A Research Synthesis Estimating the Overall Quantitative Impact of Educational Assessment on Medical Student Learning

Clarence D. Kreiter

Visiting Professor IRCME

University of Tokyo – Univ. of Iowa

Dr. Takuya Saiki – Gifu University

Context

- Medical Schools Allocate Many Resources to Assessment
- Medical Education Research Places a Strong Emphasis on R and D of Testing
- Medical Education Journals and Conferences continue to focus on Testing and Assessment topics – AAMC / RIME – Ottawa Conference

Context (con’t)

- Educational reform efforts often use assessments as a tool for reform, however
- Strong criticism of many types of testing
- Currently benefits taken on faith – need evidence-based justification to address criticism

Research Questions

- Are educational assessments an effective learning tool in medical education?
- Is research on educational assessment likely to promote educational efficiency?

How to Estimate?

- RCT – Split class into random halves – remove all assessment influences on one half – compare experimental and control
- Not doable for ethical and practical reasons
- Another approach?

Mechanism of Impact

- Review of the literature identifies 3 main ways that assessment is hypothesized to impact medical education (next slides define)
  - Direct Effect
  - Indirect Effect
  - Selection Effect

Direct Effect

- Reflects learning that occurs as part of test’s intrinsic influence on long-term retention
- Hypothesized to have mnemonic effects
- Retrieval
- Mostly unrealized potential

Indirect Effect

- Associated with summative course and licensure testing
- Operates extrinsically on learning by motivating learners and instructors
- Enables accountability mechanism
- Partly realized potential?

Selection Effect

- Gains observed by using aptitude tests to select those most likely to excel in medical education

- Mostly realized potential

Using Estimates in the Literature

- Meta-analytic approach to summarize effects of each testing effect on learning

- Effect size - standardized, scale-invariant measure to summarize and integrate studies

Effect Size – Language of Meta-Analysis

- Cohen's **d**

$$\mathbf{d} = (\mathit{Mean}_1 - \mathit{Mean}_2) / \mathit{SD}_{\mathit{Pooled}}$$

- Correlation changing r to **d**

$$\mathbf{d} = 2\,r / \sqrt{}\,(1 - r^2)$$

Literature Search

- Key words poorly defined so three methods:

    1. Traditional ERIC – Medline - PsychInfo……

    2. Ancestry approach

    3. Reverse ancestry approach (Google Scholar©)

Study Inclusion Criteria

- Conducted in-vivo
- Conducted in medical education
- When M.E. evidence limited – studies of college-level learners included
- Quantitative estimates of learning gains that can be translated into effect size (**d**)

Estimates in Literature

- Combine mean test effects to derive total potential learning effect

- $TE^{*} = (\text{đ}_{Direct}) + (\text{đ}_{Indirect}) + (\text{đ}_{Selection})$

**TE = Total Effect*

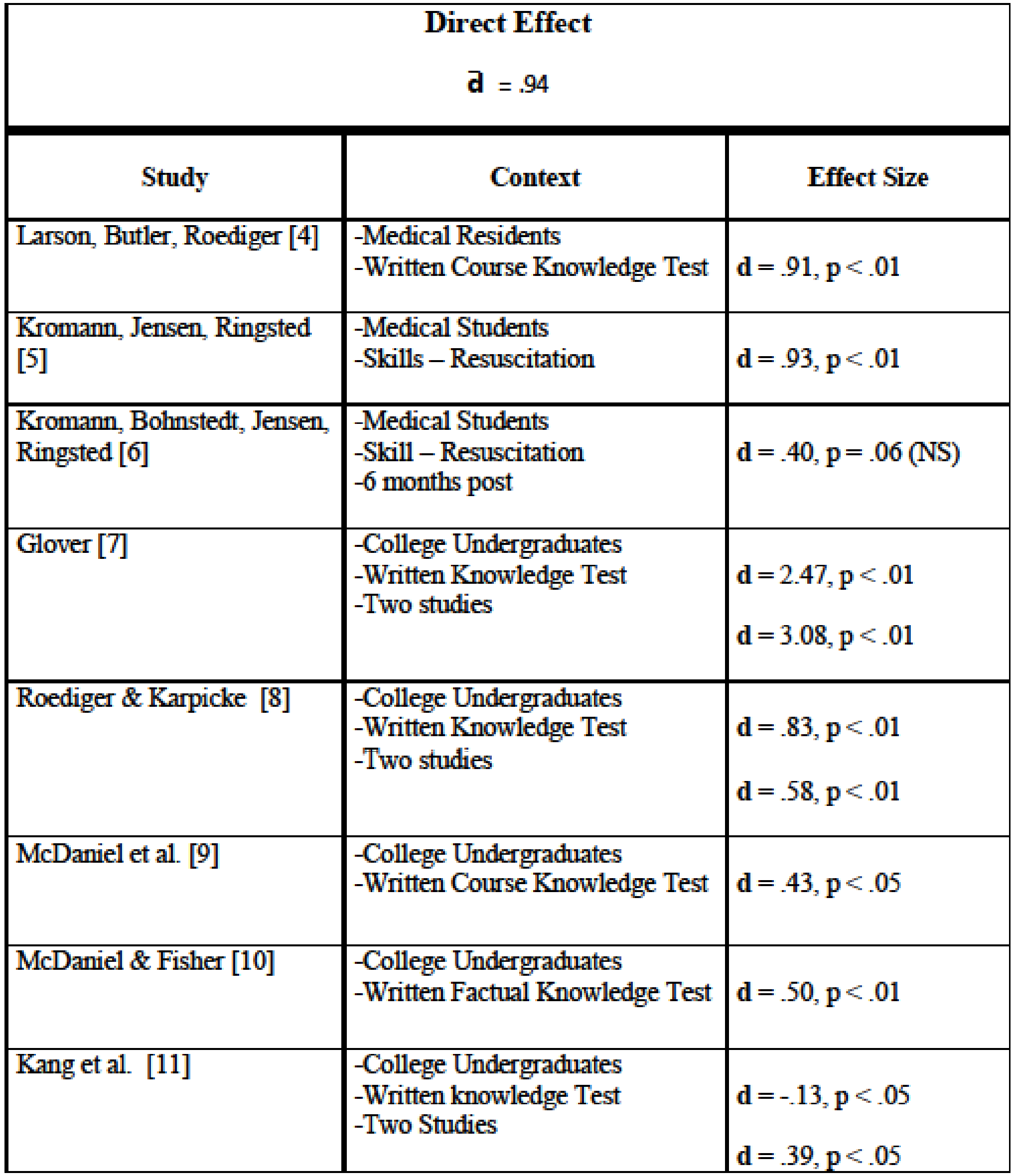
Evidence Direct Effect

- Three Studies in Medical Education (d = .91,.93,.40)
  - Larsen, D.P., Butler, A.C., Roediger, H.L. III. Repeated testing improves long-term retention relative to repeated study: a randomized controlled trial. *Medical Education,* 2009; 43:1174-1181.
  - Kronmann, C.B., Jensen, M.L., Ringsted, C. The effect of testing on skills learning. *Medical Education,* 2009; 43:21-27.
  - Kronmann, C.B., Bohnstedt, C. Jensen, M.L., Ringsted, C. The testing effect on skill learning might last 6 months. *Adv. Heal. Sci. Ed. Theory and Prac.,* 2010; 15(3):395-401.

- Many Studies in Psych and Education
  - Most laboratory-type learning task
  - 5 using undergrads and educationally relevant task (d = 2.4 (3.08), .83 (.58), .43, .50, -.13 (.39))

- Mean Effect Size = **đ** = .94.

Table 1



Direct Effect

$\bar{d}$ = .94

Study	Context	Effect Size
Larson, Butler, Roediger [4]	-Medical Residents -Written Course Knowledge Test	d = .91, p < .01
Kromann, Jensen, Ringsted [5]	-Medical Students -Skills – Resuscitation	d = .93, p < .01
Kromann, Bohnstedt, Jensen, Ringsted [6]	-Medical Students -Skill – Resuscitation -6 months post	d = .40, p = .06 (NS)

Glover [7]	-College Undergraduates
	-Written Knowledge Test
	-Two studies
		d = 2.47, p < .01
		d = 3.08, p < .01

Roediger & Karpicke [8]	-College Undergraduates
	-Written Knowledge Test
	-Two studies
		d = .83, p < .01
		d = .58, p < .01

McDaniel et al. [9]	-College Undergraduates -Written Course Knowledge Test	d = .43, p < .05
McDaniel & Fisher [10]	-College Undergraduates -Written Factual Knowledge Test	d = .50, p < .01

Kang et al. [11]	-College Undergraduates
	-Written knowledge Test
	-Two Studies
		d = -.13, p < .05
		d = .39, p < .05

Evidence of Indirect Effect

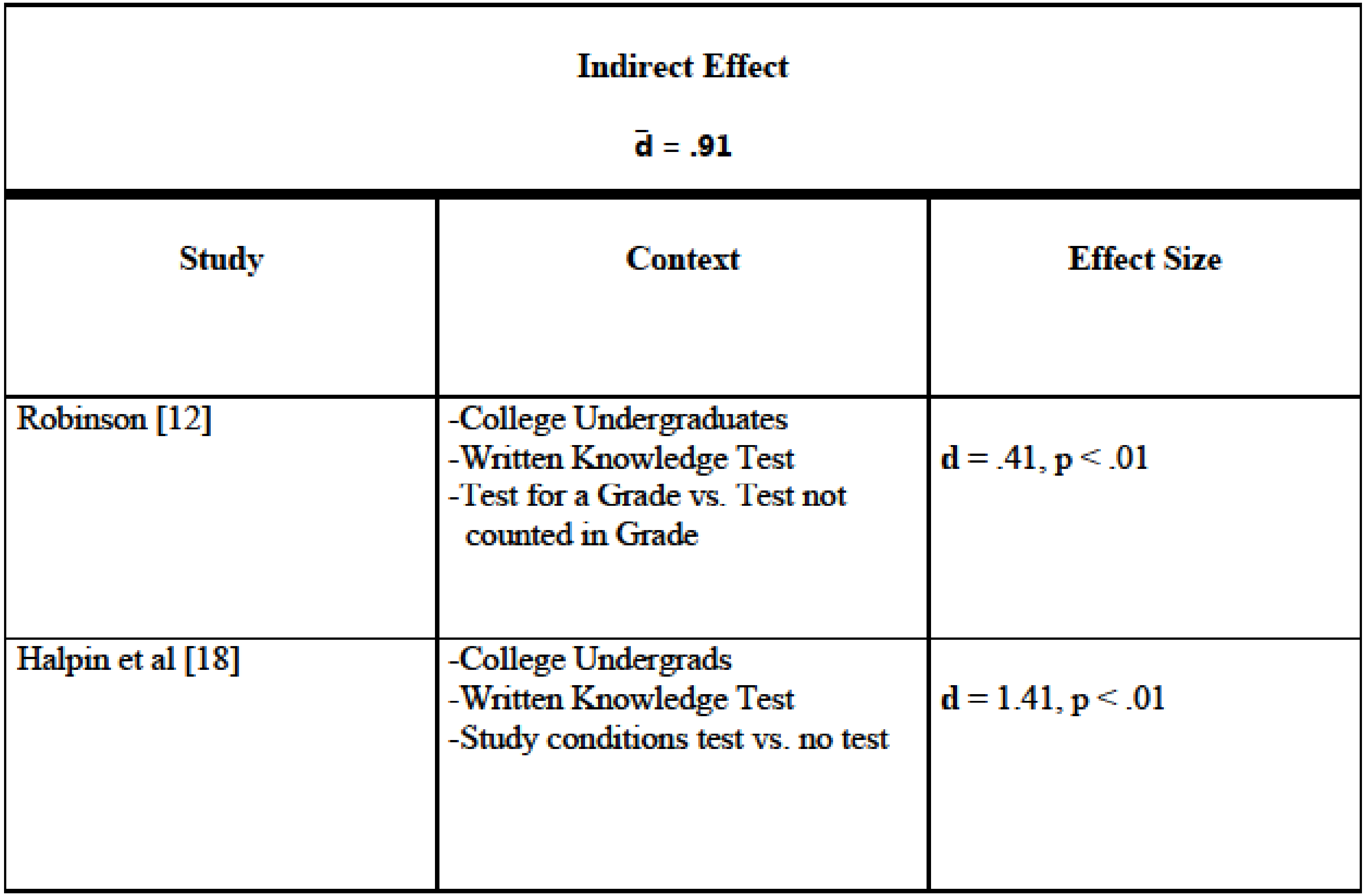
- Ubiquitous in Med Ed
- Yet no research in Medical Education
- Ethical and Methodological Challenges

Evidence for Indirect Effect

- The Effect of Testing on Achievement: Meta-Analyses – 1910-2010 : Phelps Richard **(In Press) – <u>Estimates Unrelated to ME</u>**
  - **N Studies = 170**
  - **High Stakes Testing d ~ .80 (Grade School etc)**

- Only two studies by Robinson (1972) & Halpin et al (1982)

Table 2



Indirect Effect **$\bar{d}$ = .91**		
Study	**Context**	**Effect Size**
Robinson [12]	-College Undergraduates -Written Knowledge Test -Test for a Grade vs. Test not counted in Grade	$d = .41, p < .01$
Halpin et al [18]	-College Undergrads -Written Knowledge Test -Study conditions test vs. no test	$d = 1.41, p < .01$

Evidence of Indirect Effect - questions

- Cultural differences?
- Speculation on what would happen without accountability enforced by testing…..?
- Course-based tests vs. national licensure testing
- Ways to find out?
- Qualitative Studies?

Selection Effect - Pre-existing Summaries and Other Research

- Julian, E.R. Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine,* 2005;80(10):910-917.
- Kreiter, C.D., Kreiter, Y. A validity generalization perspective on the ability of undergraduate GPA and the Medical College Admission Test to predict important outcomes. *Teaching and Learning in Medicine,* 2007;19(2):95-100.
- Donnon, T., Paolucci, E.O., Violato, C. The predictive validity of the MCAT for Medical School Performance and Medical Board Licensing Examinations: A meta-analysis of the published research. *Academic Medicine,* 2007;82(1):100-106.
- Reibnegger, G., Caluba, H.C., Ithaler, D., Manhal, S. Neges, H.N., Smolle, J. Progress of medical students after open admissions based on knowledge tests. *Medical Education,* 2010;44:205-214.

Julian Study

- 14 Medical Schools
- Med Schools Grades
- USMLE
- Range Restriction
- r = .63
- d = 1.63

Donnon et al Study

- Meta-analysis
- 23 Studies
- 1991 Version of MCAT
- USMLE
- Range Restriction
- r = .48
- d = 1.09

Kreiter et al. Study

- 29 Studies
- All versions MCAT
- Clinical Skills
- Written Tests
- Reliability Attenuation
- r = .47
- d = 1.07

Table 1. *Correlation of Outcomes With MCAT and uGPA*

Outcome/Domain (w)ritten (p)Nonwritten/Performance	Time Since Adm.	Rel. of Outcome	N	MCAT [RR Corr]*	uGPA [RR Corr]*	Both [RR Corr]*	Ref. No.
Med Sch. GPA (w)	Yr 1	.70	12 Schools n > 1,200	.54 [.66]	.40 [.53]	.64 [.73]	18
Med. Sch. Grades (w)	Yr 1–2	.70	14 Schools n > 1,400	.51 [.64]	.49 [.58]	.66 [.76]	19
Lit Review Pre-1990 Basic Sci. (w)	Yr 1–2	.70	18 Studies n > 3500			.48	17
Step 1 USMLE (w)	Yr. 2	.96	27,406	.53 [.70]	.37 [.49]	.55 [.72]	20
NBME I (w)	Yr 2	.90	1628	.45		.49	21
Lit Review Pre-1990 NBME I (w)	Yr 2	.90	16 Studies n > 4000	.58		.62	17
Step 1 USMLE (w)	Yr 2	.96	14 Schools n > 1,400	.54 [.72]	.36 [.48]	.58 [.75]	19
Step 1 USMLE (w)	Yr. 2	.96	24,000	.57	.42	.60	11
MCCE Part 1 (w)	Yr 4	.95	597			.48	23
MCCE Part 2 (w)	Yr 4	.85	597			.34	23
OSCE (p)	Yr 4	.67	137	.30	.33	.36	6
Step 2 – USMLE (w)	Yr 4	.90	26,752	.49 [.60]	.33 [.44]	.52 [.63]	20
Lit Review Pre-1990 NBME II (w)	Yr 4	.90	8 Studies n > 1500			.52	17
NBME II (w)	Yr 4	.90	1628	.42		.46	21
LMCC Part I (w)	Yr 4	.90	75	.33	.33	.36	7
Certification Exams (w)	Yr 6	.90	857	.33	.33	.40	24
LMCC Part II—OSCE (p)	Yr 6	.70	44	.07	.25	.27	7
Lit Review Pre-1990 NBME III (w)	Yr 6	.90	2 Studies n > 300			.35	17
NBME III (w)	Yr 6	.90	1188	.30		.34	21
Step 3—USMLE (w)	Yr 6–7	.90	25,170	.49 [.62]	.29 [.42]	.52 [.64]	20
Physician Disciplined [t] d = .33&.40 (p)	Yr 8–30	?	704	.30 {.15}[t]	.25 .18[t]	.34	14

t – Effect size d = .33 and .40 converted to r and corrected for dichotomization (split = 90/10).

*Reported corrected for range restriction – Range restricted value was not used in meta-analysis.

Table 2. *VG Summary Table Average Corrected Multiple (MCAT & uGPA) Correlation Coefficients*

	Written Tests of Knowledge and Clinical Reasoning				Non-Written Testing of Clinical Skills			
Attainment Level	**r**	$\bar{r}$	$\bar{r}_{yy}$	$\bar{r}_c$	r	$\bar{r}$	$\bar{r}_{yy}$	$\bar{r}_c$
Yr 1 & 2	.64 .66 .48 .55 .49 .62 .58 .60	.56	.85	.61				
Yr 3 & 4	.48 .34 .52 .46 .36	.52	.90	.58	.36	.36	.67	.44
Yr 5, 6 & 7	.40 .35 .34 .52	.51	.90	.54	.27	.27	.70	.33
Professional Practice Yrs > 7	?	?	?	?	.34		?	>.34

Reibnegger et al. Study

- Before and After the Use of Selection tests

  - Austrian medical school before and after

  - ~23 % vs. ~82 % on-time completions

  - large decrease in dropouts

  - chi – square (p <.0001)

  - d = 1.15

(Reibnegger, G., Caluba, H.C., Ithaler, d., Manhal, S., Neges, H.N., Smolle, J. Progress of medical students after open admission or admission based on knowledge tests. *Medical Education*, 2010:44:205-214.)

Table 3

<table>
<tr><th colspan="3"><b>Selection Effect</b>

$\bar{d}$ = <b>1.26</b></th></tr>
<tr><th>Study</th><th>Context</th><th>Effect Size</th></tr>
<tr><td>Donnon et al. [13]</td><td>-23 studies
-Medical Students
-Current Version of MCAT
-Med School Performance
-USMLE
-Range Restrict. Correction</td><td>r = .43 preclinical
r = .39 clerkship
r = .66 USMLE 1
r = .43 USMLE 2
r = .48 USMLE 3
Mean r = .48

<b>d</b> = 1.09</td></tr>
<tr><td>Kreiter &amp; Kreiter [14]</td><td>-29 studies
-Medical Students
-Current/past Ver. MCAT
-Undergrad GPA
-Written Testing Outcomes
-Clinical Skill PBA Outcomes
-Post Grad Performance
-Rel. Attenuation Correct</td><td>r = .61 yr1-2 written
r = .58 yr3-4 written
r = .54 yr5-7 written
r = .44 yr3-4 clinical
r = .33 yr5-7 clinical
r = .34 yr7+ clinical
Mean r = .47

<b>d</b> = 1.07</td></tr>
<tr><td>Julian [15]</td><td>-14 Medical Schools
-Med School Grades
-USMLE
-Range Restrict. Correction</td><td>r = .59 Med School Grd
r = .70 USMLE 1
r = .60 USMLE 2
r = .62 USMLE 3
Mean r = .63

<b>d</b> = 1.63</td></tr>
<tr><td>Reibnegger et al. [16]</td><td>-Medical School Before and After Test Selection
-Successful completion of study</td><td>χ^2 = 631.44, df = 1, p &lt; .0001
Mean r = .49

<b>d</b> = 1.15</td></tr>
</table>

More Real World Evidence for Selection Effect

- Variance above cut score ~.9
  - Cut Score Study
  - Cut Score MCAT = 24 / Cut Score SciGPA = 3.0

(Kreiter, C.D. A commentary on the use of cut-scores to increase the emphasis on non-cognitive variables in medical school admission – *Advances in Health Science Education*, 2006,12:315-319)

Preliminary Estimate

- Total Potential = ~.94 + ~.91 + ~1.24 = **3.09**
- Too good to be true?
- Take some effects for granted
  - Selection
  - Accountability
- Potential vs. Realized

Validity of Model

- Does equation apply?
- Are Effects Logically additive and independent?

***Total Contribution** = (Direct Effect) + (Indirect Effect)+ (Selection Effect)*

Research Questions and Answers

- Are educational assessments an effective learning tool in medical education? Yes
- Is research on educational assessment likely to promote medical education efficiency? Yes

Conclusions

- Strong evidence for testing’s ability to promote learning
- Gains only partially realized in many medical education programs
- Continued improvement in testing methods likely to yield considerable gains in learning

Conclusions

- Likely source of inexpensive and effective innovation
- Some new opportunities presented by electronic delivery– (direct effect especially)
  - Intelligent Tutoring with well timed assessment
    - (Crowley, R.S., Medvedeva, O. An intelligent tutoring system for visual classification problem solving. *Artificial Intelligence in Medicine*. 2006;36(1):85-117.
  - LabCAPS
    - (Kreiter, C. et al. A report on the piloting of a novel computer-based medical case simulation for teaching and formative assessment if diagnostic laboratory testing *Medical Education Online, 2010;15)*

Questions - Skepticism

- Wild Estimate
- Delusions of grandeur

Questions – Skepticism

- Evaluation role of findings
  - Remind Educators
  - Provide academic decision makers with hard evidence
- ? Question???
- ?
- ?
- ?