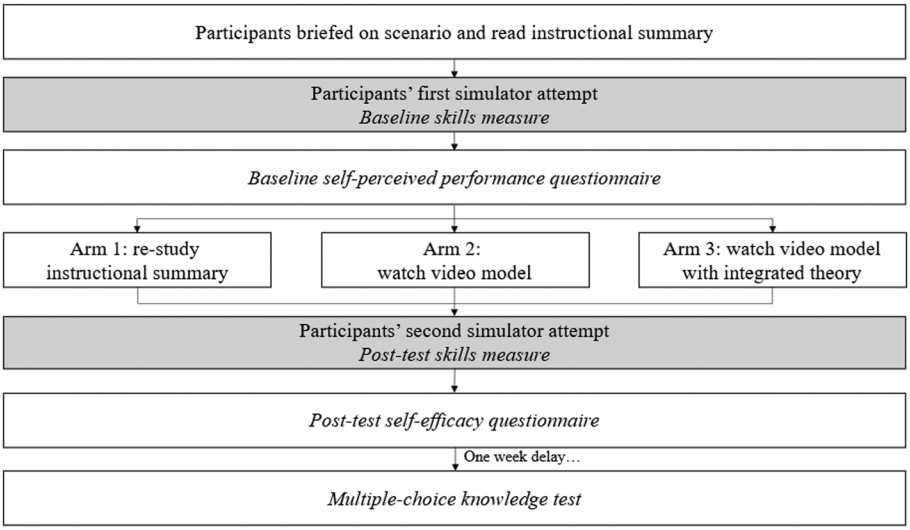

Fig. 1 *Summary of the experimental design*

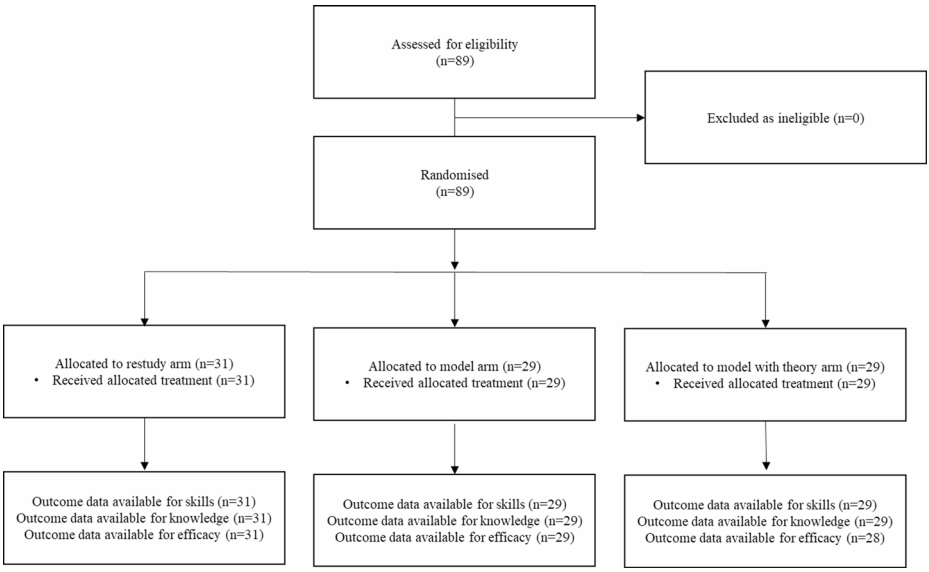

Fig. 2 Consort *diagram*

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. Pairwise correlations between all continuous variables can be found in Table S4 in the supplementary materials.

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

Model 1: $Y_i = \alpha + \beta_1 Model_i + \beta_2 Y_{i,t-1} + \beta_3 \boldsymbol{X}_i + \epsilon_i$

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 treat-

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

Where:

- $Arm3_i$ is a dummy-coded variable, 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 Fig. 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. This means that the ordinary least squares (OLS) regression coefficients are measured in terms of standard deviations and the coefficients on our binary treatment variable can be interpreted as a 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 SD), as is the R^2, and none of the experimenter dummy-coded variables 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 to 0.30

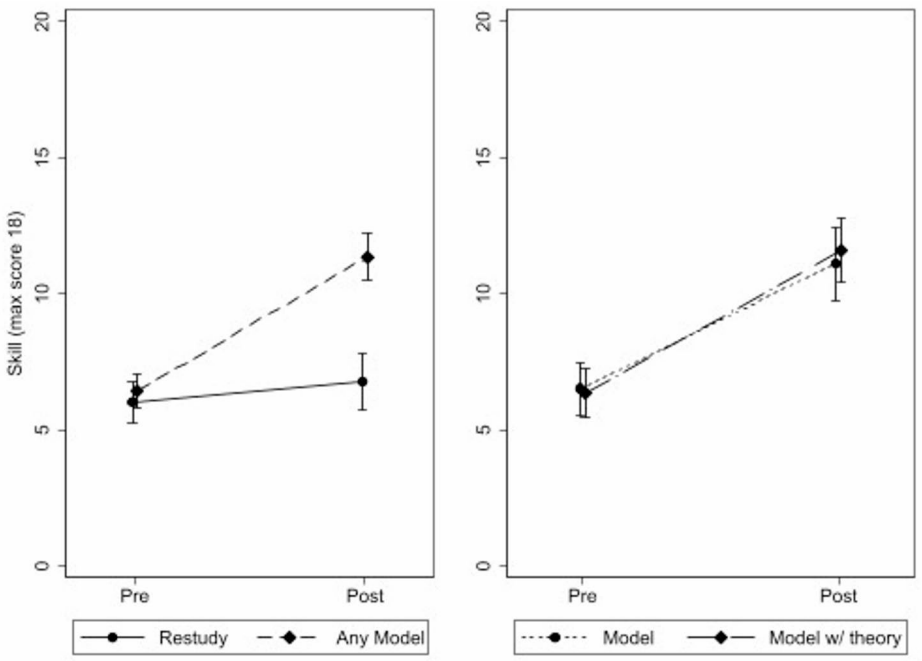

Fig. 3 Changes in teacher skills using questioning for retrieval, across treatment arms. *Note*. $N = 89$ (left panel) and 58 (right panel). Vertical error bars represent 95% confidence intervals. The measure of skill in using questioning for retrieval on the vertical axis is a raw sum score

SD). 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.

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 Fig. 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 ($p = 0.477$).

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. Column 1 of Table 3 reports 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 ($p = 0.465$).

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**	0.791**	
	(0.203)	(0.205)	
Model with theory (ref: *model*)			0.184
			(0.256)
Pre-test skills	0.295**	0.292**	0.260*
	(0.101)	(0.105)	(0.122)
Age	0.006	0.007	-0.013
	(0.013)	(0.014)	(0.018)
Female	0.141	0.127	-0.04
	(0.255)	(0.260)	(0.379)
Ethnicity: Asian	-0.659	-0.532	-0.616
	(0.575)	(0.604)	(0.642)
Ethnicity: Black	0.002	0.122	0.428
	(0.642)	(0.657)	(0.738)
Ethnicity: Mixed	-0.705	-0.601	-0.662
	(0.817)	(0.834)	(0.867)
Ethnicity: White	-0.368	-0.241	-0.387
	(0.539)	(0.566)	(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
Breusch-Pagan	$p=0.67$	$p=0.66$	$p=0.39$
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 gives 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. Ref = reference category

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). This provides some reassurance that our results do not reflect differential delays in responding to the delayed knowledge post-test across our three experimental groups.

Our fourth and final hypothesis was that exposure to any video model would increase teachers' self-efficacy in using questioning for retrieval practice. 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 coef-

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-perceived performance 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
Breusch-Pagan	$p=0.37$	$p=0.46$	$p=0.30$
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. Ref = reference category

ficients 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 ($p = 0.640$). [2]

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 experimental 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., 2023) and the proportion of non-evaluated PD that do not include modelling is likely higher still (Ofsted, 2023).

[2] While we analysed the data using the overall score calculated using confirmatory factor analysis, the sum score was used for plotting to simplify the interpretation of the plot.

We set out to provide new evidence on the effects of different types of models on initial teacher trainees' development, 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. A number of other studies have advocated for the importance of modelling in teacher education (Moore & Bell, 2019) or provided evidence that it correlates with improvements in outcomes for trainee teachers (Sims et al., 2025). However, ours is the first study to isolate the effect of modelling in teacher professional development, and this constitutes the primary contribution of this paper. Several other papers have evaluated modelling as part of a wider package of teacher educator practices (Allen et al., 2011, 2015), including in lab experiments and classroom simulator studies (Cohen et al., 2020; Mancenido et al., 2025). The findings from our study suggest that modelling may have been an important part of what made those interventions effective.

This empirical finding also provides support for two schools of thought on teacher training. First, it supports Practice Based Teacher Education theorists' argument that models (or 'representations' of practice) should be incorporated in initial teacher training. Second, it supports a recent systematic review suggested that modelling is an 'active ingredient' of effective professional development for in-service teachers (Sims et al., 2023). The present research provides the first direct experimental support for the claims made about modelling in both of these theoretical frameworks.

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, relative to a simple video model. This appears to run counter to the recommendations of various researchers who have advocated for teacher educators to label and explain video models (e.g., Brunvand & Fishman, 2006). Having said that, readers should keep in mind that all participants had already been exposed to an evidence-based guide that decomposed questioning for retrieval into five constituent parts. Lunenberg et al. (2007) and Moore and Bell (2019) distinguish between implicit models (in which teacher educators do not explicitly highlight best practices), explicit models (in which they do highlight best practices) and explicit models with connection to theory (where they both highlight and explain best practices). One way of thinking about the lack of any statistically significant difference between our two modelling conditions is therefore to note that both are examples of explicit models with connection to theory. The only difference is that in the *model* arm the explicitness and theory comes before the model, whereas in the *model with theory* arm the explicitness and theory comes before and during the model. Our study shows that this appears not to make a big difference when the delay between the two is relatively short (i.e., a matter of minutes). We return to discuss the case of implicit models in the implications section below.

We also 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 (Gorrell, 1993; Gorrell & Capron, 1990; Palmer, 2006, 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 (see limitation discussed below). Our null finding is instead driven by this increase being of equal magnitude in the modelling and non-modelling groups.

We also did not find that modelling improved knowledge. Our data does not suggest that this reflects ceiling effects or differential delays in responding to the knowledge post-test. Again, while we can only speculate, it seems plausible that the lack of an effect here stems from important differences between our domain (teacher education) and the domains in which the existing evidence comes from: maths education and medical education. For example, it may be that the clinical skills (such as surgery) in medical education are more directly visually connected to the underlying medical knowledge (such as anatomy) than is the case in teacher education. Our null findings for self-efficacy and knowledge – which both contrast with literature from other domains – are both deserving of being investigated further in future research.

Limitations

Our findings should, of course, be interpreted in light of the limitations of this study. Four stand out. Foremost amongst these is that the research took place within a 'lab' setting in an online classroom simulator, rather than out in the field. This has important advantages in terms of statistical power, experimental control, and potential reproducibility (Cohen et al., 2024; Falk & Heckman, 2009). However, there are also important limitations in terms of reduced ecological validity. For example, the simulator is an online virtual environment and the low-stakes nature of this setting may have influenced participants' 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 et al., 2023; Trafimow, 2023).

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 makes 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.

Fourth, and finally, our experimental design has limited empirical scope. We studied teachers in one country (England), in one particular career stage (initial or pre-service training), teaching one subject (primary/elementary science), using one teaching technique (questioning for retrieval). It is plausible to think that models might be less effective for more experienced teachers, who have more accumulated experiences of seeing others teach and may have better developed mental models of teaching in general. Similarly, some teaching areas of teaching, such as assessment or lesson planning, likely lend themselves less

well to visual models. Ultimately, further empirical research will be necessary to understand the generalisability of these findings.

Implications for teacher educators

These limitations notwithstanding, our findings suggest that teacher educators should consider making use of models to help early-career teachers develop evidence-based pedagogical practices. Doing so may help trainee teachers put the theory from their course into practice in their classrooms, thus helping to bridge the 'knowing-doing gap' (Knight et al., 2013). As in our experimental setup, this requires teacher educators to combine models illustrating how something should be *done* with theory to help teachers understand (or *know*) what it is they are aiming to do. Teacher educators might therefore consider developing libraries of video models exemplifying good practice, to accompany the pedagogical theory that they cover on their courses. Teacher educators should also keep in mind that teachers may need further support to reintegrate the specific techniques depicted in these models into the flow of real-world pedagogical sequences (Janssen et al., 2015; Banks et al.,2024).

Besides the development of recorded models, we see two broad ways in which teacher educators can incorporate live models into their work. The first is to provide live representations of teaching 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 support for the benefits of this sort of modelling when it comes to developing teacher skills. With this type of focused 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 integrate modelling into their work involves modelling larger lesson sequences in authentic classroom settings, perhaps via co-teaching or lesson observations. Again, we interpret our results as providing support for this type of less focused modelling. Having said that, one important difference the focused models used in our study and less focused classroom models is that, in that latter, teachers may miss the most valuable aspects of the model, or misunderstanding the reasons for their value (Brunvand & Fishman, 2006; Rich & Hannafin, 2009; Sherin & van Es, 2005; van Es & Sherin, 2002). Indeed, 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). Taking this evidence into account, we do not think our results (based on focused models) should be interpreted to mean that labelling and explaining is unnecessary when using less focused models.

Conclusions

Providing observable examples of teaching practice (models) can help teachers bridge the gap between their theoretical knowledge of pedagogical practices and their ability to enact them in practice. Where appropriate, teacher educators should consider using simple video

or live models to complement the more theoretical content on their courses. This is likely to help early-career teachers to develop evidence-based teaching practices.

Acknowledgements We thank all participants who donated their valuable time to make this research project possible. The project was funded by Ambition Institute and funding for this project was acquired by Hilary Spencer, Marie Hamer and Jennifer Barker. Thanks to Adam Cunningham and the rest of the staff at Connect Training for their support in running the simulator sessions. Thanks to Abigail Brown who supported the management of this project. Thanks to Jennifer Barker for supervising the project. Finally, thanks to Sarah Cottingham, Steve Farndon, Nick Pointer, Nick Rose, Susan Dutta, Anna Bartkiewicz, Gorana Henry, Tessa Willy, Jemima Rhys-Evans, Laura Senior, Jade Pearce, Andy Kay, Rachel Cook and the staff and pupils at Dixons Academies Trust.

CRediT author statement **Sam Sims**: Administration, Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualisation, Writing – original draft. Resources. **Harry Fletcher-Wood**: Administration, Conceptualisation, Investigation, Project administration, Software, Validation, Visualisation, Writing – review & editing. Resources. **Thomas Godfrey-Faussett**: Administration, Conceptualisation, Investigation, Project administration, Software, Validation, Visualisation, Writing – review & editing. Resources. **Peps Mccrea**: Conceptualisation, Writing – review & editing. Supervision. Funding acquisition. **Stefanie Meliss**: Investigation, Software, Validation, Writing – review & editing.

Ethics declarations

Competing interests This work was supported by Ambition Institute. All five of the authors work for organisations that provide teacher training in return for fees. All authors declare no other competing interest.

Research involving human participants This study was granted ethical approval by the UCL Institute of Education Research Ethics Committee.

Informed consent All participants provided informed consent.

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. https://doi.org/10.1126/science.1207998

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. https://doi.org/10.1080/19345747.2015.1017680

Banks, B., Sims, S., Curran, J., Meliss, S., Chowdhury, N., Altunbas, H., ... & Instone, I. (2024). Decomposition and recomposition in teacher education. UCL Centre for Education Policy and Equalising Opportunities No. 24-08.

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., ... & Anders, J. (2025). Effective teacher professional development: New theory and a meta-analytic test. *Review of Educational Research, 95*(2), 213-254.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W H Freeman.

Bandura, A., & Walters, R. H. (1977). *Social learning theory*. Prentice hall.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody, & A. Dowker (Eds.), *The development of arithmetic concepts and skills* (pp. 1–33). Routledge.

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. https://doi.org/10.1177/2372732215601691

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. https://doi.org/10.2190/L353-X356-72W7-42L9

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. https://doi.org/10.1080/00222895.1990.10735503

Cheung, J. J. H., 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. https://doi.org/10.1007/s11606-019-04959-y

Cheung, J. J. H., 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. https://doi.org/10.1097/SIH.0000000000000543

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

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231. https://doi.org/10.3102/0162373720906217

Cohen, J., Wong, V. C., Krishnamachari, A., & Erickson, S. (2024). Experimental evidence on the robustness of coaching supports in teacher education. *Educational Researcher*, *53*(1), 19–35. https://doi.org/10.3102/0013189X231198827

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. https://doi.org/10.1007/s10459-015-9646-5

Custers, E. J. F. M., 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. https://doi.org/10.1023/A:1009763210212

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. https://doi.org/10.1177/1052562912446067

Daniel, D. B., & De Bruyckere, P. (2021). Toward an ecological science of teaching. *Canadian Psychology / Psychologie Canadienne*, *62*(4), 361–366. https://doi.org/10.1037/cap0000291

Darling-Hammond, L., & Hammerness, K. (2002). Toward a pedagogy of cases in teacher education. *Teaching Education*, *13*(2), 125–135.

De Coninck, K., Valcke, M., Ophalvens, I., & Vanderlinde, R. (2019). Bridging the theory-practice gap in teacher education: The design and construction of simulation-based learning environments. Kohärenz in der Lehrerbildung: Theorien, Modelle und empirische Befunde, 263–280.

Derry, S. J., Sherin, M. G., & Sherin, B. L. (2014). Multimedia learning with video. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 785–812). Cambridge University Press.

DfE [Department for Education] (2024). Initial Teacher Training and Early Career Framework.

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. https://doi.org/10.1177/0022487102238659

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244

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.

Frei-Landau, R., Orland-Barak, L., & Muchnick-Rozonov, Y. (2022). What's in it for the observer? Mimetic aspects of learning through observation in simulation-based learning in teacher education. *Teaching and Teacher Education*, *113*, 103682.

Gelman, A. (2019). Don't calculate Post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1), e9. https://doi.org/10.1097/SLA.0000000000002908

Gibbons, L. K., & Cobb, P. (2017). Focusing on teacher learning opportunities to identify potentially productive coaching activities. *Journal of Teacher Education*, *68*(4), 411–425.

Goldsmith-Pinkham, P., Hull, P., & Kolesár, M. (2022). *Contamination Bias in Linear Regressions (Working Paper 30108)*. National bureau of economic research. https://doi.org/10.3386/w30108

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633. https://doi.org/10.1525/aa.1994.96.3.02a00100

Gorrell, J. (1993). Cognitive modeling and implicit rules: Effects on problem-solving performance. *The American Journal of Psychology*, *106*(1), 51–65. https://doi.org/10.2307/1422865

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. https://doi.org/10.1177/002248719004100503

Green, F. (2011). *What is Skill? An Inter-Disciplinary Synthesis*. Centre for Learning and Life Chances in Knowledge Economies and Societies. https://www.llakes.ac.uk/publication/what-is-skill-an-inter-disciplinary-synthesis/

Grossman, P. L. (1992). Why models matter: An alternate view on professional growth in teaching. *Review of Educational Research*, *62*(2), 171–179. https://doi.org/10.3102/00346543062002171

Grossman, P. L. (Ed.). (2018). *Teaching core practices in teacher education*. Harvard Education.

Grossman, P. L., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching: Theory and Practice*, *15*(2), 273–289. https://doi.org/10.1080/13540600902875340

Han, Y., Ali, S., 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), Article 16. https://doi.org/10.3390/ijerph191610109

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. *British Journal of Surgery*, *105*(13), 1713–1720. https://doi.org/10.1002/bjs.10991

Hauser, M., & Kavanagh, S. S. (2019). Practice-Based Teacher Education. In *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.419

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, Crossroads, and challenges. *Educational Researcher*, *42*(9), 476–487. https://doi.org/10.3102/0013189X13512674

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. https://doi.org/10.26300/zhhb-j781

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, *17*(6), 722–738. https://doi.org/10.1016/j.learninstruc.2007.09.013

Hoogerheide, V., & Sepp, S. (2024). Six Evidence-Informed tips on how to optimize learning from instructional videos. In Gegenfurtner, A., & Kollar, I. (Eds.), *Designing effective digital learning environments* (pp. 75–89). Routledge.

Hoogerheide, V., Loyens, S. M., & van Gog, T. (2016). Learning from video modeling examples: Does gender matter? *Instructional Science*, *44*, 69–86.

Hoy, W. A., Hoy, W. K., & Davis, H. A. (2009). Teachers' Self-Efficacy beliefs. In Wentzel, K. (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. https://doi.org/10.1016/j.tate.2015.06.009

Jenkins, J. M. (2014). Pre-service teachers' observations of experienced teachers. *Physical Educator*, *71*(2), 303.

Johnson, D. (2010). Learning to teach: The influence of a University-School partnership project on Pre-Service elementary efficacy for literacy instructionteachers'. *Reading Horizons*, *50*(1), 23-33. https://scholarworks.wmich.edu/reading_horizons/vol50/iss1/4

Juszczak, E., Altman, D. G., Hopewell, S., & Schulz, K. (2019). Reporting of Multi-Arm Parallel-Group randomized trials: Extension of the CONSORT 2010 statement. *Journal of the American Medical Association*, *321*(16), 1610–1620. https://doi.org/10.1001/jama.2019.3087

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. https://doi.org/10.1007/s10212-018-0408-7

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, *62*(2), 129–169. https://doi.org/10.3102/00346543062002129

Kalamar, K. (2018). *Questioning techniques that increase student engagement during the mathematics lesson* [PhD Thesis]. Moravian College.

Kennedy, M. M. (1999). The role of preservice teacher education. In L. Darling-Hammond, & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of teaching and policy* (pp. 54–86). Jossey Bass.

Kennedy, C., & Mann, C. B. (2017). *RANDOMIZE: Stata module to create random assignments for experimental trials, including blocking, balance checking, and automated rerandomization* [Computer software]. https://econpapers.repec.org/software/bocbocode/s458028.htm

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. https://doi.org/10.1177/0022487103054003005

Knight, J. (2021). *If it ain't broke, handle with care: Report by the Special Interest Group on Initial Teacher Training (ITT) All Party Parliamentary Group for the Teaching Profession*. IRIS Press. https://research.leedstrinity.ac.uk/en/publications/if-it-aint-broke-handle-with-care-report-by-the-special-interest-

Knight, B., Turner, D., & Dekkers, J. (2013). The future of the practicum: Addressing the knowing-doing gap. In D. E. Lynch, & T. Yeigh (Eds.), *Teacher education in australia: Investigations into Programming, practicum and partnership* (pp. 63–76). Oxford Global. https://hdl.handle.net/10018/1014453

Kornell, N., & Vaughn, K. E. (2016). Chapter Five - How Retrieval Attempts Affect Learning: A Review and Synthesis. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Academic Press. https://doi.org/10.1016/bs.plm.2016.03.003

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), 283–294. https://doi.org/10.1037/a0037850

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. https://doi.org/10.1177/0022487120939544

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. https://doi.org/10.1097/ACM.0b013e3182a45def

Labone, E. (2004). Teacher efficacy: Maturing the construct through research in alternative paradigms. *Teaching and Teacher Education*, *20*(4), 341–359. https://doi.org/10.1016/j.tate.2004.02.013

(

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

Loughran, J. (1995). Practising what I preach: Modelling reflective practice to student teachers. *Research in Science Education*, *25*(4), 431–451. https://doi.org/10.1007/BF02357386

Loughran, J., & Berry, A. (2005). Modelling by teacher educators. *Teaching and Teacher Education*, *21*(2), 193–203. https://doi.org/10.1016/j.tate.2004.12.005

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 & Treatment of Children*, *38*(2), 211–240. https://psycnet.apa.org/record/2015-26532-004

Mancenido, Z. (2024). Impact evaluations of teacher Preparation practices: Challenges and opportunities for more rigorous research. *Review of Educational Research*, *94*(2), 268–307. https://doi.org/10.3102/00346543231174413

Mancenido, Z., Hill, H. C., Garcia Coppersmith, J., Carter, H., Pollard, C., & Monschauer, C. (2025). Practice-Based teacher education pedagogies improve responsiveness: Evidence from a lab experiment. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2025.2456716

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. https://doi.org/10.1177/0022487113493807

McFadden, J., Ellis, J., Anwar, T., & Roehrig, G. (2014). Beginning science teachers' use of a digital video annotation tool to promote reflective practices. *Journal of Science Education and Technology*, *23*, 458–470.

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. https://doi.org/10.1080/02607476.2020.1733398

McGrew, S., Alston, C., & Fogo, B. (2018). Modeling as an example of representation. In P. L. Grossman (Ed.), *Teaching core practices in teacher education* (pp. 33–55). Harvard Education.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia*, *138*, 107296. https://doi.org/10.1016/j.neuropsychologia.2019.107296

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Moore, E. J., & Bell, S. M. (2019). Is instructor (faculty) modeling an effective practice for teacher education? Insights and supports for new research. *Action in Teacher Education*, *41*(4), 325–343.

Myers, R. E. (1978). *An analysis and synthesis of experimental teacher training studies (1963–1974) which*. tested elements of Bandura's social learning theory.

Noble, J. M. (1997). *Observational learning: Is a picture really worth a thousand words?* [PhD Thesis]. University of Colorado at Boulder.

Ofsted (2023). *Independent review of teachers' professional development in schools: Phase 1 findings*. GOV. UK. https://www.gov.uk/government/publications/teachers-professional-development-in-schools/independent-review-of-teachers-professional-development-in-schools-phase-1-findings

Orchard, J., & Winch, C. (2015). What training do teachers need? Why theory is necessary to good teaching. *Impact*, *2015*(22), 1–43. https://doi.org/10.1111/2048-416X.2015.12002.x

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. https://doi.org/10.1007/s11165-005-9007-0

Palmer, D. H. (2011). Sources of efficacy information in an inservice program for elementary teachers. *Science Education*, *95*(4), 577–600. https://doi.org/10.1002/sce.20434

Perry, T., Lea, R., Rübner Jørgensen, C., Cordingley, P., Shapiro, K., & Youdell, D. (2021). *Cognitive science in the classroom: Evidence and practice review*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/cognitive-science-approaches-in-the-classroom/

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. https://doi.org/10.1177/0022487108328486

Richardson, J. R., & Lee, T. D. (1999). The effects of proactive and retroactive demonstrations on learning signed letters. *Acta Psychologica*, *101*(1), 79–90. https://doi.org/10.1016/S0001-6918(98)00046-8

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. Cohen, Kadosh, & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 1118–1134). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.014

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. https://doi.org/10.1007/s10648-015-9302-x

Saclarides, E. S., & Munson, J. (2021). Exploring the foci and depth of coach-teacher interactions during modeled lessons. *Teaching and Teacher Education*, *105*, 103418. https://doi.org/10.1016/j.tate.2021.103418

Schunk, D. H., & DiBenedetto, M. K. (2021). Chapter Four—Self-efficacy and human motivation. In A. J. Elliot (Ed.), *Advances in Motivation Science* (Vol. 8, pp. 153–179). Elsevier. https://doi.org/10.1016/bs.adms.2020.10.001

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*(3), 313–322. https://doi.org/10.1037/0022-0663.77.3.313

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. https://doi.org/10.1007/s10648-019-09461-9

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. G., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20–37. https://doi.org/10.1177/0022487108328155

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 (Working Paper No. 23–07)* (23–07). UCL Centre for Education Policy and Equalising Opportunities. https://EconPapers.repec.org/RePEc:ucl:cepeow:23-07

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J., & Anders, J. (2025). Effective teacher professional development: New theory and a Meta-Analytic test. *Review of Educational Research*, *95*(2), 213–254. https://doi.org/10.3102/00346543231217480

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. https://doi.org/10.1007/s10212-020-00494-1

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of education sciences. *International Journal of Research & Method in Education*, *39*(3), 255–267. https://doi.org/10.1080/1743727X.2016.1150454

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science*, *8*(2), 282–293. https://doi.org/10.1037/tps0000332

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, *16*(2), 165–169. https://doi.org/10.1016/j.learninstruc.2006.02.005

Tobin, K. (1987). The role of wait time in higher cognitive level learning. *Review of Educational Research*, *57*(1), 69–95. https://doi.org/10.3102/00346543057001069

Trafimow, D. (2023). A new way to think about internal and external validity. *Perspectives on Psychological Science*, *18*(5), 1028–1046. https://doi.org/10.1177/17456916221136117

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, *68*(2), 202–248. https://doi.org/10.3102/00346543068002202

Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to student mistakes. *Teaching and Teacher Education*, *33*, 56–68. https://doi.org/10.1016/j.tate.2013.02.003

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 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. https://doi.org/10.1007/s10648-008-9094-3

van Kesteren, M. T. R., Fernandez, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences*, *107*(16), 7550–7555. https://doi.org/10.1073/pnas.0914892107

van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., Morris, R. G. M., & Fernández, G. (2014). Building on prior knowledge: Schema-dependent encoding processes relate to academic performance. *Journal of Cognitive Neuroscience*, *26*(10), 2250–2261. https://doi.org/10.1162/jocn_a_00630

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications*, *4*(1), 35. https://doi.org/10.1186/s41235-019-0187-y

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. https://doi.org/10.1002/acp.3929

Warner, L. M., & French, D. P. (2020). Self-efficacy interventions. In M. S. Hagger, L. D. Cameron, K. Hamilton, N. Hankonen, & T. Lintunen (Eds.), *The handbook of behavior change* (pp. 461–478). Cambridge University Press.

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. https://doi.org/10.1016/S0001-6918(00)00039-1

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, *3*(1), 2. https://doi.org/10.1186/s41235-017-0087-y

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. https://doi.org/10.1080/00461520.2018.1501693

Woods, N. N., Brooks, L. R., & Norman, G. R. (2007). It all makes sense: Biomedical knowledge, causal connections and memory in the novice diagnostician. *Advances in Health Sciences Education*, *12*(4), 405–415. https://doi.org/10.1007/s10459-006-9055-x

Wulf, G., Shea, C., & Lewthwaite, R. (2010). Motor skill learning and performance: A review of influential factors. *Medical Education*, *44*(1), 75–84. https://doi.org/10.1111/j.1365-2923.2009.03421.x

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), 399–435. https://doi.org/10.1037/bul0000309

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*, 326–340. https://doi.org/10.1177/0022487105285893