1. Cover Sheet

Title: The effectiveness of test-enhanced learning in health professions education—a systematic review and meta-analysis

Reviewers:

Michael Green, MD, ScM (lead reviewer)	Professor of Medicine Associate Director for Student Assessment Yale University School of Medicine Teaching and Learning Center 367 Cedar Street, ESH-A 2nd floor New Haven, CT 06510 203-785-3327 michael.green@yale.edu
Jeremy Moeller, MD, MS (master of science in clinical education)	Assistant Professor of Neurology Neurology residency program director Yale School of Medicine jeremy.moeller@yale.edu
Judy Spak. BA, MLS	Curriculum Services Librarian Cushing/Whitney Medical Library Yale School of Medicine 333 Cedar Street, SHM-L111 New Haven, CT 06510 203.737.2961 judy.spak@yale.edu
Statistician	To be named

2. Background to the Topic

Educators commonly think of assessment "of" learning, which occurs at the end of a course of study and determines what the students learned. Recently, educators have turned their attention to assessment "for" learning,[1] considering assessment as a pedagogical strategy in and if itself. Looming assessments *indirectly* enhance learning by driving students' study behaviors (*rehearsal effect*). 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 consistently demonstrate that recalling previously learned information (*retrieval practice*) enhances the ability to recall the information in the future (*retrieval effect*).[2-6] 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. This effect occurs even without feedback.

More recently, investigators observed demonstrated the retrieval effect in health professions education. Trainees allocated to testing (versus studying) demonstrate superior medical knowledge[7,8] (by multiple choice tests) and better skills (cardiopulmonary resuscitation,[9,10] radiograph interpretation[11]), with effects lasting up to six months. Different type of testing (e.g. multiple choice versus standardized patients) may have different effects on learning.[12] The standardized effect size of TEL has been estimated at 0.9,[13] indicating large practical importance.

As indicated in the review questions below, our systematic review will address the effectiveness of TEL in enhancing learning (including retention and transfer); the magnitude of this effect; and possible differential effects with different populations, test and testing timing and procedures, and co-interventions. Medical educators commonly understand that assessment indirectly enhances learning by driving study behavior. Our review should increase their awareness that assessment also directly enhances learning. Our findings will help educators design formative assessments and assessment systems that maximize learning, both indirectly and directly. Furthermore, our review may reveal gaps in TEL science that will direct further research.

3. Review topic/question(s), objectives and key words

Review questions

Our review will address the question: Does TEL enhance learning in health professions education?

Within this main question, we will consider several sub questions"

Systematic Review:

- What is the magnitude of effect of TEL?
- What are the effects on short term recall, retention, and transfer?
- What are the effects of testing conditions of TEL? (number, frequency, format, context of "tests")
- What are the effects of different types of "testing" (multiple choice examination, essay or short answer examinations, objective structured clinical examinations with standardized patients, simulation, etc.)?
- Are there differential impacts for learning knowledge, skills, attitudes, competencies?
- Are there differential impacts for the level of learner (undergraduate medical education, graduate medical education, continuing medical education?
- Are there differential impacts for different health professions? (allopathic medicine, nursing, allied health professions, dentistry)
- What are the effects of TEL "enhancements," such as retrieval clues, feedback, and self-explanation?

Mapping Review:

- What gaps in the TEL literature must medical education researchers target to answer the outstanding questions?
- What methodologic shortcomings in the literature should be improved to make higher confidence inferences about the effectiveness of TEL

Review objective

This review will help educators plan TEL interventions to meet their pedagogical needs and maximize the impact on learning. Information gaps in TEL will illuminate directions for medical education research.

Key words: **"**test-enhanced learning," "retrieval effect," "retrieval practice," "retention,"

4. Search sources and strategies

Search for previous systematic reviews of TEL

Searches of the Cochrane Database of Systematic Reviews, Campbell Collaboration Library of Systematic Reviews, PubMed (using systematic review filter), and PROSPERO captured no existing systematic reviews of TEL. An examination of the BEME website in the Published Reviews, Reviews in Progress and Just Registered Topics sections did not turn up any systematic reviews on this topic. We are aware of a non-systematic review[13] that performed a very limited search and restricted the outcome to a single measure (average impact of TEL estimated by standardized effect size).

Database searching

We plan to search the following databases from year 2000 to present. In our extensive preliminary searching we did not identify any relevant articles before year 2007:

MEDLINE (Ovid) (National Library of Medicine)	CINAHL (EBSCO) (Cumulative Index to Nursing and Allied Health Literature)
Embase (Ovid)	ERIC (Education resources information center)
AMED (Ovid) (allied and complementary medicine)	Education Research Complete (EBSCO)
PsycINFO (Ovid)	Scopus
Academic Search Premier	ProQuest Dissertations & Theses Global
	Proceedings Citation Index Science and Conference Proceedings Citation Index Social Science & Humanities (Web of Science)

We performed a MeSH (Medical Subject Headings) analysis of 30 representative articles and developed this preliminary concept table of search terms for use in searching MEDLINE. This will be adapted to most efficiently search the other databases. Of note Scopus and EMBASE contain conference proceedings

Concept table

	Field of Study		***Intervention***		***Outcome***
Controlled Vocabulary	Education, medical Education, Nursing Education, Pharmacy Education, Veterinary Students, Health Occupations Osteopathic Medicine/education	***AND***	Educational Measurement Self-Evaluation Programs Test Taking Skills Practice Reinforcement Feedback	***AND***	Learning Comprehension Reinforcement Knowledge Health Knowledge, Attitudes, Practice Professional Competence Clinical Competence ***OR*** Memory Mental Recall Recognition (Psychology) Repetition Priming Retention (Psychology) ***OR*** Transfer Time Factors
Free Text	(intern or interns*).tw. ((medic* or nurs* or dent* or pharmac*) adj residen*).tw. PGY*.tw. ((medic* or nurs* or physician assistant* or allied health or dent* or pharmac*) adj2 (student* or trainee*)).tw.		test-enhanced learning.tw. TEL.tw. testing effect*.tw. (repetitive adj test*).tw. (repeat* adj2 test*).tw. testing adj2 learning.tw. repeat* retrieval.tw. repeat* adj2 (quiz* or exam* or test* or study*).tw. (space* adj2 (test* or assess* or exam* or quiz*)).tw. retrieval effect.tw. retrieval practice.tw.		long term learning.tw. long term retention.tw. retain*.tw. clinical skill*.tw. enhance* adj2 knowledge.tw. clinical* adj2 competen*.tw.

Reference Searching:

We will use *Scopus* to identify articles cited in included articles and "future" articles that cite included articles

Manual searching:

We will hand search the tables of contents of the following medical education journals from the year 2000 until present:

Academic Medicine	Teaching and Learning in Medicine
Medical Education	BMC medical education
Medical Education On-Line	Medical Science Educator
Advances in Health Professions Education	The Clinical Teacher
Journal of Continuing Education in the Health Professions	Journal of Graduate Medical Education
Medical Teacher	Perspectives on Medical Education
Nurse Education Today	Nurse Educator
Medical Education Research Network*	Research on Medical Education Outcomes (ROMEO)*

*Parochial databases of medical education studies

5. Study selection criteria

This review will consider all relevant primary research studies using the following inclusion/exclusion criteria:

Inclusion criteria

Subjects

- Trainees in health professions education (allopathic medicine, osteopathic medicine, allied health professions, dentistry, nursing)
- Trainees at all levels of health professions education (undergraduate, graduate, and continuing education programs)

AND

Intervention

- Test enhanced learning defined as one or more "tests" intended to enhance learning by the retrieval effect.
- Test enhanced learning using any other types of assessment

AND

Study Methods

- Experimental trials or observational studies of TEL
- May include controlled or uncontrolled studies
- May include studies with co-interventions

AND
Outcome

- Learning as documented assessed by an examination (OR)
- Learning documented by any other type of assessment

Exclusion criteria

- Studies of *indirect* effects of assessment on learning (*selection effect* or *rehearsal effect*)
- Purely descriptive or opinion pieces
- Reviews

*No study will be excluded on the grounds of study design, geographical location or, as far as is possible, language.

Reproducibility of the inclusion process

Two readers will independently screen the title and abstracts to winnow the search output to a smaller number of articles for full-text reviews. Two readers will also perform the full text review, indicating the articles for inclusion and exclusion, including reasons for exclusions. We will determine the level of agreement with a Kappa statistic and resolve differences by consensus. Differences will be resolved by consensus.

6. Assessment of study quality

While no study will be excluded from the review based on study design, we plan to assess the methodologic quality of included studies. Analyzing the subgroup of higher quality studies may allow higher confidence inferences. We will determine the MERSQI[14-16] score for each of the included studies. Two raters will independently extract data for the MERSQI score and subscores. Differences will be resolved through consensus. We do not plan to weight studies based on quality scores in the meta-analysis. Experience in clinical meta-analyses shows no consistent relationships between quality scores and pooled estimates of treatment effects.[17,18]

7. Procedure for extracting data

We will develop an extraction form and pilot it using non-included studies. The pilot will determine the clarity of the items and calibrate raters by resolving inter-rater differences. Once the abstraction form is finalized, two raters will abstract data from the full text of each included article. We will determine appropriate statistical measures of agreement (such as the Kappa statistic). Differences will be resolved through consensus.

We will extract the following data from the included studies:

Data extraction Form

General information	(1) Year of study	
	(2) Location (country)	
	(3) Health profession (*circle one*)	Allopathic medicine
		Osteopathic medicine
		Nursing
		Pharmacy
		Dentistry
		Allied health professions
		Complimentary medicine
	(4) Level of training (*circle one*)	Undergraduate education
		Graduate education
		Continuing education
	(5) Discipline for GME or CME	(*text*)
		NA
Intervention ("tests")	(6) Content or objectives of initial teaching session	(*text*)
	(7) Knowledge or skills	knowledge
		skills
	(8) Duration of teaching session	__ hours
	(9) Format of teaching session	lecture
		Interactive / participatory sessions
		OSCE / SP
		TBL
		Simulation
		Other (*text*)
	(10) Type of assessment (*circle one*)	Examination (TF or MCQ)
		Examination (short answer or essay)
		Standardized patient (OSCE)
		Direct observation (work based)
		Simulation
		Other
	(11) Number of items	
	(12) Scoring	Answer key (TF or MCQ)
		checklist
		Global rating
		other
	(13) Number of tests[1]	
	(14) Intervals	Exact timing
	(15) Intervals (*circle one*)	Days (7 or less)
		Weeks (4 or less)
		Months

[1] Not including final test take by both cases (TEL) and controls) studying

	(16) Duration of testing	Exact timing
	(17) Duration of testing (*circle one*)	Days (7 or less)
		Weeks (4 or less)
		Months
	(18) Co-interventions (*circle one*)	None
		Feedback
		Self-explanation
		Other
	(19)	Type of feedback, (*text*)
	Control intervention	(20) Study sheet: yes no
		(21) other (*text*)
		(22) Identical material to TEL tests: yes no
	Sample size	(23) Total
		(24) Cases
		(25) Controls
	(26) Study design (*circle one*)	RCT
		Trial Controlled (simultaneous)
		Trial Controlled (historical)
		Trial Uncontrolled
		Observational Controlled (simultaneous)
		Observational Controlled (historical)
		Observational Uncontrolled
		Other
Outcome	Timing	(27) From end of teaching session: __ days
		(28) From last TEL "test": __ days
	Type of learning	Immediate recall
		Retention
		Transfer
	(29) Type of assessment (*circle one*)	Examination (TF or MCQ)
		Examination (short answer or essay)
		Standardized patient (OSCE)
		Direct observation (work based)
		Simulation
		Other (*text*)
	(30) number of items	
	(31) scoring	Answer key (TF or MCQ)
		checklist
		Global rating
		Other (*text*)
	(32) Validity evidence	Yes or no
	(33) Validity evidence (*circle all that apply)*	Internal consistency reliability
		Inter-rater reliability
		Content
		Response processes

		Internal structure (internal consistency)
		Internal structure (dimensionality)
		Other variables (criterion)
		Other variables (discrimination)
		Other variables (response)
		Consequential
		Other
	(34) Validity evidence data	(*text*)
	Adjust for confounding	(35) Yes or no
		(36) Compare frequencies: yes or no
		(37) regression: yes or no
		(38) list confounding variables: (*text*)
	Magnitude and significance of effect	(39) Case score
		(40) Case confidence interval
		(41) Case standard deviation
		(42) Case standard error
		(43) Control score
		(44) Control confidence interval
		(45) Control standard deviation
		(46) Control standard error
		(47) Absolute difference
		(48) Relative difference
		(49) Difference p value
		(50) Difference confidence interval
		(51) Difference standard deviation
		(52) Difference standard error
		(53) Standardized mean difference*
		(54) Statistical test
Quality Score		(55) Overall Score
		(56) study design score
		(57) sampling score
		(58) type of data score
		(59) validity of evaluation score
		(60) data analysis score
		(61) outcomes score
		Comments

*To allow comparison among heterogeneous studies, we will determine a standardized effect size (standard mean difference SMD)[19] if the study provides the requisite data. For a comparison of means:

$SMD = (mean_1 - mean_2) / SD_{pooled}$

If variance data is not provided, we will determine estimates[20] for standard deviations in SMD determinations.

8. Synthesis of extracted evidence

Systematic Review

We will provide descriptive statistics to demonstrate an overview of the range and categories of TEL interventions and outcomes. We will also compare subgroups to make qualitative inferences to presumptively answer our study questions. For instance, we might find that studies with frequent and briefer TEL interventions appear to show greater effects on learning. Or we might find that certain co-interventions enhance TEL.

Meta-analysis

The learning outcomes of our included studies will clearly be too heterogeneous to combine directly. We will determine the standardized mean difference (also called Cohen's effect size) for each study if sufficient information is provided. This unitless measure of "impact" an be compared and combined among studies. We will than determine if statistical (Cochran's Q and I^2)[21] heterogeneity are sufficiently low to permit a quantitative synthesis (meta-analysis). If heterogeneity is not prohibitive, we will use a random-effects model to pool weighted outcomes of standardized mean differences (SMD).

Mapping Review

We will characterize the quantity, quality, and focus of the existing scientific literature on TEL. We expect to identify gaps that will help set an agenda for future research.

9. Scoping search

In order to evaluate the availability of evidence and develop a potential final search strategy, we conducted scoping searches in May 2016 in the following databases. Preliminary results appear below:

MEDLINE (Ovid)	490
Embase (Ovid)	656
CINAHL with Full Text (EBSCO)	868
ERIC (ProQuest)	6774

The scoping searches yielded sufficient results to proceed with this Systematic Review across these and the other databases mentioned in the Database Searching section above. We will refine these strategies and translate them into the proper syntax required for each of the following databases: MEDLINE, Embase, PsycINFO, AMED, CINAHL with Full Text, ERIC, and Education Research Complete. We believe that customizing the search strategies to the exact constraints of each of the databases will result in a reasonable retrieval set. Based on a cursory look at the captured papers, we estimate that 40 to 60 studies will ultimately be included in the systematic review.

10. Translation into practice

We suspect many educators remain unaware of the "retrieval effect" and TEL, as some studies have languished in obscure medical education or psychology journals. Publication of this review will illuminate the potent pedagogical potential of assessment. Furthermore, this review will provide practical guidelines for educators seeking to implement TEL interventions. Our review may reveal that certain types of testing, testing conditions, and co-interventions may be more suited to different learners, settings, or learning outcomes. Finally, students should find gratification in the knowledge that assessments are not merely administrative exercises imposed by external stakeholders. On the contrary, assessments promote learning in ways that studying cannot.

11. Project timetable

	May 2016	June 2016	July 2016	August 2016	September 2016	October 2016	November 2016	December 2016
BEME registration								
BEME protocol								
Literature search								
Inclusion								
Data extraction								
Quality ratings								
Descriptive tables								
Quantitative synthesis								
Manuscript writing								

12. Conflict of interest statement

None of the authors have any financial, institutional, political, personal or other conflicts of interest.

13. Plans for updating the review

Before and after we begin the manuscript writing process, we will rerun our search strategy to ensure that we include the most recent articles. We will also set up auto-alerts for our search strategies to capture the latest articles as they get indexed.

14. Changes to the Protocol–

Any changes to the protocol will be recorded as amendments and communicated to BEME with a rationale.

References

1. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher.* 2011;33(6):478-485.
2. Butler AC. Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition.* 2010;36(5):1118-1133.
3. Butler AC, Karpicke JD, Roediger HL, III. Correcting a Metacognitive Error: Feedback Increases Retention of Low-Confidence Correct Responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition.* 2008;34(4):918-928.
4. Roediger HL, III, Karpicke JD. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science.* 2006;17(3):249-255.
5. Cepeda NJ, Vul E, Rohrer D, Wixted JT, Pashler H. Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychol Sci.* 2008;19(11):1095-1102.
6. Halpin G, Halpin G. Experimental investigation on the effects of study and testing on student learning, retention, and ratings of instruction. *Journal of Educational Psychology.* 1982;74(1):32-38.
7. Larsen DP, Butler AC, Roediger HL, 3rd. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Med Educ.* 2009;43(12):1174-1181.
8. Messineo L, Gentile M, Allegra M. Test-enhanced learning: analysis of an experience with undergraduate nursing students. *BMC medical education.* 2015;15(1):182.
9. Kromann CB, Bohnstedt C, Jensen ML, Ringsted C. The testing effect on skills learning might last 6 months. *Advances in health sciences education : theory and practice.* 2010;15(3):395-401.
10. Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills learning. *Med Educ.* 2009;43(1):21-27.
11. Baghdady M, Carnahan H, Lam EWN, Woods NN. Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Medical Education.* 2014;48(2):181-188.
12. Larsen DP, Butler AC, Lawson AL, Roediger HL, 3rd. The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in health sciences education : theory and practice.* 2013;18(3):409-425.
13. Kreiter C, Green J, Lenoch S, Saiki T. The overall impact of testing on medical student learning: quantitative estimation of consequential validity. *Adv in Health Sci Educ.* 2013;18(4):835-844.
14. Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. 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.* 2008;23(7):903-907.
15. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association Between Funding and Quality of Published Medical Education Research. *JAMA.* 2007;298(9):1002-1009.
16. Cook DA, Reed DA. Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle–Ottawa Scale-Education. *Acad Med.* 2015;90(8):1067-1076.
17. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282(11):1054-1060.

18. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology.* 2006;59(12):1249.e1241-1249.e1211.
19. Hojat M, Xu G. A Visitor's Guide to Effect Sizes - Statistical Significance Versus Practical (Clinical) Importance of Research Findings. *Adv Health Sci Educ.* 2004;9(3):241-249.
20. Higgins JPT DJ. Chapter 7: Selecting studies and collecting data. In: Higgins JPT GS, editors, ed. *Cochrane Handbook for Systematic Reviews of Interventions* 2011.
21. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557-560.