AC 2007-1712: "WHAT WORKS" IN ENGINEERING EDUCATION? A META-ANALYSIS OF VANTH/ERC BIOMEDICAL ENGINEERING MODULES

David Cordray, Vanderbilt University

David S. Cordray PhD is Professor of Psychology and Public Policy at Vanderbilt University. He is currently the Thrust Leader in Assessment and Evaluation for the VaNTH ERC. Professor Cordray has written extensively on research and evaluation methodology in education and human services areas. He has conducted experimental, quasi-experimental and meta-analytic assessments of intervention effectiveness in education, health, welfare, and other human service areas.

Thomas Harris, Vanderbilt University

Jennifer Gilbert is graduate student in the Department of Special Education within Peabody College at Vanderbilt University.

Jennifer Gilbert, Vanderbilt University

Thomas R. Harris MD PhD is the Orrin Henry Ingram Distinguished Professor of Engineering and Professor of Biomedical Engineering, Chemical Engineering and Medicine at Vanderbilt University. He is currently Chair of the Department of Biomedical Engineering. His current interests focus on the development of learning sciences and learning technology for bioengineering. He is currently the director of the National Science Foundation Engineering Research Center in Bioengineering Technologies.

"What Works" in Engineering Education? A Meta-analysis of VaNTH/ERC Biomedical Engineering Modules

Abstract

The *V*anderbilt-*N*orthwestern-*T*exas-*H*arvard/MIT Engineering Research Center (VaNTH/ERC) for Bioengineering Educational Technologies has undertaken a series of studies to examine the effects of instructional innovation on learning outcomes. In this paper we summarize the nature, scope and results of these assessments. In the spirit of identifying evidence-based practices in education, we present estimates of the overall and conditional effects from 28 studies and sub-studies reported in 19 evaluation studies. The results suggest that VaNTH-sponsored innovations are effective, although some of the effects may be exaggerated or understated due to technical and procedural problems. This paper identifies which effects are trustworthy and which require additional examination before they can be incorporated (or not) into the knowledge-base on "What Works" in biomedical engineering education.

I. Introduction

Established in 1999 with a grant from the National Science Foundation, the *V*anderbilt-*N*orthwestern-*T*exas-*H*arvard/MIT Engineering Research Center (VaNTH/ERC) for Bioengineering Educational Technologies is aimed at improving the short- and long-term learning outcomes of bioengineering education at many levels with a particular emphasis on undergraduates. To achieve this goal, the center has enlisted teams composed of faculty in bioengineering, learning sciences, learning technology and assessment and evaluation to develop innovative instructional strategies and to test their effectiveness relative to traditional instruction in bioengineering (See Harris, Bransford & Brophy[1]). These innovations are based on the model of learning and instruction described by Bransford, Brown & Cocking[2] in a volume issued by the National Academy of Sciences entitled *How People Learn: Brain, Mind, Experience, and School* (popularly known as the "HPL model"). Using the HPL model to guide the creation of innovative instructional materials, over 60 modules and course enhancements have been developed within VaNTH covering a variety of bioengineering areas. This paper examines the quality of the evidence underlying VaNTH-sponsored studies and summarizes the quantitative effects of these innovations that have been derived from experimental and quasi-experimental evaluation efforts to date.

A primary motivator behind this paper is to contribute to a relatively new global interest in identifying evidence-based practices (see Cottingham, Maynard & Stagner[3], Slavin[4]). In doing so, we follow an evolving set of guidelines and practices that are being developed by the major organization responsible for conducting similar types of reviews. We have drawn from practices espoused by the Cochrane Collaboration (see http://www.cochrane.org) in medicine, the Campbell Collaboration in social and educational areas (see http://www.campbellcollaboration.org), the Institute of Education Sciences' (IES) What Works Clearinghouse (see http://www.ed.gov/ies/whatworks/), and the Coalition for Evidence-based Practices (see hhttp://www.evidencebasedprograms.org).

For each of these organizations, a necessary first step is a review of the quality of the evidence underlying each study.

In addition to serving as a update of the results reported earlier in Cordray, Pion, A. Harris & Norris[5], this paper adds new information on the quality of the evidence underlying each study. As with the prior analysis, we regard these results as preliminary estimates, pending additional reanalyses guided by advances in statistical practices. Final results will be presented later this year as part of a comprehensive summative appraisal of the impact of the VaNTH ERC project.

II. Meta-analytic Methods

Although the VaNTH/ERC educational innovations share a common pedagogical model (HPL), they represent a broad array of bioengineering topics (e.g., biomechanics, biotransport, optics, ethics). They have been developed by many different faculty at the partner institutions for college and high school students, and they represent single modules delivered as part of a course, collections of modules, and full-scale college courses. They use outcome measures that are designed to gauge the degree to which participants *understand* bioengineering principles and practices, a key objective of the HPL model (Schwartz et al.[6]). As such, it is not possible to use standardized learning outcomes. Moreover, individual studies are, by necessity, limited to specific operationalizations of cause-effect relationships. An important set of inferences involve the generality of the results across types of students, materials, topics or content areas, and time. Further, conducting experimental tests of educational innovations in classroom settings often result in small-sample sizes for intervention and control conditions, making it difficult to detect effects (due to low statistical power).

To summarize what has been learned from this diverse collection of innovations, it was necessary to capture consistent information about study content, procedures, and results. We use a particular method known as meta-analysis (e.g., Cook, Cooper, Cordray, Hartman, Hedges, Louis, & Mosteller[7], Cooper & Hedges[8], Hedges & Olkin[9], Lipsey & Wilson[10]) to quantitatively summarize the results of multiple studies. Although meta-analysis methods are used extensively in many areas (e.g., medicine, education, and job training), their application in engineering and in engineering education is almost non-existent. While novel to engineering, they have been used to assess the cumulative effects of problem-based instruction (Colliver[11], Dochy, Segers, Van den Bossche, & Gijbels[12], Gijbels, Dochy, Van den Bossche & Segers[13]).

Quality of evidence. Fundamental to claims that an educational innovation "works" (i.e., the learning outcome for the group exposed to the innovation is greater, on average, than that of participants exposed to traditional instruction) is the quality of the research design used to derive the relative effect. Consistent with the guidelines for identifying evidence-based educational practices, we first examined the quality of the research design used in each of the VaNTH-sponsored studies. Designs were classified into two broad categories: (1) randomized experiments; and (2) quasi-experiments. It is widely held that results from randomized experiments (where participants have been

assigned to the innovation and counterfactual conditions at random) are more trustworthy than designs that entail non-random assignment (e.g., Boruch[14], Shadish, Cook & Campbell[15]). Quasi-experiments are a class of research designs that resemble an experiment because they entail the manipulation of conditions and systematic measurement of outcomes but do not allocate participants to conditions at random (Campbell & Stanley[16]). Quasi-experiments vary in their ability to control the influence of rival explanations. As such, we distinguished two subclasses of quasi-experiments: those where the innovation and control conditions were based on cohorts of participants in the *same* (or comparable) institution and those quasi-experiments that were based on participants in intact classes from *different* institutions. The former is generally less susceptible to the influence of selection bias than the latter (see Shadish et al[15]).

The effect size. To quantitatively synthesize the results of many diverse modules, a common metric for the results, the effect size or ES, was derived for each study result, where

$ES = (M^T - M^C)/ SD_{pooled.;}$
M^T = *M*ean of the treatment or experimental group;
M^C = *M*ean of the control group;
$SD_{pooled.}$ = Pooled standard deviation of both groups.

The effects size (ES) is the mean difference between innovation control conditions, normalized to the pooled standard deviation for each condition[a] . Cohen[17] asserts that effect sizes of 0.20, 0.50 and 0.80 can be interpreted as representing small, medium and large effects, respectively. These values are used as benchmarks for the interpretation of effect sizes. Statistical tests for the null hypothesis and tests for between study homogeneity are also used to evaluate results. A comprehensive treatment of these issues is provided by Cooper and Hedges[8], Lipsey & Wilson[10], and the extensive references included in both of the classic texts on the topic.

Selecting studies for inclusion in the meta-analysis. Because claims of what works depend on identifying the most trustworthy studies, the focus of this paper is on the extent which the effects differ systematically depending on the research design that was used. Rather than making an *a priori* determination to use only those results that stem from well executed randomized experiments, we treat the decision about inclusion of study results based on different designs as an empirical matter.

III. Description of the Studies

Table 1 presents a brief summary of each of the 28 estimates derived from 19 VaNTH/ERC studies that attempted to estimate the effectiveness of a given module or set of modules[b]. Comprehensive descriptions of the modules and the studies used to derive these effect sizes can be found at the VaNTH website (http://www.vanderbilt.vanth.org).

[a] The calculation of effect sizes depends on the statistical model that was used in each study. The definition offered here is the most generic version. Lipsey and Wilson[10] provide a thorough discussion of the derivation and analysis of effect size estimates.

Table 1: Study Characteristics, Topics and Preliminary Effect Sizes

Study	Topic	Design	No. Modules	Type of Measure	N^C	N^T	Effect Size
1	Virtual Biology Labs	Exp[1]	2	KBQ	14	14	0.351
2	Virtual Biology Labs, Iron Cross, Jumping Jack	Exp[1]	3	KBQ	16	16	0.381
3	Metabolism	Exp[1]	1	KBQ	18	19	0.600
4	Ethics and Adaptive Expertise	Exp[1]	1	Adapt. Expertise	15	15	0.825
5	Ultrasound	Exp[1]	1	KBQ	22	21	-0.366
6	Jumping Jack	Exp[1]	1	Adapt. Expertise	10	11	0.766
7	Matlab-based homework	Exp[1]	1	KBQ	20	20	0.760
8	Calorimetry	Exp[1]	2	KBQ	30	47	0.435
9	Bioreactor	Q-Exp[2]	2 sections	KBQ	11	11	0.650
10	Microbial Kinetics	Q-Exp[2]	1	KBQ	11	11	2.00*
11	Spectral Analysis	Q-Exp[2]	1	KBQ	24	22	0.693
12	Capillary Filtration	Q-Exp[2]	1	KBQ	39	46	0.377
13	Homework-less course	Q-Exp[2]	2	KBQ	69	39	0.361
14	Biomechanics course 1	Q-Exp[2]	Full Course	KBQ	~50	~50	0.680
	Biomechanics course 2	Q-Exp[2]	Full Course	KBQ	~50	~50	0.180
	Biomechanics course 3	Q-Exp[2]	Full Course	KBQ	~50	~50	0.490
	Biomechanics course 4	Q-Exp[2]	Full Course	KBQ	~50	~50	0.230
15	Metabolic flux	Q-Exp[2]	1	KBQ	10	10	0.780
16	Biotransport	Q-Exp[2]	Full Course	Adapt. Expertise	52	54	1.440
17	Port Wine Stain	Q-Exp[3]	3 sections	KBQ	57	57	0.830
18	Balance beam 1	Q-Exp[4]	1	Facts +KBQ	44	151	0.416
	ECG	Q-Exp[4]	1	Facts +KBQ	62	85	0.764
	Iron cross, imaging, swim	Q-Exp[4]	3	Facts +KBQ	99	12	2.00*
19	ECG-Physics	Q-Exp[4]	1	Facts+App+KBQ	47	37	1.060
	Balance beam 2	Q-Exp[4]	1	Facts+App+KBQ	15	19	0.853
	ECG-Biology	Q-Exp[4]	1	Facts+App+KBQ	57	43	0.550
	Iron cross 2	Q-Exp[4]	1	Facts+App+KBQ	34	69	0.873
	Image, Swim, Optics, Hemodynamics	Q-Exp[4]	4	Facts+App+KBQ	14	6	2.00*

Notes: [1] true experimental design, with randomization to conditions;
[2] Quasi-experimental design based on nonrandom enrollment in spring or fall courses (either intervention or control condition);
[3] Quasi-experimental design with students as their own control in relevant and irrelevant conditions/outcomes;
[4] Quasi-experimental design with students in intact classes, different schools.

Eight of the estimates were derived from experimental studies; 11 estimates were obtained from studies employing the same/comparable institution cohort type of quasi-experiment; and 9 of the estimates were derived from studies using intact classes as control conditions or another form of control (e.g., Study 17 using multiple dependent variables).

[b] Multiple estimates were derived from studies if the assessments involved independent samples of participants, otherwise an average effect size was derived across multiple outcomes or where a common control condition was used for two or more ESs.

In addition to the type of experimental or quasi-experimental research design that was used, these descriptions show that the modules/courses focused on a variety of bioengineering topics and that some topics were tested (replicated) multiple times. The studies differed in size, ranging from a total of 20 participants to nearly 200 participants. The statistical methods underlying the meta-analysis methods used here weight the effect size estimates by the precision of the individual estimate (sample size), so larger studies affect the aggregate estimate (the average effect size) more than smaller studies, all else considered.

As noted earlier, a common or standardized outcome could not be used to assess educational outcomes across modules or courses. Rather, deep understanding of basic bioengineering knowledge was assessed through a combination of problem solving and performance transfer tasks. As shown in Table 1, these outcomes were scored with rubrics designed to assess the depth of knowledge and understanding through knowledge-based questions (KBQ), weighted scores producing an index of adaptive expertise (variously defined), or as weighted composite effects (Facts+ Application+ KBQ). Effects are reported for each study or sub-study. In several cases (denoted with an "*") the effect sizes exceeded conventional bounds. These effect estimates were winsorized (truncated) at a maximum value of ES=2.00.

VI. Meta-analytic Results

Table 1 presents evidence on the effectiveness of dozens of modules (from 19 experimental or quasi-experimental studies). As shown in Table 2, for this set of VaNTH-sponsored modules, the weighted average effect size is 0.644 standard deviation units (95% CI = 0.496 →0.792). That is, on average, the participants exposed to VaNTH-sponsored modules or courses out-performed their counterparts (exposed to the same material but through traditional pedagogy) by about two-thirds of a standard deviation. Not surprising, the test of statistical significance ($z = ES_{\text{wt average}}/SE = 13.46$) for this weighted estimate is highly significant; the weighted average is statistically different than 0.

Moreover, by Cohen's norms[17], this weighted average ES represents an effect that is solidly between a medium and large effect. Empirically, prior meta-analyses show that innovations produce effects that are on average in the small range (i.e., 0.25).

Table 2. Overall Effects and Effects by Type of Design

Study type	Studies	Estimates	Weighted Average	95 % Confidence Intervals ES -.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.1 1.2
Experiment	8	8	0.416	------x------
Cohort-based Quasi-experiment	8	11	0.562	----x----
Intact Class-based Quasi-experiment	3	9	0.803	----x----
Overall	19	28	0.644	---x---

The effect of research design quality on the magnitude of the ES estimates. Additional analyses show that studies using higher quality designs (experiments) produce effects that are smaller (weighted average ES = 0.417, 95% CI = 0.188 →0.646) than studies using the cohort-based quasi-experimental design (weighted average ES = 0.562, 95% CI = 0.419 →0.704). Studies using the intact classes to construct comparisons produced larger estimates (weighted average ES = 0.803, 95% CI = 0.655 → 0.951). All of the weighted averages are statistically different than zero at p <.001.

V. Discussion and Future Directions

Although this disparity in effects due to design quality is common in meta-analytic studies (i.e., poorer designs produce biased effects), more careful examination of the studies suggests that design quality and treatment fidelity may be confounded. Some of the strongest interventions were studied with designs that appear, on the surface, to be weaker than the true experiments. Preliminary examination and reanalyses of some of the studies and sub-studies (e.g., the ECG –Physics sub-study from Study 19 in Table 2, one of the weaker quasi-experiments) shows that the groups are remarkably similar on pre-test measures of knowledge and in gender/racial compositions. Here, the innovation was tested on 37 studies from four classes and the comparison group also contained 47 students from four separate classes. As such, the overall statistical results are not accurate because students were nested within classes. This is a common problem in educational research that is only now being addressed with more advanced statistical methods (e.g., Hierarchical Linear Modeling). Before the final summative meta-analysis is undertaken, these studies will be reanalyzed.

Returning to the design/fidelity confounding, some of the small effects produced by strong experimental designs may involve substantial infidelity in the implementation of the HPL model. For example, Studies 5 and 7 did not entail all the HPL elements. Instead, these studies investigated the influence of various forms of technology enhancements as their primary source of innovation. Other studies did not actively invoke all aspects of the Legacy Cycle model that was used in 15 of the 19 studies. Simultaneous consideration of the strength of the research design and fidelity of the intervention should provide more useful estimates of the underlying effects of this set of modules.

Acknowledgements:

This work was supported primarily by the Engineering Research centers Program of the National Science Foundation under Award Number EEC-9876363. Jennifer Gilbert's efforts were also supported by the Institute for Education Sciences (IES) as part of Vanderbilt's predoctoral training grant titled the Experimental Education Research Training (ExpERT) program (Award Number IES R305B04110, Professor David S. Cordray, Director).

References

[1] Harris, T.R., Bransford, J.D. & Brophy, S.P. (2002). Roles for learning sciences and learning technologies in biomedical engineering education: A review of recent advances. *Annual Review of Biomedical Engineering, 4,* 29-48.

[2] Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

[3] Cottingham, P., Maynard, R., & Stagner, M. (2005). Generating and using evidence to guide public policy and practice: Lessons form the Campbell test-bed project. *Journal of Experimental Criminology*, 1, 279-294.

[4] Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31 (7), 15-21.

[5] Cordray, D.S., Pion, G.M., Harris, A. & Norris, P. (2003). The value of the VaNTH Engineering Center. *IEEE Engineering in Medicine and Biology Magazine,* **22**: 47-54.

[6] Schwartz, D.L. Lin, X., Brophy, S., & Bransford, J.D. (1999). Toward the development of flexibly adaptive instructional designs. In C,M, Reigelut (Ed.), *Instructional design theories and models: Volume 11.* Hillsdale, NJ: Erlbaum.

[7] Cook, T.D., Cooper, H. Cordray, D.S., Hartmann, H., Hedges, L.V., Light,. R.J., Louis, T.A., & Mosteller, F. (1992) *Meta-analysis for explanation: A casebook.* New York, NY: Russell Sage Foundation.

[8] Cooper, H. & Hedges, L.V. (Eds.) (1993). *Handbook of Research Synthesis.* New York, NY: Russell Sage Foundation.

[9] Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

[10] Lipsey, M. W. & Wilson, D. B.(2001). *Meta-analysis: A practical guide. Thousand Oaks, CA: Sage.*

[11] Colliver, J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine*, 75(3), 259-266.

[12] Dochy, F., Seger,M., Van den Bossche, P., & Gijbels, D. (2003). *Learning and Instruction,* 13, 533-568.

[13] Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). *Review of Educational Research*, 75(1), 27-61.

[14] Boruch, R.F. (1997). *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage.

[15] Shadish, W. Cook, T.D, & Campbell, D.T. (2002) *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Co.

[16] Campbell, D.T. & Stanley, J. C. (1963) *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

[17] Cohen, J. (1988) *Statistical power analysis for the behavioral sciences (2nd Ed.).* Hillsdale NJ: Lawrence Erlbaum Associates.