Interpretation of Results of Research in Medical Education: Threats to Validity and Magnitude of Effect

Jerry A. Colliver, Ph.D.

Southern Illinois University

School of Medicine

Society of Directors of Research in Medical Education

Santa Fe, New Mexico

June 16, 2009

Purpose:

- to address the reputation of research in medical education
- to clarify what I think is the basis for this reputation
- to consider how to turn this around

Interpretation of Results of Research in Medical Education: Threats to Validity and Magnitude of Effect

Reputation of Research

Kaestle CF. The awful reputation of education research. *Educational Researcher* 1993;22:23-31.

Whitcomb ME. Research in medical education: What do we know about the link between what doctors are taught and what they do? *Academic Medicine* 2002; 77: 1067-1068.

Bligh J. Editorial. *Medical Education* 2003; 37: 184-5.

Colliver JA. The research enterprise in medical education. *Teaching and Learning in Medicine* 2003;15:154-155.

Lurie SJ. Raising the passing grade for studies of medical education. *JAMA* 2003;290:1210-1212.

Albert M, Hodges B, Regehr G. Research in medical education: Balancing service and science. *Advances in Health Sciences Education* 2006;12:103-115.

Schuwirth LWT, van der Vleuten CPM. Challenges for educationalists. *BMJ* 2006;333:544-546.

Todres M, Stephenson A, Jones R. Medical education research remains the poor relation. *BMJ* 2007;335:333-335.

Norman G. Editorial – How bad is medical education research anyway? *Advances in Health Sciences Education* 2007;12:1-5.

Colliver JA, McGaghie WC. The Reputation of Medical Education Research: Quai-experimentation and Unresolved Threats to Validity. *Teaching and Learning in Medicne* 2008;20:101-103.

- Critics say the research is methodologically flawed, because it lacks randomization and control

- I agree because research in medical education is mostly quasi-experimental

- which by definition lacks randomization and control

Campbell DT, Stanley JC. ***Experimental and Quasi-experimental Designs for Research.*** **Chicago; Rand McNally, 1966.**

Cook TD, Campbell DT. ***Quasi-experimentation: Design and Analysis Issues for Field Settings.*** **Chicago, Rand McNally, 1979.**

Shadish WR, Cook TD, Campbell DT. ***Experimental and Quasi-experimental Designs for Generalized Causal Inference.*** **Boston, Houghton Mifflin 2002.**

Quasi-experimentation

- to provide an alternative to randomized controlled trials
for researchers in applied field settings,
like psychology and education

- in which randomization and control are limited
if not impossible due to practical and ethical constrains

- so quasi-experimentation would seem to be
made to order for research in medical education.

TABLE 4.1 Quasi-Experimental Designs Without Control Groups

The One-Group Posttest-Only Design

$X \quad O_1$

The One-Group Posttest-Only Design With Multiple Substantive Posttests

$X_1 \quad \{O_{1A}\ O_{1B} \ldots O_{1N}\}$

The One-Group Pretest–Posttest Design

$O_1 \quad X \quad O_2$

The One-Group Pretest–Posttest Design Using a Double Pretest

$O_1 \quad O_2 \quad X \quad O_3$

The One-Group Pretest–Posttest Design Using a Nonequivalent Dependent Variable

$\{O_{1A}, O_{1B}\} \quad X \quad \{O_{2A}, O_{2B}\}$

The Removed-Treatment Design

$O_1 \quad X \quad O_2 \quad O_3 \quad \not{X} \quad O_4$

The Repeated-Treatment Design

$O_1 \quad X \quad O_2 \quad \not{X} \quad O_3 \quad X \quad O_4$

TABLE 4.2 Quasi-Experimental Designs That Use Control Groups But No Pretest

Posttest-Only Design With Nonequivalent Groups

NR	X	O_1
NR		O_2

Posttest-Only Design Using an Independent Pretest Sample

NR	O_1	X	O_2
NR	O_1		O_2

Posttest-Only Design Using Proxy Pretests

NR	O_{A1}	X	O_{B2}
NR	O_{A1}		O_{B2}

Quasi-Experimentation

First Part – The designs

Second Part –

- **The crucial part of quasi-experimentation that attempts to make up for its shortcomings**
- **The part that addresses the problems created by the lack of randomization and control**
- **Involves the "theory of the threats to validity"**
- **Threats to the validity of the research conclusion**

TABLE 1

SOURCES OF INVALIDITY FOR DESIGNS 1 THROUGH 6

	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Pre-Experimental Designs:												
1. One-Shot Case Study X O	−	−				−	−			−		
2. One-Group Pretest-Posttest Design O X O	−	−	−	−	?	+	+	−		−	−	?
3. Static-Group Comparison X O --- O	+	?	+	+	+	−	−	−		−		
True Experimental Designs:												
4. Pretest-Posttest Control Group Design R O X O R O O	+	+	+	+	+	+	+	+		−	?	?
5. Solomon Four-Group Design R O X O R O O R X O R O	+	+	+	+	+	+	+	+		+	?	?
6. Posttest-Only Control Group Design R X O R O	+	+	+	+	+	+	+	+		+	?	?

Threats to Validity

- **Flaws and confounds**
- **Must be ruled out to support research conclusion**
- **Must be empirically tested and ruled out**

Otherwise, results are just as plausibly attributed to threats as to research hypothesis

The problem is NOT that we are using flawed research methods

-WHICH WE ARE-

but that we are disregarding the flaws in drawing research conclusions

Example 1.

A study of the effectiveness of problem-based learning found the PBL students performed better on:

- **Step 1 (d = .18)**
- **Step 2 (d = .39)**
- **clerkship ratings (d = .50)**
- **and post clerkship SP exam (d=.30)**
- **MCAT (d= .46)**

Example 2.

Another Non-randomized comparison of PBL and standard track students NBME Part III (d= +.33)

- **A subgroup of students who requested PBL were randomly assigned NBME Part III (d= -.33)**

- **Conclusion: "in the long run, the more student-centered problem-based curriculum better prepared the students for NBME III."**

Dochy F, Segers M, Ven den Bossch P, Gijbels D. Effects of Problem-based Learning: a Meta-analysis. *Learning and Instruction* 2003;13:553-68.

Gijbels D, Dochy F, Van den Bossch P, Segars M. Effects of Problem-based Learning: A Meta-analysis from the angle of assessment. *Review of Educational Research* 2005;75:27-61

Table 1. Summary of Design, Biases/Confounds, Threats to Validity, and Effect Size (d) of Studies in "Principles" Category of Gijbels, et al. Meta-analysis.

Studies from Giibles et al., Meta-analysis	***n***		***Comparison of***					***Biases / Confounds***				***Validity Threats***		**Outcome Effect Sizes *(d)***
	PBL	**LBL**	**Two Tracks**	**Consec-utive Classes**	**Different Schools**	**Small Subgroups of Curriculum Groups**	**Random-ized Groups**	**Selection Bias**	**Intervention / Outcome Confound**	**Time on Task Confound**	**Testing Lag & Response Rates**	**Favor Interven-tion**	**Not Ruled Out**	
#1.Boshuizen[16]	4	4			X	X		?	X			X	X	2.268
#2.Distlehorst[17]	47	154	X					X				X	X	0.445
#3. Doucet[18]	21	26	X					X	X	X		X	X	1.293
#4. Finch[19]	21	26		X					X			X	X	1.904
#5. Goodman[20]	36	297	X					X				X	X	-0.133
#6. Hmelo[21]	20	20				X*		X	X			X	X	0.7305
#7. Hmelo[22]	39	37	X			X		X	X			X	X	0.768
#8.Martenson[23]	1,651	818		X							X	X	X	0.00
#9. Mennin[24]	144	447	X					X				X	X	0.046
(randomized)	(67)	(27)	(X)				(X)					NA	NA	(-0.16)
#10. Richards[25]	88	364	X					X				X	X	0.3375
#11 Schmidt[26]	PBL ~204	Org Syst ~204			X		X					NA	NA	0.310
		Lect ~204												

*PBL elective vs. non PBL elective

mean = .795

NA: not applicable

- **Our "tentative conclusion is that the field of medical education might be better served in most instances by systematic narrative reviews that describe and critically evaluate individual studies and their results in light of threats to their validity."**

Colliver JA, Kucera K, Verhulst S. Meta-analysis of quasi-experimental research: Are systematic narrative reviews indicated? Medical Education 2008;42:858-825

Magnitude of Effect

Statistical Significance

- **unlikely results are due to sampling error**
- **nice to know, but more is needed to provide insight into practical value of results**

Effect-size Measures

- **are in standardized units**
- **may not give picture of meaning of results for practical purposes**

Research on Empathy

A DISTURBING CONCLUSION

- **empathy is said to decline during medical school and residency training**

A STRONG REACTION

- **Is there hardening of the heart during medical school?**
- **Vanquishing virtue**
- **the relevant question is not how to create humane qualities but how it comes about that medical education destroys them**

BLAME

- **the medical school selection process**
- **prevalent teaching methods**
- **self protection cynicism**

FRUSTRATION

- **How do we stop the rot?**

The empathy decline studies

- Used three empathy measures
- All were self-report instruments
- Scored by summing ratings across Likert items

First thing we did was transform the sums back to means:

- Which are in the original units of the rating scales
- In order to base conclusions on the scale actually rated by the students and residents
- And to make the magnitude of effect clearer by reporting results in the metric of the original anchors that give meaning to the ratings.
- Also the sums –compared to the means – magnify group differences (declines) and distort the appearance of effects

	Mean Rating (response rate)									Decline in	
	Pre MS	MS1	MS2	MS3	MS4	PGY1	PGY2	PGY3	PGY4	Mean rating	Response rate
1.						3.1 (98%)	3.0 (79%)			-0.1	-19%
2.						3.2 (98%)	3.0 (79%)	2.9 (72%)	2.9 (64%)	-0.3.	-34%
3.						3.2	2.8			-0.4	
4.	3.2			3.0							
5.			6.2	6.0						-0.2	
6.		4.0	4.1		4.1					+0.1	
7.						5.9	5.7	5.7		-0.2	
8.		6.2 (80%)	5.7 (54%)	5.6 (50%)	5.8 (58%)					-0.4	-22%
9.	5.8 (96%)	5.9 (86%)	5.9 (96%)	5.6 (70%)	5.3 (59%)					-0.5	-27%
10.		1.5	1.4	1.5	1.3					-0.2	
11.		1.5	1.3	1.3	1.1					-0.4	
										-0.2 mean	-26% mean

Papadakis MA, Hodgson CS, Teherani A, Kohastu ND. Unprofessional Behavior in Medical School is Associated with Subsequent Disciplinary Action by a State Medical Board. *Academic Medicine* 2004;79:244-249.

Papadakis MA, Teherani A, Banach MA, Knettler TR, Rattner SL, Sterm DT, Veloski JJ, Hodgson CS. Disciplinary Action by Medical Boards and Prior Behavior in Medical School. *New England Journal of Medicine* 2005;335;25:2673-2682.

Papadakis MA, Arnold GK, Blank LL, Holmboe ES, Lipner RS. Performance During Internal medicine Residency Training and Subsequent Disciplinary Action by State Licensing Boards. *Annals of Internal Medicine* 2008;148:869-876.

"disciplinary action by a Medical Board was strongly associated with unprofessional behavior in medical school."

In all three studies:

- **Attributable Risk = 1%**

- **Only one percent of students or residents who tested positive on predictor (in medical school or residency) later subject to Board Disciplinary Action attributable to positive test.**

- **98% who tested positive receive remediations or dismissal even though not subject to future Board Action**

- **For all practical purposes, magnitude of effect shows little of practical value.**

1. **Should these studies have been conducted? Yes, definitely**

2. **Should they have been published? Resounding Yes**

3. **Should the conclusions of the reports better emphasize that the association was very weak and probably of little practical value?**

This is my point YES

Ten years ago, an important study compared students in three different schools

- With different curricula
- Over five years of training
- Using a diagnostic accuracy test as the outcome
- A huge, significant effect of years of training
- Tiny effect of curriculum, but very significant ($p<.0001$).
- Roughly equivalent to only an additional 3 weeks or so of medical school training.

Interpretation of Results of Research in Medical Education: Threats to Validity and Magnitude of Effect

So what can be done about this?

- first, read Campbell and Stanley

So what can be done about this?

- first, read Campbell and Stanley
- second, develop a sensitivity to outcomes – the measures used and the magnitude of effect –to get at the practical consequences of what we can do in medical education
- Medical education research is applied research.