






# Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48

Michael L. Green, Jeremy J. Moeller & Judy M. Spak



To cite this article: Michael L. Green, Jeremy J. Moeller & Judy M. Spak (2018): Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48, Medical Teacher, DOI: [10.1080/0142159X.2018.1430354](https://doi.org/10.1080/0142159X.2018.1430354)

To link to this article: <https://doi.org/10.1080/0142159X.2018.1430354>

 View supplementary material 

 Published online: 01 Feb 2018.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 

## Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48

Michael L. Green<sup>a</sup>, Jeremy J. Moeller<sup>b</sup> and Judy M. Spak<sup>c</sup>

<sup>a</sup>Department of Internal Medicine and Teaching and Learning Center, Yale School of Medicine, New Haven, CT, USA; <sup>b</sup>Department of Neurology, Yale School of Medicine, New Haven, CT, USA; <sup>c</sup>Cushing-Whitney Medical Library, Yale School of Medicine, New Haven, CT, USA

### ABSTRACT

**Background:** Cognitive psychology studies demonstrate that subjects who attempt to recall information show better learning, retention, and transfer than subjects who spend the same time studying the same material (test-enhanced learning, TEL). We systematically reviewed TEL interventions in health professions education.

**Methods:** We searched 13 databases, 14 medical education journals, and reference lists. Inclusion criteria included controlled studies of TEL that compared TEL to studying the same material or to a different TEL strategy. Two raters screened articles for inclusion, abstracted information, determined quality scores, and calculated the standardized mean difference (SMD) for the learning outcomes.

**Results:** Inter-rater agreement was excellent for all comparisons. The 19 included studies reported 41 outcomes with data sufficient to determine a SMD. TEL interventions included short answer questions, multiple choice questions, simulation, and standardized patients. Five of six *immediate learning* outcomes (SMD 0.09–0.44), 21 of 23 *retention* outcomes (SMD 0.12–2.5), and all seven *transfer* outcomes (SMD 0.33–1.1) favored TEL over studying.

**Conclusions:** TEL demonstrates robust effects across health professions, learners, TEL formats, and learning outcomes. The effectiveness of TEL extends beyond knowledge assessed by examinations to clinical applications. Educators should include TEL in health professions curricula to enhance recall, retention, and transfer.

### Introduction

Educators commonly think of assessment “of” learning. At the end of a course of study, students recall information previously learned through studying. Recently, educators have turned their attention to assessment “for” learning (Schuwirth and Van der Vleuten 2011), considering assessment as a pedagogical strategy in and of itself. Looming assessments *indirectly* enhance learning by driving students’ study behaviors (rehearsal effect) (Fitch et al. 1951; Newble and Jaeger 1983). This effect also operates after assessments as students receive the results and direct further study in areas of poor performance.

Assessment also *directly* enhances learning. Studies in cognitive psychology laboratories (Roediger and Butler 2011; Karpicke and Grimaldi 2012) and classrooms (Roediger et al. 2011; Agarwal et al. 2012) consistently demonstrate that recalling previously learned information (retrieval practice) enhances the ability to recall the information in the future (retrieval effect). Students who engage in effortful, deliberate attempts to recall information show better learning, retention, and transfer than students who spend the same time repeatedly studying the same material. This effect is also known as “test-enhanced learning (TEL)” when the retrieval practice occurs in the context of a test. The effect size for retrieval practice in laboratory settings and primary, secondary, and post-secondary classrooms has been estimated to be 0.5 (Rowland 2014; Adesope et al. 2017).

Several test formats enhance the retrieval effect. Repeated testing promotes better recall than a long single

### Practice points

- TEL demonstrates consistent and robust effects across different health professions, learner levels, TEL formats, and learning outcomes.
- The effectiveness of TEL extends beyond knowledge assessed by examinations to clinical applications.
- Educators should include TEL in health professions curricula to enhance recall, retention, and transfer. Ideally, “tests” should be repeated, spaced over time, utilize items that require production of information, and include feedback with the correct responses and rationale.
- Students should consider retrieval practice as an effective study strategy.

test (Wheeler and Roediger 1992; Karpicke and Roediger 2008). Spacing tests over time intervals results in better recall compared with back-to-back consecutive testing (Landauer and Bjork 1978; Cepeda et al. 2006; Karpicke and Roediger 2007). These effects persists at spacing intervals from one minute to 30 days. Equal spacing intervals produce better long-term retention than expanding intervals. Items that require production of information (short answer, essay) perform better than items that require recognition of information (multiple choice, true/false). (McDaniel, Anderson, et al. 2007; McDaniel, Roediger, et al. 2007; Pyc and Rawson 2009) Students who receive feedback recall

more information and enjoy enhanced metacognitive awareness (Bangert-Drowns et al. 1991; Butler et al. 2008).

These findings have implications for learners. College undergraduates appear to lack an awareness of the retrieval effect as they most commonly employ rereading as a learning strategy and very rarely engage in retrieval practice (Karpicke et al. 2009). Incorporating TEL may thus improve learning outcomes as well as help students develop study strategies that aid in recall. Such results may be particularly important in health professions education, where the volume of information is high and accurate recall has important implications for patient safety.

More recently, investigators have demonstrated the retrieval effect in health professions education. Trainees allocated to testing (versus studying) demonstrate superior medical knowledge (Larsen et al. 2009; Messineo et al. 2015) (by multiple choice and short answer tests) and better skills [cardiopulmonary resuscitation (Kromann et al. 2009, 2010) and radiograph interpretation (Baghdady et al. 2014)], with effects lasting up to 6 months. The standardized effect size of TEL has been estimated at 0.9, indicating large practical importance (Kreiter et al. 2013).

Previous reviews of TEL were not systematic (Larsen et al. 2008; Roediger and Butler 2011; Kreiter et al. 2013; Augustin 2014; Brame and Biel 2015; Yeh and Park 2015) and/or confined to cognitive psychology studies (Roediger and Butler 2011; Eisenkraemer et al. 2013; Brame and Biel 2015; Yeh and Park 2015). Herein we report a systematic review of TEL in the health professions. As reflected in the review questions below, we document the effectiveness of TEL in enhancing learning (including recall, retention, and transfer); the magnitude of this effect; and differential effects with different settings, test formats and timing, and co-interventions.

Medical educators commonly understand that assessment indirectly enhances learning by driving study behavior. Our review should raise their awareness of the potent direct effects of assessment (retrieval effect) and help them design learning and assessments strategies that maximize learning, retention, and transfer. Furthermore, our review reveals gaps in TEL science that illuminate directions for medical education research.

## Methods

### Review questions

Our systematic review addressed the question: In health professions students or providers (Subjects), does TEL (Intervention) compared to studying the same material (Comparison) increase learning, retention, or transfer, and what is the magnitude of the effect (Outcome)? Several subsidiary questions are listed in [Supplementary Appendix 1](#). We also illuminated areas for further study in TEL. In this *mapping review*, we sought to identify gaps and methodologic shortcomings in the TEL literature ([Supplementary Appendix 1](#))

### Search strategy

We performed a Medical Subject Headings (MeSH) analysis of 30 representative articles using the Yale MeSH Analyzer (Grossetta-Nardini and Wang, 2017) and developed a concept table of search terms ([Supplementary Appendix 2](#)),

including controlled vocabulary and free text terms. Using these terms, we searched thirteen electronic bibliographic citation databases from 2000 to July 2016 ([Supplementary Appendix 3](#)), adjusting the search strategies for the syntax appropriate for each database/platform. We updated the searches in May of 2017 to capture articles published during the time of the review. We also manually searched the tables of contents of fourteen medical education journals for the same time period ([Supplementary Appendix 4](#)). Finally, we screened the references of and, looking forward, the articles citing the initially included articles.

### Inclusion and exclusion criteria

We sought primary research studies of the effectiveness of TEL on learning (including recall, retention, and transfer) in health professions education. Detailed criteria appear in [Supplementary Appendix 5](#). Inclusion criteria included controlled studies of TEL interventions in health professions education that reported an objective learning outcome and compared TEL to studying the same material or to a different TEL strategy.

These criteria were necessary to isolate the direct *retrieval effect* of TEL. A control group, in general, accounted for temporal trends and co-interventions. Furthermore, a control group that specifically studied the same material eliminated the *rehearsal effect*, in which looming assessments (prospectively) and the results of past assessments (retrospectively) *indirectly* effect learning via their influence on study behavior. We specified an objective learning outcome to allow comparisons between various types of TEL interventions. Finally, the basic study design in our included studies mirrored the design in retrieval practice studies in cognitive psychology. This allowed us to determine if the cognitive psychology laboratory and classroom findings translate into health professions education studies.

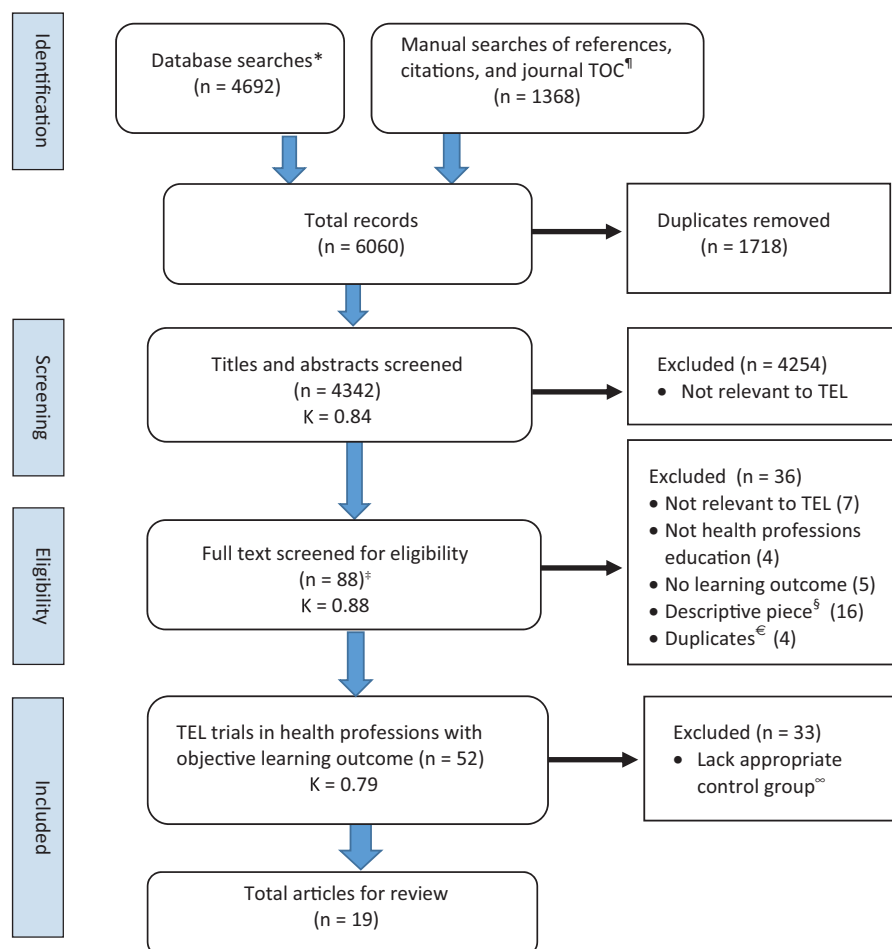
### Screening and selection of studies

The screening and selection process is shown in [Figure 1](#). Two readers (M. L. G. and J. J. M.) independently screened the titles and abstracts of the search output to exclude articles that were obviously irrelevant and not at all related to TEL. The same two readers independently performed a full text review of the remaining articles, applying the inclusion and exclusion criteria and indicating the reasons for exclusion. At each winnowing, we determine the inter-rater agreement with the Kappa statistic and resolved differences by consensus.

### Data extraction and coding

To refine the coding process, two raters independently and sequentially abstracted data from four articles using a provisional coding sheet. After each coding session, the raters met to revise the coding sheet by adding, eliminating, or editing items to resolve ambiguity and ensure consistency and completeness. We declared the coding sheet final after the fourth revision and translated it into an electronic survey, which we used for the remainder of the process.

Items on the coding sheet included study authorship, study citation, five study design variables, study quality



**Figure 1.** Flow diagram of study screening and eligibility. \*See [Supplementary Appendix 3](#) for databases. † References in captured articles, articles citing captured articles, and TOC of journals. See [Supplementary Appendix 4](#) for journals. § Reviews, editorials, letters. ‡ 76 from database and 12 from manual searching. ¶ 2 duplicated articles and 2 articles – meeting abstract pairs. ∞ Studying the same material or an alternate TEL strategy.

score, health profession, level of training, sample size, response rate, teaching session format and content, TEL interventions, control interventions, learning outcomes, and results (including magnitude, variance, and significance of effects). We used commonly accepted criteria for validity evidence of outcome instruments (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association; the American Psychological Association; and the National Council on Measurement in Education 1999; Beckman et al. 2005; Cook and Beckman 2006) (Table 1 and [Supplementary Appendix 6](#)). For the learning outcomes, we distinguished between *immediate learning* (determined immediately after the TEL intervention) and *retention* (determined sometime after the TEL intervention). *Transfer* outcomes required application of learned concepts to new inferential questions in the same or new knowledge domains.

To allow comparison of effects among heterogeneous studies, we determined the effect size (Hojat and Xu 2004; Leppink et al. 2016) if the study provided the requisite data. The effect size for comparing means is determined as the standardized mean difference (SMD). For comparing an intervention group to a control group:  $SMD = (\text{mean}_{\text{cases}} - \text{mean}_{\text{controls}}) / SD_{\text{controls}}$ . For comparing two interventions:  $SMD = (\text{mean}_{\text{cases}} - \text{mean}_{\text{controls}}) / SD_{\text{pooled}}$ . As a measure of the impact of an intervention, the SMD is operationally interpreted as:

SMD = 0.20 (SMALL, negligible practical importance)

SMD = 0.50 (MEDIUM, moderate practical importance)

SMD = 0.80 (LARGE, crucial practical importance)

If standard deviations were not provided, we calculated them from standard errors, *p* values, *t* values, and confidence intervals ([Supplementary Appendix 7](#)) (Higgins and Green 2011).

### Assessment of study methodological quality

We determined the Medical Education Research Study Quality Instrument (MERSQI) (Reed et al. 2007, 2008; Cook and Reed 2015) score each of the included studies ([Supplementary Appendix 8](#)). Prior research showed that manuscripts with higher scores on the MERSQI were more likely to receive positive editorial decisions in the peer review process. (Reed et al. 2008) MERSQI scores were also associated with 3-year citation rate, journal impact factors, and study funding (Reed et al. 2007).

### Reproducibility of data abstraction processes

Two raters (M. L. G. and J. J. M.) independently abstracted data from the included studies. We determined the inter-rater reliability for key variables, including the validity evidence for outcome instruments, the MERSQI quality score, and the calculated standardized mean difference for the main comparisons. We used the kappa statistic

Table 1. Study design and quality.

Study	Study design <sup>a,b</sup>	Adjustment for confounding <sup>c</sup>	Time on task (TOT)	Validity evidence for outcome instrument <sup>d</sup>	Co-intervention <sup>e</sup>	Quality Score <sup>f</sup>
Ali and Ruit (2014)	RCT Crossover Within subjects	NA	E-learning platform allowed 1 min per question	None	Feedback (interactive discussion of questions)	9
Baghdady et al. (2014)	RCT	Multivariable analysis: educational background (dentistry versus dental hygiene program)	E-learning program ensured equal TOT for the TEL and control groups	Content Internal consistency reliability Relationship to other variables (responsiveness) None	Feedback (answers)	14.5
Chen and Chuang (2012)	RCT	Multivariable analysis: birth order, body mass index, and parental age, education, employment status, and income	None	None	None	11
Cook et al. (2014)	RCT Crossover Within subjects	Multivariable analysis: residency program, postgraduate year, gender	Subjects reported the time to complete module	Content Internal consistency reliability	None	15.5
DeSignore et al. (2016)	RCT	Compared frequencies: Age, field of training, ICU experience, ventilator experience	Three study groups watched the same video simultaneously	Content Inter-rater reliability Relationship to other variables	None	15
Dobson and Linderholm (2015)	RCT Crossover Within subjects	Multivariable analysis: year of training, gender	Students were limited to 20 min per passage and encouraged to spend 5–7 min on each studying strategy	Internal consistency reliability	None	11
Dobson et al. (2015)	RCT Crossover Within subjects	NA	Students were timed to ensure equal time studying and testing. Eight minutes for each session.	None	None	11
Kromann et al. (2010)	RCT	None	TEL tested and controls studied for 30 min each	Inter-rater reliability Content Relationship to other variables (discrimination and responsiveness)	Feedback (general to the group, not reflecting individual performance)	14
Kromann et al. (2009)	RCT	None	TEL tested and controls studied for 30 min each	Inter-rater reliability Content Relationship to other variables (discrimination and responsiveness)	Feedback (general to the group, not reflecting individual performance)	14
Larsen et al. (2015)	RCT Crossover Within subjects	NA	On average, participants read through the review items 1.96 times and read through the answers to the quiz questions 1.54 times.	Inter-rater reliability	Feedback (answers and explanations)	11

(continued)

Table 1. Continued

Study	Study design <sup>a,b</sup>	Adjustment for confounding <sup>c</sup>	Time on task (TOT)	Validity evidence for outcome instrument <sup>d</sup>	Co-intervention <sup>e</sup>	Quality Score <sup>f</sup>
Larsen et al. (2013a)	RCT Crossover Within subjects	NA	None	Inter-rater reliability Content	Feedback (answers)	13
Larsen et al. (2009)	RCT Crossover Within subjects	NA. Documented clinical exposures to the content topics	None	Inter-rater reliability Content	Feedback (answers)	12
Larsen et al. (2013b)	RCT Crossover Within subjects	NA	No time limits placed on learning activities. But students completed testing and study session in about an hour.	Inter-rater reliability Content	Self-explanation	11
McConnell et al. (2015a)	RCT	None	None	None	None	11
McConnell et al. (2015b)	RCT Crossover Within subjects	NA	90 s for quiz items	Internal structure Content	Feedback (answers)	14
Messineo et al. (2015)	RCT	Multivariable analysis for test anxiety	Time limit for completion of TEL tests. E-learning program recorded time control students spent restudying the information	None	Feedback (answers)	11
Oglesby (2013)	RCT Crossover Within subjects	NA	Controlled class room setting. Student completed TEL tests for 20min then read study sheet on the other topic	None	Feedback (answers)	11
Raupach et al. (2016)	RCT Crossover Within subjects	NA	Seminars with test items – mean 21 min Seminars with study only items – mean 13 min	None	Feedback (answers and explanations)	12
Schmidmaier et al. (2011)	RCT	Compared frequencies for college grades and profession of parents	Both groups used same amount of time on task each session	None	None	11

<sup>a</sup>RCT = randomized controlled trial.

<sup>b</sup>Crossover within-subjects design = Student serve as cases (TEL) for some content areas and controls for other content areas.

<sup>c</sup>Within subject designs not vulnerable to confounding as students serve as their own controls.

<sup>d</sup>Commonly accepted categories of validity evidence (Supplementary Appendix 6).

<sup>e</sup>Feedback could consist of the correct answer alone or the correct answer with an explanation.

<sup>f</sup>Medical education research study quality instrument (Supplementary Appendix 8).



(dichotomous variables) and the intra-class correlation coefficient (continuous variables) to determine the level of agreement. We resolved differences through consensus.

## Results

### Reproducibility analyzes

The inter-rater reliability was excellent for the title and abstract screening ( $\kappa = 0.82$ ,  $p < 0.0001$ ), full text review for study inclusion ( $\kappa = 0.88$ ,  $p < 0.0001$ ), and full text review for meta-analysis ( $\kappa = 0.79$ ,  $p < 0.0001$ ). In the abstraction process, the two raters showed strong agreement in determining the SMD for the primary outcome measure ( $\kappa = 0.86$ ,  $p < 0.0001$ ) and the quality score (intraclass correlation coefficient 0.93,  $p < 0.0001$ ).

### Study retrieval, screening, and inclusion

Figure 1 shows the flow diagram of study screening and eligibility. Notably, there were four abstracts from scientific meetings among the 88 articles selected for full text review. We identified the corresponding full text for two of them through contacting authors and additional searching. Also, we translated one article written in Danish (Google translate<sup>®</sup>) and one article written in German (bilingual translator) into English. We winnowed the 88 articles down to 52 trials of TEL interventions in the health professions that reported an objective learning outcome and then to 19 controlled trials that also compared TEL to studying the same material or to a different TEL strategy. The characteristics of the 19 studies (including two PhD theses) appear in Table 1 (study design), Table 2 (demographics, control, and TEL interventions), and Table 3 (outcome measures and results).

### Study characteristics (study design)

The design and conduct of the studies, all randomized controlled trials (RCTs), were generally methodologically sound (Table 1). Notably, 11 of the RCTs studies employed a crossover, within-subjects design (Larsen et al. 2009; Schmidmaier et al. 2011; Larsen, Butler, Lawson, et al. 2013; Larsen, Butler, and Roediger 2013; Oglesby 2013; Ali and Ruit 2014; Cook et al. 2014; Dobson and Linderholm 2015; Dobson et al. 2015; Larsen et al. 2015; McConnell et al. 2015a; Raupach et al. 2016). That is, all subjects participated in all arms of the study (controls and one or more TEL intervention) but engaged in different content areas for each. In the analysis, investigators combined results for the different content areas, effectively doubling (for two arms) or tripling (for three arms, etc.) the sample size.

Quality scores ranged from 9 to 15.5 out of 18. Some studies lost points on the quality instrument for confining the research to a single institution, losing subjects to follow up, lacking multiple sources of validity evidence for the outcome instrument, or reporting knowledge but not behavior or health outcomes.

Three studies lost more than 40% of subjects to follow up. Most studies attempted to equalize the time on learning task between TEL and control groups. Among the standard RCTs, five accounted for confounding factors,

either by adjusting the results with multivariable analyzes or descriptively comparing frequencies between the TEL and control groups. Eleven studies reported some reliability or validity evidence for the learning outcome instrument, most commonly “inter-rater reliability” and “content validity” followed by “relationship to other variables.”

### Study characteristics (demographics, control, and TEL interventions)

Subjects included medical students (8 studies), nursing students (3), allied health students (2), residents (3), physicians in CME programs (2), and dental and dental hygiene students (1) (Table 2). TEL interventions included short answer questions (SAQs), multiple choice questions (MCQs), simulation (cardiopulmonary resuscitation), standardized patients, and key features questions (clinical reasoning). A *key feature* represents a difficult step in the identification and resolution of a problem in practice in which examinees are likely to make errors. Key features are embedded in case-based questions with short answer response formats (Page et al. 1995; Hrynchak et al. 2014). The short answer items were both cued (response to a question) and non-cued (free recall of relevant information).

Among the 12 studies in which students took multiple TEL tests, seven included identical items on repeated tests (Larsen, Butler, Lawson, et al. 2013; Larsen, Butler, Roediger 2013; Oglesby, 2013; Dobson et al. 2015; Larsen et al. 2015; McConnell et al. 2015b), one included different items on the same topics (Ali and Ruit 2014), and four included different items on different topics (Schmidmaier et al. 2011; Cook et al. 2014; Messineo et al. 2015; Raupach et al. 2016), linked to a new lecture or e-learning session. All of these studies spaced repeated TEL tests at 1–2 week intervals, with the exception of one study (Dobson et al. 2015) that employed consecutive TEL testing. Co-interventions included feedback for assessment items (correct answers with or without rationale) and *self-explanation*. For the latter, students generated explanations about why a particular piece of information is important and how it relates to their existing knowledge (Chamberland et al. 2015).

### Outcome measures and results

The 19 studies reported 49 learning outcomes, including examinations (MCQs, SAQs, essay, key features questions), radiograph interpretation, simulation (cardiac arrest scenario), and standardized patient assessment (Table 3). Forty-one outcomes included sufficient data to determine a SMD. Of the 29 outcomes comparing TEL to studying the same material, 26 showed SMDs favoring TEL (0.09–2.5), with some 95% CIs crossing zero due to low sample sizes. The SMD of 2.5 represents an outlier, as the next highest was 1.1. Twenty-four outcomes included either a direct statistical comparison between TEL and studying or determined the “effect” of TEL in a multivariable analysis. All but six of these comparisons showed statistical significance at  $p < 0.05$ .

Among the six *immediate learning* outcomes, five favored TEL over studying (SMD 0.09–0.44). Among the 23 *retention* outcomes (1 week–6 months), 21 favored TEL over studying (SMD 0.10–2.5). Three studies demonstrated the

**Table 2.** Study demographics, teaching sessions, and interventions.

Study	Learners	Response rate	Teaching or learning session(s)/topics	Controls	TEL intervention
Ali and Ruit (2014)	Medical students (n = 69)	Not reported	Ongoing anatomy course 16 radiologic anatomy and 12 clinical anatomy topics from lecture and laboratory. Topics randomly assigned to three groups	One test 8 MCQs Radiologic anatomy clinical anatomy topics	Three tests weekly 8 MCQs (same topics but different questions for each test) Three tests weekly 8 SAQs (same topics but different questions for each test)
Baghdady et al. (2014)	Undergraduate dental and dental hygiene students (n = 123) (two institutions)	(n = 112) 91%	E-learning software with learning material, radiographic images, audio, and testing. Radiographic features and pathophysiology underlying four bony abnormalities, representing four disease categories	Study sheet with information identical to TEL tests	One test at end of learning sessions 23 MCQs Pathophysiology underlying the four disease categories
Chen and Chuang (2012)	Nursing students (n = 151)	(n = 146) 97%	Ongoing community health nursing course	Study sheet with information identical to TEL tests	Computer-based testing (items or format not specified)
Cook et al. (2014)	Internal medicine and family medicine residents (n = 197)	(n = 180) 91%	Four web-based learning modules with embedded case-based self-assessment questions	Module with no questions	Modules with 1, 5, 10, or 15 questions
DeSignore et al. (2016)	Pediatrics residents (n = 49) (three institutions)	(n = 35) 71%	TEL items embedded in 23-min teaching video. High frequency oscillatory ventilation.	Uninterrupted video	Video interrupted with questions at 3–5 min intervals 10 SAQs Video interrupted logic puzzles at 3–5 min intervals
Dobson and Linderholm (2015)	Allied health professions students (n = 147)	(n = 120) 82%	TEL and control interventions integrated into 20-min self-study Cardiac electrophysiology, ventilation, endocrinology (knowledge)	For one of three topics, read study sheet three times in one sitting (R-R-R)	For one of three topics, read study sheet, read again, read a third time while taking notes in one sitting (R-R-N) For one of three topics, read study sheet, free recall (TEL), read again in one sitting (R-T-R). For free recall, students wrote down as many of the definitions and concepts as they could
Dobson et al. (2015)	Allied health professions students (n = 88)	(n = 83) 94%	TEL and control interventions integrated into 1-h self-study. Origins, insertions, actions, and innervations of six sets of skeletal muscle sets (2 familiar, 2 mixed, 2 unfamiliar)	For one set of muscles within each group of “familiarity,” 2 min reading the information four times consecutively in one setting (R-R-R-R) <sup>a</sup>	For one set of muscles within each group of familiarity, 2 min reading, 2 min free recall, 2 min reading, 2 min free recall in one setting (R-T-R-T) <sup>a</sup> . For free recall, students wrote down as many of the definitions and concepts as they could
Kromann et al. (2010)	Medical students (n = 180)	(n = 89) 50%	31/2-h CPR training, including simulated cardiac arrest scenarios	Study sheet with information identical to TEL tests Immediately after session, students ran through 3–4 scenarios that were identical to those used for the test in the intervention group, for 30 min	Immediately after session, students tested individually in one of six 5-min cardiac arrest scenarios. Spent another 5 min assisting another participant’s test. Scenarios differed in case story, but the scenarios and checklist were essentially the same. Tester rated the students’ performance on each item of 25-item checklist
Kromann et al. (2009)	Medical students (n = 140)	(n = 81) 59%	31/2-h CPR training, including simulated cardiac arrest scenarios (knowledge and skills)	Study sheet with information identical to TEL tests Immediately after session, students ran through 3–4 scenarios that were identical to those used for the test in the intervention group, for 30 min	Immediately after session, students tested individually in one of six 5-min cardiac arrest scenarios. Scenarios differed in case story, but the scenarios and checklist were essentially the same. Tester rated the students’ performance on each item of 25-item checklist
Larsen et al. (2015)	Participants in continuing education courses (n = 96)	(n = 35) 36%	Lectures in four CME Courses Neurology: epilepsy, headache, child neurology, multiple sclerosis (knowledge)	Study sheet with information identical to TEL tests Four sessions at 1 week intervals	Examination SAQs (number of items not specified) Four assessments at 1 week intervals

(continued)



Table 2. Continued

Study	Learners	Response rate	Teaching or learning session(s)/topics	Controls	TEL intervention
Larsen et al. (2013a)	First year medical students (n = 41)	(n = 41) 100%	2-h interactive session Migraine, seizures, and myasthenia gravis	Study sheet with information identical to TEL tests Four times weekly	Standardized patient assessments OSCE 27–29 items Students scored their own assessment by checklist Examination 27–29 SAQs Four exams at 1-week intervals Examination 40 SAQs Three identical examinations at 2-week intervals
Larsen et al. (2009)	Pediatrics and emergency medicine residents (n = 44)	(n = 40) 90%	1-h interactive session Status epilepticus, myasthenia gravis Knowledge	Study sheet with information identical to TEL tests	Four assessments at 1-week intervals Examination 27–29 SAQs Four exams at 1-week intervals Examination 40 SAQs Three identical examinations at 2-week intervals
Larsen et al. (2013b)	Medical students (n = 49)	(n = 47) 96%	2-h Interactive teaching session Seizure, optic neuritis, migraine, myasthenia gravis (knowledge)	Study sheet with information identical to TEL tests: Study (S)	Four sessions at 1-week intervals Examination: 26–30 SAQs (T) Self-explanation (E) <sup>b</sup> Exam plus self-explanation (TE) Study plus self-explanation (SE) Examination 10 SAQs
McConnell et al. (2015a)	Internists in continuing medical education course (n = 83)	(n = 56) 68%	One 20-min lecture Constipation management (knowledge)	Study sheet with information identical to TEL tests Explanation of the correct answer Study sheet with information identical to TEL tests	Examination 10 SAQs
McConnell et al. (2015b)	Medical students at four medical schools (n = 424) Four institutions	(n = 224) 52%	Context of preparing for Medical Counsel of Canada qualifying exam Population health, ethics, legal, and organization aspects of medicine (knowledge)	Study sheet with information identical to TEL tests Items built from MCQs with question and correct answer used to make a statement	Four examinations weekly 28 items (two for each learning objective) Context free (CF) MCQs <sup>c</sup> Context rich (CR) MCQs SAQs
Messineo et al. (2015)	Nursing students (n = 201)	(n = 161) 80%	Eight lectures in psychology course, timing not specified (knowledge)	Study sheet with information identical to TEL tests	Eight examinations weekly 10 MCQs
Oglesby (2013)	Nursing students (n = 36)	(n = 31) 81%	Two 50-min interactive sessions on the same day	Study sheet with information identical to TEL tests	Two examinations 2 weeks apart 20 MCQ
Raupach et al. (2016)	Medical students (n = 124)	(n = 87) 70%	Oxygenation and sepsis (knowledge) Ten 45-min case-based computer “e-seminars” Cardiology, pulmonary, nephrology, rheumatology, oncology (knowledge)	Study sheet with information identical to TEL tests Five e-seminars with four cases each Feedback information from TEL items included in different format (alternate weekly with test condition)	Five e-seminars with four cases each 20 key feature <sup>d</sup> questions embedded in each case 100 total items (alternate weekly with study condition)
Schmidmaier et al. (2011)	Medical students Years 2–5 (n = 80)	(n = 76) 95%	All students participated in four learning cycles using flashcards, viewing total of 30 electronic flash cards with total of 98 items in clinical nephrology. <sup>e</sup>	Correctly recalled flashcards were “studied” repetitively and the corresponding test cards were dropped from subsequent cycles	Correctly recalled flashcards were “tested” repetitively and the corresponding study cards were dropped from subsequent cycles

RCT: randomized control trial; MCQ: multiple choice question; SAQ: short answer question.

<sup>a</sup>In the outcome, we collapsed the results reported separately for the familiar, mixed, and unfamiliar muscle groups.

<sup>b</sup>Students generate explanations about why a particular piece of information is important and how it relates to their existing knowledge (Chamberland et al. 2015).

<sup>c</sup>Context rich MCQs require application of clinical knowledge; Context free MCQs require recognition of fact.

<sup>d</sup>A key feature focuses on a step in which examinees are likely to make errors in the resolution of a problem. It is a difficult aspect of the identification and management of the problem in practice. Key features are embedded in a case-based question with short answer response format (Page et al. 1995; Hrynchak et al. 2014).

<sup>e</sup>All students first engaged in a study session reading “study” flash cards, which included a learning objective and a list of “targets.” Then, after a distractor task, students engaged in “testing” trials in which they viewed learning objectives but had to type in the targets in blank spaces. Subsequently, the way in which correctly recalled flashcards were repeated differed between the control and TEL groups.

**Table 3.** Outcome assessments and results.

Study	Outcome assessment <sup>a</sup>	Results <sup>b,c</sup>	SMD (95% CI)
Ali and Ruit (2014)	<i>Retention</i> 4-weeks from last TEL test Examination: MCQs different from TEL MCQ arm (number not specified) <sup>j</sup>	Radiologic anatomy topics SAQ: 80% ± 32 MCQ: 65% ± 41	$d = 0.41$ (0.080, 0.75)
		Clinical anatomy topics SAQ: 97% + 18 MCQ: 62% ± 39	$d = 1.1$ (0.77, 1.5)
		Radiologic anatomy topics MCQ 3 TEL tests = 65% ± 41 MCQ 1 TEL test = 78% ± 39	$d = -0.33$ (-0.66, -0.01)
		Clinical anatomy topics MCQ 3 TEL tests = 62% ± 39 MCQ 1 TEL test = 78% ± 40	$d = -0.41$ (-0.41, -0.75)
	<i>Retention</i> 2-7 months from last TEL test Examination: MCQs different from TEL MCQ arm (number not specified) <sup>j</sup>	Radiologic anatomy topics SAQ: 52% MCQ: 61%	$d = -0.27$ (-0.61, 0.10)
		Non-radiologic anatomy topics SAQ: 67% MCQ: 73%	$d = -0.19$ (-0.53, 0.14)
Baghdady et al. (2014)	<i>Immediate learning</i> Examination: Matching – Choose the correct radiographic features among list of 12 for four intrabony abnormalities Different from TEL test	TEL = 73% ± 12 Controls = 75% ± 10 $p = 0.3^d$	$d = -0.19$ (-0.55, 0.19)
	<i>Retention</i> 1 week from last TEL test Same examination as above	TEL = 60% ± 10 Controls = 61% ± 10 $p = 0.3^d$	$d = -0.1$ (-0.47, 0.27)
	<i>Immediate (transfer)</i> Diagnostic accuracy assessment Matching: Choose correct diagnosis among list of four abnormalities for 22 radiographs Different from TEL test	TEL = 74% ± 15 Controls = 67% ± 16 $p = 0.04^d$	$d = 0.44$ (0.09, 0.81)
	<i>Retention (transfer)</i> 1 week from last TEL test Same format and content as above but different items	TEL = 72% +.15 Controls = 67% ± 15 $p = 0.04^d$	$d = 0.33$ (-0.02, 0.69)
Chen and Chuang (2012)	<i>Retention</i> 11 weeks from last TEL test Examination (format and content not specified)	TEL = 69% ± 13 Control = 63% ± 9.3 $p = 0.001$	$d = 0.64$ (0.26, 1.0)
	<i>Retention</i> 18 weeks from last TEL test Examination (format and content not specified)	TEL = 71% ± 9.9 Control = 68% ± 11 $p = 0.42$	$d = 0.23$ (-0.16, 0.62)
Cook et al. (2014)	<i>Immediate learning</i> Examination with case-based MCQ written to test knowledge application	TEL-1 question = 73% TEL-5 questions = 72% TEL-10 questions = 76% TEL-15 questions = 74% Control = 73% $p = 0.04^d$	$d = 0.09$ (TEL-10 versus control)
DelSignore et al. (2016)	<i>Immediate learning</i> Examination: 10 SAQs Different from TEL test	$p < 0.05$ TEL-10 versus control TEL = 67% Controls = 63% ± 13.6; Logic puzzle = 63%	$d = 0.29$ (-0.3, 0.95) (TEL versus control)
	<i>Retention</i> 6 months from last TEL test Same examination as above	TEL = 39% Controls = 39% ± 6.9 Logic puzzle = 30%	$d = 0$ (-0.80, 0.80) (TEL versus control)
	<i>Immediate learning</i> Examination: 30 MCQs (10 for each of the three study topics) Different from TEL test	TEL = 64% ± 19 Controls = 59% ± 17 $p = 0.03^d$	$d = 0.25$ (0, 0.50)
Dobson and Linderholm (2015)	<i>Retention</i> 1 week from last TEL test Same examination as above	TEL = 53% ± 18 Controls = 48% ± 20 $p = 0.03^d$	$d = 0.29$ (0.048, 0.56)
	<i>Immediate learning</i> Examination SAQs (recall as much information as they could for six sets of muscles) <sup>e</sup>	TEL = 34% ± 26 Controls = 28% ± 25 $p < 0.001^d$	$d = 0.24$ (-0.07, 0.52)
	Free recall for TEL and outcome		
Dobson et al. (2015)	<i>Retention</i> 1 week from last TEL test Same examination as above	TEL = 14% ± 11 Controls = 11% ± 11 $p < 0.001^d$	$d = 0.28$ (-0.02, 0.58)
	<i>Retention</i> 3 weeks from last TEL test Same examination as above	TEL = 11% ± 11 Controls = 9.7% ± 10 $p < 0.001^d$	$d = 0.12$ (-0.17, 0.42)
	<i>Retention</i> 6 months after CPR course and testing	TEL = 75% ± 11 Controls = 70% ± 17 $p = 0.06$	$d = 0.4$ (-0.087, 0.75)
	Simulated cardiac arrest scenario. 25-point checklist. Item scored 0-5, >2 indicated acceptable performance. Testing scenario differed from TEL but the checklists were same		
Kromann et al. (2010)	<i>Retention</i> 6 months after CPR course and testing	TEL = 82% ± Controls = 73% ± 10.2 $p < 0.001$	$d = 0.93$ (0.47, 1.4)
Kromann et al. (2009)	<i>Retention</i> 2 weeks after CPR course and testing		
	Simulated cardiac arrest scenario. 25-point checklist. Item scored 0-5, >2 indicated acceptable performance. Testing scenario differed from TEL but the checklists were same		

(continued)

Table 3. Continued

Study	Outcome assessment <sup>a</sup>	Results <sup>b,c</sup>	SMD (95% CI)
Larsen et al. (2015)	<i>Retention</i> 4.5 months from last TEL test Examination: SAQs (number of items not specified) Different from TEL test	TEL = 55% ± 19 Controls = 46% ± 15 <i>p</i> = 0.01	<i>d</i> = 0.60 (0.12, 1.1)
Larsen et al. (2013a)	<i>Retention (transfer)</i> 5 months from last TEL test Standardized patient assessment (Checklist identical to TEL but patient demographic data and history was different. Scored by faculty coders)	TEL (SP) = 59% ± 14 Controls = 43% ± 14 <i>p</i> < 0.0001 TEL (SAQ) = 49% ± 14 Controls = 43% ± 14 <i>p</i> = 0.04 TEL (SP) = 59% ± 14 TEL (SAQ) = 49% ± 14 <i>p</i> = 0.002	<i>d</i> = 1.1 (0.68, 1.6) <i>d</i> = 0.43 (0.0, 0.87) <i>d</i> = 0.71 (0.27, 1.2)
	<i>Retention</i> 5 months from last TEL test Examination: 27–29 SAQs Same as TEL items	TEL (SP) = 61% ± 14 Controls = 48% ± 14 <i>p</i> = 0.0002 TEL (SAQ) = 61% ± 14 Controls = 48% ± 14 <i>p</i> = 0.0001 TEL (SP) = 61% ± 14 TEL (SAQ) = 61% ± 14 NS	<i>d</i> = 0.93 (0.47, 1.4) <i>d</i> = 0.93 (0.47, 1.4) <i>d</i> = 0 (−0.43, 0.43)
Larsen et al. (2009)	<i>Retention</i> 5 months from last TEL test Examination: 80 SAQs items Same as TEL items	TEL = 39% ± 11 Controls = 26% ± 14 <i>p</i> < 0.001	<i>d</i> = 0.91 (0.46, 1.4)
Larsen et al. (2013b)	<i>Retention (transfer)</i> 5 months from last TEL test Essay examination. Outline approach to managing patient's condition, given a short clinical scenario. Different from TEL test	TE = 40%, T = 36%, SE = 29%, S = 20% TE > SE ( <i>p</i> = 0.001) T > SE ( <i>p</i> = 0.02) TE > T ( <i>p</i> = 0.08) SE > S ( <i>p</i> = 0.001) T > S ( <i>p</i> not reported)	<i>d</i> = 0.70 <i>d</i> = 0.48 <i>d</i> = 0.28 <i>d</i> = 0.68 <i>d</i> = 1.01 (0.63, 1.5)
McConnell et al. (2015a)	<i>Retention</i> 4 weeks from last TEL test Examination: 10 new SAQs Different from TEL test	TEL = 43% ± 20 Control = 41% ± 20 <i>p</i> = 0.71	<i>d</i> = 0.10 (−0.42, 0.62)
McConnell et al. (2015b)	<i>Retention</i> After TEL but time not specified Mock licensure examination: 116 Context rich MCQs Some verbatim questions from TEL (CR MCQ) and some with same objectives but different context and content <sup>f,g</sup>	SAQ > CF MCQ ( <i>p</i> < 0.001) SAQ > study ( <i>p</i> < 0.001) CR MCQ > CF MCQ ( <i>p</i> < 0.001) CR MCQ > study ( <i>p</i> < 0.001) SAQ versus CR MCQ ( <i>p</i> > 0.05) CF MCQ versus study ( <i>p</i> > 0.05)	Insufficient data
Messineo et al. (2015)	<i>Retention</i> 2 weeks from last TEL test 80 MCQs Different from TEL test	TEL = 23.1 ± 5.4 <sup>h</sup> Control = 20.5 ± 7.4 <i>p</i> < 0.05	<i>d</i> = 0.39 (0.031, 0.68)
Oglesby (2013)	<i>Retention</i> 7 weeks from last TEL test Examination: 40 MCQ Same as TEL test items	TEL = 70% ± 11.5 Control items = 49% ± 8.5 <i>p</i> = 0.005	<i>d</i> = 2.5 (1.8, 3.1)
Raupach et al. (2016)	<i>Retention</i> 1 week from last TEL test Examination: 30 key features questions <i>Retention</i> 9 months Same examination	Case items 59.3% ± 27.7 Control items 46.7% ± 24.8 <i>p</i> < 0.001 Case items 56% ± 25.8 Control items 48.8% ± 24.7 <i>p</i> < 0.001	<i>d</i> = 0.51 (0.21, 0.81) <i>d</i> = 0.29 (0, 0.59)
Schmidmaier et al. (2011)	<i>Immediate learning</i> Examination: 13 tests cards from learning sessions <i>Retention</i> 1 week from last TEL test Examination: 30 test cards <i>Retention</i> 6 months from last TEL test Examination: 30 test cards	TEL = 86.97% <sup>i</sup> Controls = 88.56% NS TEL = 67.2% <sup>i</sup> Controls = 57.3% ± 13 <i>p</i> < 0.001 TEL = 34.28% <sup>i</sup> Controls = 38.29% NS	Insufficient data <i>d</i> = 0.76 (0.43, 1.1) Insufficient data

<sup>a</sup>The timing of outcome assessments represents the time elapsed after last TEL intervention. *Immediate learning* is determined immediately after the TEL intervention. *Retention* is determined sometime after the TEL intervention. *Transfer* outcomes required application of learned concepts to new inferential questions in the same or new knowledge domains.

<sup>b</sup>*p* value relates to pairwise comparisons, unless otherwise indicated. Some studies failed to report pairwise comparisons at all or reported them without *p* values.

<sup>c</sup>Scores reported as proportion correct are presented percent correct (%). Otherwise the scores represent raw scores, as reported in the studies.

<sup>d</sup>*p* value relates to an "effect" of learning strategy (TEL versus studying), as determined by an ANOVA, in a model including other variables such as assessment interval, content area, familiarity with content. It does not directly relate to pairwise comparisons.

<sup>e</sup>Means of scores reported separately for familiar muscles, mixed familiar and unfamiliar muscles, and unfamiliar muscles.

<sup>f</sup>Scores not reported.

<sup>g</sup>Higher mean scores found on previously seen relative to novel questions.

<sup>h</sup>Test anxiety inversely associated with scores ( $\beta = -0.47$ ).

<sup>i</sup>Items used as unit of analysis.

<sup>j</sup>Mean scores reported separately for individual topics. We calculated mean scores for each study arm.

impact of TEL on *transfer*: (1) SAQ TEL assessed with standardized patient outcome (Larsen et al. 2013), (2) MCQ TEL assessed with radiograph interpretation outcome (Baghdady et al. 2014), and (3) SAQ TEL assessed with essay examination (students described the approach to managing a patient's condition) (Larsen et al. 2013). All seven *transfer* outcomes in these studies favored TEL over studying (SMD 0.33–1.1). All of the studies that repeatedly measured outcomes over time found a decay in the TEL effect. At 1 week, five studies documented 2.7–29% decreases in scores. Two studies showed reductions in scores by 42 and 60% at 6 months.

Fifteen outcomes compared different TEL strategies. The only study that varied the number of tests failed to show an advantage of three TEL tests over one (Ali and Ruit 2014). One study varied the number of questions per TEL test. The learning outcome scores increased, albeit within a very narrow margin, as the number of questions progressed from 1 to 5 to 10 questions, with no further advantage beyond that (Cook et al. 2014). SAQ tests resulted in better learning than MCQ tests in four outcomes in two studies (SMDs 0.86 and 1.1) (Ali and Ruit 2014; McConnell et al. 2015b). “Context rich” MCQs (requiring application of knowledge) performed better than “context free” MCQs (requiring recognition of facts) (McConnell et al. 2015b). Students with standardized patients TEL retained more than students with SAQs TEL on a standardized patient outcome test (SMD 0.71) but not on a written examination outcome test (Larsen 2013a) Self-explanation as a co-intervention enhanced studying (SMD 0.68) to a greater degree than it enhanced testing (SMD 0.28) in medical students (Larsen et al. 2013b).

Six studies examined the effect of content on TEL. Three descriptively showed different SMDs for different medical topics (Larsen et al. 2009; Ali and Ruit 2014; Raupach et al. 2016) Three isolated a content effect in multivariable analyses ( $\eta^2$  0.06–0.12), indicating a modest effect (Larsen, Butler, Lawson, et al. 2013; Larsen, Butler, and Roediger 2013; McConnell et al. 2015b).

## Discussion

Studies in cognitive psychology consistently demonstrate that recalling previously learned information (*retrieval practice*) enhances the ability to recall the information in the future (*retrieval effect or test-enhanced learning*). Students who engage in effortful, deliberate attempts to recall information show better learning, retention, and transfer than students who spend the same time studying the same material.

In our systematic review, we initially identified 52 reports of TEL interventions in health professions education. Of these, only 19 compared TEL to studying the same material or to an alternate TEL strategy. TEL, in these studies, demonstrated consistent and robust effects across different health professions, learner levels, TEL formats, and learning outcomes.

The design and conduct of the 19 studies, all randomized controlled trials, was generally sound, as indicated by relatively high quality scores. Methodologic shortcomings, in some studies, included small sample sizes, low response rates, limited validity evidence for the outcome instrument,

and insufficient descriptions of TEL interventions and outcome measures.

Notably, 11 studies employed a within-subjects design. This design benefits from larger analytical sample sizes (and thus greater power), as each subject contributes data to all arms of the study. These studies also remain invulnerable to confounding as each subject serves as her own control. On the other hand, participation in one arm of the study may *carry over* to performance in another arm. For instance, students who first participate in the TEL arm may be inclined to test themselves when they subsequently participate in the “study only” control arm. This would be expected to bias the results toward the null hypothesis. If subjects are tested immediately after participating in each arm of a trial, their scores may improve over time due to *practice effects* rather than a real effect of TEL. This was not the case in the studies herein as subjects took the outcome assessments after participating in the control and TEL arms.

TEL has been extensively studied by cognitive and educational psychologists (Roediger and Butler 2011; Brame and Biel 2015; Eisenkraemer et al. 2013). The effect size for retrieval practice in laboratory settings and primary, secondary, and post-secondary classrooms has been estimated to be 0.5 (Rowland 2014; Adesope et al. 2017) Our review revealed that health professions educators internalized and extended these findings in their TEL research. Firstly, both literatures find consistent and robust effects of TEL across multiple learners, settings, and TEL formats. While the cognitive psychology studies confined TEL interventions and learning outcomes to “examinations” of some type, health professions education studies included a wider array of clinical assessments, such as radiograph interpretation (Baghdady et al. 2014), cardiopulmonary resuscitation simulation (Kromann et al. 2009, 2010), standardized patient encounters (Larsen et al. 2013a), and clinical reasoning (Raupach et al. 2016).

Cognitive psychology experiments demonstrate that repeated TEL tests are more effective than just one (Wheeler and Roediger 1992; Roediger and Karpicke 2006; Karpicke and Roediger 2008). In addition, spacing the tests over time is superior to consecutive testing (Landauer and Bjork 1978; Cepeda et al. 2006; Karpicke and Roediger 2007). Students in seven of our health professions studies took one only one TEL test, while the remainder took 2–8 TEL tests at 1–2 week intervals. The only health professions study that compared repeated testing (3 weekly) with a single test did not reveal an advantage to repeated testing (Ali and Ruit 2014). This unexpected finding might be explained by the extreme content specificity in the data. Across the study groups, performance varied widely for different anatomic topics. It is possible that the 1-week arm topics were easier than others. Also, the repeated testing arm included different items on same topic for the three TEL tests. This may have attenuated learning compared to the more common practice of including identical items over repeated testing.

In cognitive psychology studies, items requiring production of information perform better than items requiring recognition of information (McDaniel, Anderson, et al. 2007; McDaniel, Roediger, et al. 2007; Pyc and Rawson 2009). This effect has been called “desirable difficulty.” Our review confirmed this finding in health professions education, demonstrating an advantage of short answer

questions over multiple choice questions, and an advantage of context rich multiple choice questions (which require application of knowledge) over context free multiple choice questions.

Feedback after a retrieval attempt increases the mnemonic effect of testing (Bangert-Drowns et al. 1991). It should include the correct answer (not merely right or wrong) to prevent the student from retrieving and “learning” an incorrect response (Butler et al. 2007; Butler and Roediger 2008). This is particularly important for recognition items because the students are “exposed” to incorrect information in the distractor options (Butler et al. 2006). Feedback also enhances the retention of correct but low confidence responses (Butler et al. 2008). Several studies suggest that delayed feedback may be more effective than immediate feedback (Butler et al. 2007). Students in the majority of the health professions studies received feedback, which included the correct response and, in many instances, the rationale for the correct response. None of the studies, however, investigated the effect of feedback on learning and retention.

Psychologists have proposed several theories to explain the retrieval effect (Roediger and Butler 2011; Karpicke and Grimaldi 2012; Yeh and Park 2015) Memory may have two dimensions: storage strength and retrieval strength, which, according to the *deficient processing* theory, are negatively correlated during initial learning. More difficult retrieval (lower retrieval strength) results in higher gains in storage strength. Recalling information is more “difficult” than reading or recognizing it. This may also explain the spacing effect as it is more “difficult” to repeatedly recall information with intervening time gaps than to recall it on consecutive trials. Memory performance, per the *transfer-appropriate processing* theory, is enhanced to the extent that the learning context matches the retrieval context. The act of testing as practice more closely approximates the conditions on the final test than simply rereading the material. Finally, retrieval of information from memory may elaborate the memory trace and create additional retrieval routes (cues), which make it more likely that the information will be successfully retrieved again in the future. At the neuronal level, these effects presumably accompany both molecular changes at the level of individual synapses and more widespread modifications of the neuronal network (Friedlander et al. 2011).

Interpretation of our finding should be tempered by a few potential limitations. As with any systematic review, we could have missed studies and suffered publication bias. In this case, our review may have overestimated the effect of TEL, as negative studies are less likely to be published. However, our search included many measures to ensure exhaustive capture, including librarian assistance with search terms and strategies, searching multiple databases, manual searching of medical education journals, reviewing references in and articles citing captured articles, translating articles in other languages, including unpublished PhD theses, and contacting authors of abstracts. In addition, the heterogeneity of the studies precluded a quantitative synthesis or formal meta-analysis. However, when possible, we determined the SMD as a common comparable measure of “impact.”

While our restrictive inclusion criteria were necessary to isolate the effect of retrieval practice and allow

comparisons to cognitive psychology studies, we acknowledge that these restrictions limited the scope of the review. In particular, we likely neglected potentially illuminating reports of uncontrolled TEL interventions and descriptions of innovative TEL strategies.

In summary, TEL in the health professions demonstrated consistent and robust effects across different professions, learner levels, TEL formats, and learning outcomes. These results extend the findings in cognitive psychology studies to several clinical applications. These findings have several implications for educators in the health professions. Educators should consider including TEL in their curricula to enhance recall, retention, and transfer of medical information. Ideally, TEL “tests” should be repeated, spaced over time, utilize items that require production of information, and include feedback with the correct responses and rationale.

It may be challenging for educators to integrate TEL into already overfilled curricula. They might appropriate some didactic lecture time for a quiz, with a few short answer questions, inked to each lecture or other learning session. The students might “borrow” some of their dedicated study time to take the quiz. Even though they might see the correct answers after completing the quiz, they would nonetheless retake it over specified time intervals. Finally, the teacher would revisit the quiz in class time to provide explanations for the correct answers. In this role, faculty should provide clear explanations for the range of “correct” responses to the short answer questions. The findings in our review suggest that such tradeoffs in curricular time would be worthwhile.

Furthermore, retrieval practice need not be restricted to formal “tests.” Whenever a student considers the material she is learning, sets it aside, and actively reconstructs it, she is engaging in retrieval practice. These might include e-learning modules with interspersed questions (DeSignore et al. 2016; Raupach et al. 2016), electronic flashcards with questions (Schmidmaier et al. 2011), and various question generating applications. Finally, educators can modify existing educational “homework” strategies to incorporate retrieval practice. Educators can change “open book” learning activities to “closed book” activities followed by feedback. In cognitive psychology studies, these “closed book” modifications applied to take home quizzes (Agarwal et al. 2008) and concept mapping (Blunt and Karpicke 2014) greatly enhanced learning.

Our findings also have implications for health professions students and trainees. College undergraduate students demonstrate poor metacognitive awareness. While they *actually* learn more with repeated retrieval practice, they *predict*, prior to engaging, that they will learn more with repeated reading (Roediger and Karpicke 2006). Consistent with this belief, undergraduates most commonly employ repeated reading as a study strategy and very rarely engage in retrieval practice (Karpicke et al. 2009). Consequently, health professions students may not readily embrace multiple TEL quizzes or retrieval practice as an effective study strategy. Educators may need to share the persuasive data or invite students to try retrieval practice and see if it improves their retention compared to repeated reading (Dobson and Linderholm 2015). Once persuaded of the virtues of TEL, students should find gratification in the knowledge that assessment is not merely an administrative



exercise imposed by external stakeholders. On the contrary, assessment promotes learning in ways that studying cannot.

Health professions education researchers can focus their efforts on the gaps and shortcomings revealed in our review. Studies of health professions students' metacognitive awareness of retrieval based learning should inform TEL curriculum development. Larger sample sizes with better response rates will narrow the wide confidence intervals in our review and avoid the necessity of within-subjects designs and their potential biases. Researchers should explore the effects of different TEL test formats, strategies, and co-interventions, including feedback. Outcome measures should be supported by multiple types of validity evidence. Finally, multi-institutional studies will reinforce the generalizability of TEL effects.

### Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

### Notes on contributors

**Michael L. Green**, MD, ScM, is professor of medicine and director of student assessment at the Teaching and Learning Center at Yale School of Medicine.

**Jeremy J. Moeller**, MD, MSc, is assistant professor, program director, and associate vice-chair for education in the Department of Neurology at Yale School of Medicine.

**Judy M. Spak**, MLS, is assistant director of research and education services at the Cushing-Whitney Medical Library at Yale School of Medicine.

### References

- Adesope OO, Trevisan DA, Sundararajan N. 2017. Rethinking the use of tests: a meta-analysis of practice testing. *Rev Educ Res.* 87:659–701.
- Agarwal PK, Bain PM, Chamberlain RW. 2012. The value of applied research: retrieval practice improves classroom learning and recommendations from a Teacher, a Principal, and a Scientist. *Educ Psychol Rev.* 24:437–448.
- Agarwal PK, Karpicke JD, Kang SHK, Roediger HL, McDermott KB. 2008. Examining the testing effect with open- and closed-book tests. *Appl Cognit Psychol.* 22:861–876.
- Ali SH, Ruit KG. 2014. Psychometrics and test-enhanced learning in a patient-centered learning curriculum. *Ann Arbor: The University of North Dakota.*
- Augustin M. 2014. How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. *Yale J Biol Med.* 87:207–212.
- Baghdady M, Carnahan H, Lam EW, Woods NN. 2014. Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Med Educ.* 48:181–188.
- Bangert-Drowns RL, Kulik C-LC, Kulik JA, Morgan M. 1991. The instructional effect of feedback in test-like events. *Rev Educ Res.* 61:213–238.
- Beckman TJ, Cook DA, Mandrekar JN. 2005. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med.* 20:1159–1164.
- Blunt JR, Karpicke JD. 2014. Learning with retrieval-based concept mapping. *J Educ Psychol.* 106:849–858.
- Brame CJ, Biel R. 2015. Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science courses. *CBE-Life Sci Educ.* 14:es4.
- Butler AC, Karpicke JD, Roediger HL. 2007. The effect of type and timing of feedback on learning from multiple-choice tests. *J Exp Psychol Appl.* 13:273–281.
- Butler AC, Karpicke JD, Roediger HL, III. 2008. Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *J Exp Psychol Learn Mem Cogn.* 34:918–928.
- Butler AC, Marsh EJ, Goode MK, Roediger HL, III. 2006. When additional multiple-choice lures aid versus hinder later memory. *Appl Cognit Psychol.* 20:941–956.
- Butler AC, Roediger HL, III. 2008. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Mem Cogn.* 36:604–616.
- Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. 2006. Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol Bull.* 132:354–380.
- Chamberland M, Mamede S, St-Onge C, Setrakian J, Bergeron L, Schmidt H. 2015. Self-explanation in learning clinical reasoning: the added value of examples and prompts. *Med Educ.* 49:193–202.
- Chen HY, Chuang CH. 2012. The learning effectiveness of nursing students using online testing as an assistant tool: a cluster randomized controlled trial. *Nurs Educ Today.* 32:208–213.
- Cook DA, Beckman TJ. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 119:166.e7–166.e16.
- Cook DA, Reed DA. 2015. Appraising the quality of medical education research methods: the medical education research study quality instrument and the Newcastle-Ottawa Scale-Education. *Acad Med.* 90:1067–1076.
- Cook DA, Thompson WG, Thomas KG. 2014. Test-enhanced web-based learning: optimizing the number of questions (a randomized cross-over trial). *Acad Med.* 89:169–175.
- DelSignore LA, Wolbrink TA, Zurkowski D, Burns JP. 2016. Test-enhanced e-learning strategies in postgraduate medical education: a randomized cohort study. *J Med Internet Res.* 18:e299.
- Dobson J, Linderholm T. 2015. Self-testing promotes superior retention of anatomy and physiology information. *Adv Health Sci Educ.* 20:149–161.
- Dobson JL, Linderholm T, Yarbrough MB. 2015. Self-testing produces superior recall of both familiar and unfamiliar muscle information. *Adv Physiol Educ.* 39:309–314.
- Eisenkraemer RE, Jaeger A, Stein LM. 2013. A systematic review of the testing effect in learning. *Paidéia (Ribeirão Preto).* 23:397–406.
- Fitch ML, Drucker AJ, Norton JA, Jr. 1951. Frequent testing as a motivating factor in large lecture classes. *J Educ Psychol.* 42:1–20.
- Friedlander MJ, Andrews L, Armstrong EG, Aschenbrenner C, Kass JS, Ogden P, Schwartzstein R, Viggiano TR. 2011. What can medical education learn from the neurobiology of learning? *Acad Med.* 86:415–420.
- Grossetta-Nardini HK, Wang L. 2017. The Yale MeSH Analyzer; [accessed 2017 April]. <http://mesh.med.yale.edu/>.
- Higgins JPT, Green S, editors. 2011. *Cochrane handbook for systematic reviews of interventions version 5.1* [Data extraction for continuous outcomes]. London: The Cochrane Collaboration. Available from <http://handbook.cochrane.org/>.
- Hojat M, Xu G. 2004. A visitor's guide to effect sizes – statistical significance versus practical (clinical) importance of research findings. *Adv Health Sci Educ.* 9:241–249.
- Hrynchak P, Glover Takahashi S, Nayer M. 2014. Key-feature questions for assessment of clinical reasoning: a literature review. *Med Educ.* 48:870–883.
- Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association; the American Psychological Association; and the National Council on Measurement in Education. 1999. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.
- Karpicke J, Grimaldi P. 2012. Retrieval-based learning: a perspective for enhancing meaningful learning. *Educ Psychol Rev.* 24:401–418.
- Karpicke JD, Butler AC, Roediger HL, III. 2009. Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory.* 17:471–479.
- Karpicke JD, Roediger HL. 2008. The critical importance of retrieval for learning. *Science.* 319:966–968.
- Karpicke JD, Roediger HL, III. 2007. Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *J Exp Psychol Learn Mem Cogn.* 33:704–719.



- Kreiter CD, Green J, Lenocho S, Saiki T. 2013. The overall impact of testing on medical student learning: quantitative estimation of consequential validity. *Adv Health Sci Educ.* 18:835–844.
- Kromann CB, Bohnstedt C, Jensen ML, Ringsted C. 2010. The testing effect on skills learning might last 6 months. *Adv Health Sci Educ.* 15:395–401.
- Kromann CB, Jensen ML, Ringsted C. 2009. The effect of testing on skills learning. *Med Educ.* 43:21–27.
- Landauer T, Bjork R. 1978. Optimum rehearsal patterns and name learning. In: Gruneburg M, Morris P, Sykes R, editors. *Practical aspects of memory.* London: Academic Press. p. 625–632.
- Larsen DP, Butler AC, Aung WY, Corboy JR, Friedman DI, Sperling MR. 2015. The effects of test-enhanced learning on long-term retention in AAN annual meeting courses. *Neurology.* 84:748–754.
- Larsen DP, Butler AC, Lawson AL, Roediger HL. 2013a. The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Adv Health Sci Educ.* 18:409–425.
- Larsen DP, Butler AC, Roediger HL, III. 2009. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Med Educ.* 43:1174–1181.
- Larsen DP, Butler AC, Roediger HL. 2013b. Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Med Educ.* 47:674–682.
- Larsen DP, Butler AC, Roediger HL, III. 2008. Test-enhanced learning in medical education. *Med Educ.* 42:959–966.
- Leppink J, O'Sullivan P, Winston K. 2016. Effect size - large, medium, and small. *Perspect Med Educ.* 5:347–349.
- McConnell MM, Azzam K, Xenodemetropoulos T, Panju A. 2015a. Effectiveness of Test-enhanced learning in continuing health sciences education: a randomized controlled trial. *J Contin Educ Health Prof.* 35:119–122.
- McConnell MM, St-Onge C, Young ME. 2015b. The benefits of testing for learning on later performance. *Adv Health Sci Educ.* 20:305–320.
- McDaniel MA, Anderson JL, Derbish MH, Morrisette N. 2007a. Testing the testing effect in the classroom. *Eur J Cogn Psychol.* 19:494–513.
- McDaniel MA, Roediger HL, McDermott KB. 2007b. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychon Bull Rev.* 14:200–206.
- Messineo L, Gentile M, Allegra M. 2015. Test-enhanced learning: analysis of an experience with undergraduate nursing students approaches to teaching and learning. *BMC Med Educ.* 15:182.
- Newble DI, Jaeger K. 1983. The effect of assessments and examinations on the learning of medical students. *Med Educ.* 17:165–171.
- Oglesby R. 2013. Examining nursing students' retention of taught content by repeat study and repeat testing: a replicated study (D.N.P.), Gardner-Webb University. <http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=109863261&site=ehost-live&scope=site>.
- Page G, Bordage G, Allen T. 1995. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med.* 70:194–201.
- Pyc MA, Rawson KA. 2009. Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *J Mem Lang.* 60:437–447.
- Raupach T, Andresen JC, Meyer K, Strobel L, Koziolok M, Jung W, Brown J, Anders S. 2016. Test-enhanced learning of clinical reasoning: a crossover randomised trial. *Med Educ.* 50:711–720.
- Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. 2008. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's Medical Education Special Issue. *J Gen Intern Med.* 23:903–907.
- Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. 2007. Association between funding and quality of published medical education research. *JAMA.* 298:1002–1009.
- Roediger HL, Agarwal PK, McDaniel MA, McDermott KB. 2011. Test-enhanced learning in the classroom: long-term improvements from quizzing. *J Exp Psychol Appl.* 17:382–395.
- Roediger HL, Butler AC. 2011. The critical role of retrieval practice in long-term retention. *Trends Cogn Sci.* 15:20–27.
- Roediger HL, III, Karpicke JD. 2006. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci.* 17:249–255.
- Rowland CA. 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol Bull.* 140:1432–1463.
- Schmidmaier R, Ebersbach R, Schiller M, Hege I, Holzer M, Fischer MR. 2011. Using electronic flashcards to promote learning in medical students: retesting versus restudying. *Med Educ.* 45:1101–1110.
- Schuwirth LWT, Van der Vleuten CPM. 2011. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach.* 33:478–485.
- Wheeler MA, Roediger HL. 1992. Disparate effects of repeated testing: reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychol Sci.* 3:240–245.
- Yeh DD, Park YS. 2015. Improving learning efficiency of factual knowledge in medical education. *J Surg Educ.* 72:882–889.