

# QUANTITATIVE RESEARCH ON TEACHING METHODS IN TERTIARY EDUCATION

William E. Becker

For presentation as a keynote address at the Midwest Conference on Student Learning in Economics, Innovation, Assessment and Classroom Research, University of Akron, Akron, Ohio, November 7, 2003, and forthcoming as Chapter 11 in W. E. Becker and M. L. Andrews (Eds.), *The Scholarship of Teaching and Learning in Higher Education: Contributions of Research Universities*, Indiana University Press, 2004.

Advocates and promoters of specific education methods are heard to say “the research shows that different teaching pedagogy really matters.” Education specialist Ramsden (1998, p. 355) asserts: “The picture of what encourages students to learn effectively at university is now almost complete.” Anecdotal evidence and arguments based on theory are often provided to support such claims, but quantitative studies of the effects of one teaching method versus another are either not cited or are few in number. DeNeve and Heppner (1997), for example, found only 12 of the 175 studies identified in a 1992 through 1995 search for “active learning” in the Educational Resources Information Center (ERIC) data base made comparisons of active learning techniques with another teaching method. An ERIC search for “Classroom Assessment Techniques” (CATs) undertaken for me by Jillian Kinzicat in 2000 yielded a similar outcome. My own (March 2000) request to CATs specialist Tom Angelo for direction to quantitative studies supporting the effectiveness of CATs yielded some good leads, but in the end there were few quantitative studies employing inferential statistics.

Even when references to quantitative studies are provided, they typically appear with no critique. When advocates point to individual studies or to meta-analyses summarizing quantitative studies, they give little or no attention to the quality or comparability of studies encompassed. When critics, on the other hand, point to a block of literature showing “no significant difference,” the meaning of statistical significance is overlooked.<sup>1</sup>

In this study I address the quality of research aimed at assessing alternative teaching methods. I advance the scholarship of teaching and learning by separating empirical results with statistical inference from conjecture about the student outcomes associated with CATs and other teaching strategies aimed at actively engaging students in the learning process. I provide specific criteria for both conducting and

exploring the strength of discipline-specific quantitative research into the teaching and learning process. Examples employing statistical inference in the teaching and learning process are used to identify teaching strategies that appear to increase student learning. I focus only on those studies that were identified in the literature searches mentioned above or were called to my attention by education researchers as noteworthy.

Although my review of the education assessment literature is not comprehensive and none of the studies I reviewed were perfect when viewed through the lens of theoretical statistics, there is inferential evidence supporting the hypothesis that periodic use of things like the one-minute paper (wherein an instructor stops class and asks each student to write down what he or she thought was the key point and what still needed clarification at the end of a class period) increases student learning. Similar support could not be found for claims that group activities increase learning or that other time intensive methods are effective or efficient. This does not say, however, that these alternative teaching techniques do not matter. It simply says that there is not yet compelling statistical evidence saying that they do.

### **CRITERIA**

A casual review of discipline-specific journals as well as general higher education journals is sufficient for a reader to appreciate the magnitude of literature that provides prescriptions for engaging students in the educational process. Classroom assessment techniques, as popularized by Angelo and Cross (1993), as well as active-learning strategies that build on the seven principles of Chickering and Gamson (1987) are advanced as worthwhile alternatives to chalk and talk. The relative dearth of quantitative work aimed at measuring changes in student outcomes associated with one teaching method versus another is surprising given the rhetoric surrounding CATs and the numerous methods that fit under the banner of active and group learning.

A review of the readily available published studies involving statistical inference shows that the intent and methods of inquiry, analysis, and evaluation vary greatly from discipline to discipline. Thus, any attempt to impose a fixed and unique paradigm for aggregating the empirical work on education practices across disciplines is destined to fail.<sup>2</sup> Use of flexible criteria holds some promise for critiquing empirical work involving statistical inferences across diverse studies. For my work

here, I employ an 11-point set of criteria that all inferential studies can be expected to address in varying degrees of detail:

- 1) Statement of topic, with clear hypotheses;
- 2) Literature review, which establishes the need for and context of the study;
- 3) Attention to unit of analysis (e.g., individual student versus classroom versus department, etc.), with clear definition of variables and valid measurement;
- 4) Third-party supplied versus self-reported data;
- 5) Outcomes and behavioral change measures;
- 6) Multivariate analyses, which include diverse controls for things other than exposure to the treatment that may influence outcomes (e.g., instructor differences, student aptitude), but that cannot be dismissed by randomization (which typically is not possible in education settings);
- 7) Truly independent explanatory variables (i.e., recognition of endogeneity problems including simultaneous determination of variables within a system, errors in measuring explanatory variables, etc.);
- 8) Attention to nonrandomness, including sample selection issues and missing data problems;
- 9) Appropriate statistical methods of estimation, testing, and interpretation;
- 10) Robustness of results -- check on the sensitivity of results to alternative data sets (replication), alternative model specifications and different methods of estimation and testing; and
- 11) Nature and strength of claims and conclusions.

These criteria will be discussed in the context of selected studies from the education assessment literature.

### **TOPICS AND HYPOTHESES**

The topic of inquiry and associated hypotheses typically are well specified. For example, Hake (1998) in a large scale study involving data from some 62 different physics courses seeks an answer to the single question: "Can the classroom use of IE (interactive engagement of students in activities that yield immediate feedback) methods increase the effectiveness of introductory mechanics courses well beyond that attained by traditional methods?" (p. 65) The Hake study is somewhat unique in its attempt to measure the learning effect of one set of teaching strategies versus another across a broad set of institutions.<sup>3</sup>

In contrast to Hake's multi-institution study are the typical single-institution and single-course studies as found, for example, in Harwood (1999). Harwood is interested in assessing student response to the introduction of a new feedback form in an accounting course at one institution. Her new feedback form is a variation on the widely used one-minute paper (a CAT) in which an instructor stops class and asks each student to write down what he or she thought was the key point and what still needed clarification at the end of a class period. The instructor collects the students' papers, tabulates the responses (without grading), and discusses the results in the next class meeting. (Wilson 1986, p. 199). Harwood (1999) puts forward two explicit hypotheses related to student classroom participation and use of her feedback form:

H<sub>1</sub>: Feedback forms have no effect on student participation in class.

H<sub>2</sub>: Feedback forms and oral in-class participation are equally effective means of eliciting student questions. (p. 57)

Unfortunately, Harwood's final hypothesis involves a compound event (effective and important), which is difficult to interpretation:

H<sub>3</sub>: The effect of feedback forms on student participation and the relative importance of feedback forms as compared to oral in-class participation decline when feedback forms are used all of the time. (p. 58)

Harwood does not address the relationship between class participation and learning in accounting, but Almer, Jones, and Moeckel (1998) do. They provide five hypotheses related to student exam performance and use of the one-minute paper:

H<sub>1</sub>: Students who write one-minute papers will perform better on a subsequent quiz than students who do not write one-minute papers.

H<sub>1a</sub>: Students who write one-minute papers will perform better on a subsequent essay quiz than students who do not write one-minute papers.

H<sub>1b</sub>: Students who write one-minute papers will perform better on a subsequent multiple-choice quiz than students who do not write one-minute papers.

H<sub>2</sub>: Students who address their one-minute papers to a novice audience will perform better on a subsequent quiz than students who address their papers to the instructor.

H<sub>3</sub>: Students whose one-minute papers are graded will perform better on a subsequent quiz than students whose one-minute papers are not graded. (p. 493)

Rather than student performance on tests, course grades are often used as an outcome measure and explicitly identified in the hypothesis to be tested. For example, Trautwein, Racke, and Hillman (1996/1997, p. 186) ask: “Is there a significant difference in lab grades of students in cooperative learning settings versus the traditional, individual approach?” The null hypothesis and alternative hypotheses here are “no difference in grades” versus “a difference in grades.” There is no direction in the alternative hypothesis so, at least conceptually, student learning could be negative and still be consistent with the alternative hypothesis. That is, this two-tail test is not as powerful as a one-tail test in which the alternative is “cooperative learning led to higher grades,” which is what Trautwein, Racke and Hillman actually conclude. (More will be said about the use of grades as an outcome measure later.)

Not all empirical work involves clear questions and unique hypotheses for testing. For example, Fabry et al. (1997, p. 9) state “The main purpose of this study was to determine whether our students thought CATs contributed to their level of learning and involvement in the course.” Learning and involvement are not two distinct items of analysis in this statement of purpose. One can surely be involved and not learn. Furthermore, what does the “level of learning” mean? If knowledge (or a set of skills, or other attributes of interest) is what one possesses at a point in time (as in a snap-shot, single-frame picture), then learning is the change in knowledge from one time period to another (as in moving from one frame to another in a motion picture). The language employed by authors is not always clear on the distinction between knowledge (which is a stock) and learning (which is a flow) as the foregoing examples illustrate.

## **LITERATURE REVIEW**

By and large, authors of empirical studies do a good job summarizing the literature and establishing the need for their work. In some cases, much of an article is devoted to reviewing and extending the theoretical work of the education specialists. For instance, Chen and Hoshower (1998) devoted approximately one-third of their 13 pages of text to discussing the work of educationalists. Harwood (1999), before or in conjunction with the publication of her empirical work, co-

authored descriptive pieces with Cottel (1998) and Cohen (1999) that shared their views and the theories of others about the merits of CAT. In stark contrast, Chizmar and Ostrosky (1999) wasted no words in stating that as of the time of their study no prior empirical studies addressed the learning effectiveness (as measured by test scores) of the one-minute paper (see their endnote 3); thus, they established the need for their study.<sup>4</sup>

### **VALID AND RELIABLE UNITS OF ANALYSIS**

The 1960s and 1970s saw debate over the appropriate unit of measurement for assessing the validity of student evaluations of teaching (as reflected, for example, in the relationship between student evaluations of teaching and student outcomes). In the case of end-of-term student evaluations of instructors, an administrator's interest may not be how students as individuals rate the instructor but how the class as a whole rates the instructor. Thus, the unit of measure is an aggregate for the class. There is no unique aggregate, although the class mean or median response is typically used.<sup>5</sup>

For the assessment of CATs and other instructional methods, however, the unit of measurement may arguably be the individual student in a class and not the class as a unit. Is the question: how is the  $i^{th}$  student's learning affected by being in a classroom where one versus another teaching method is employed? Or is the question: how is the class's learning affected by one method versus another? The question (and answer) has implications for the statistics employed.

Hake (1998) reports that he has test scores for 6,542 individual students in 62 introductory physics courses. He works only with mean scores for the classes; thus, his effective sample size is 62, and not 6,542. The 6,542 students are not irrelevant, but they enter in a way that I did not find mentioned by Hake. The amount of variability around a mean test score for a class of 20 students versus a mean for 200 students cannot be expected to be the same. Estimation of a standard error for a sample of 62, where each of the 62 means receives an equal weight, ignores this heterogeneity.<sup>6</sup> Francisco, Trautman, and Nicoll (1998) recognize that the number of subjects in each group implies heterogeneity in their analysis of average gain scores in an introductory chemistry course. Similarly, Kennedy and Siegfried (1997) make an adjustment for heterogeneity in their study of class size on student learning in economics.

Fleisher, Hashimoto, and Weinberg (2002) consider the effectiveness (in terms of course grade and persistence) of 47 foreign graduate student instructors versus 21 native English speaking graduate student instructors in an environment in which English is the language of the majority of their undergraduate students. Fleisher, Hashimoto, and Weinberg recognize the loss of information in using the 92 mean class grades for these 68 graduate student instructors, although they do report aggregate mean class grade effects with the corrected heterogeneity adjustment for standard errors based on class size. They prefer to look at 2,680 individual undergraduate results conditional on which one of the 68 graduate student instructors each of the undergraduates had in any one of 92 sections of the course. To ensure that their standard errors did not overstate the precision of their estimates when using the individual student data, Fleisher, Hashimoto, and Weinberg explicitly adjusted their standard errors for the clustering of the individual student observations into classes using a procedure developed by Moulton (1986).<sup>7</sup>

Whatever the unit of measure for the dependent variable (aggregate or individual) the important point here is recognition of the need for one of two adjustments that must be made to get the correct standard errors. If an aggregate unit is employed (e.g., class means) then an adjustment for the number of observations making up the aggregate is required. If individual observations share a common component (e.g., students grouped into classes) then the standard errors reflect this clustering. Computer programs like STATA and LIMDEP can automatically perform both of these adjustments.

No matter how appealing the questions posed by the study are, answering the questions depends on the researcher's ability to articulate the dependent and independent variables involved and to define them in a measurable way. The care with which researchers introduce their variables is mixed, but in one way or another they must address the measurement issue: What is the stochastic event that gives rise to the numerical values of interest (the random process)? Does the instrument measure what it reports to measure (validity)? Are the responses consistent within the instrument, across examinees, and/or over time (reliability)?<sup>8</sup>

Standardized aptitude or achievement test scores may be the most studied measure of academic performance. I suspect that there are nationally normed testing instruments at the introductory college levels in every major discipline – at a minimum, the Advanced Placement exams of ETS. There are volumes written on the

validity and reliability of the SAT, ACT, GRE, and the like. Later in this chapter I comment on the appropriate use of standardized test scores, assuming that those who construct a discipline-specific, nationally normed exam at least strive for face validity (a group of experts say the exam questions and answers are correct) and internal reliability (each question tends to rank students as does the overall test).

Historically, standardized tests tend to be multiple-choice, although national tests like the advanced placement (AP) exams now also have essay components. Wright et al. (1997) report the use of a unique test score measure: 25 volunteer faculty members from external departments conducted independent oral examinations of students. As with the grading of written essay-exam answers, maintaining reliability across examiners is a problem that requires elaborate protocols for scoring. Wright et al. (1997) employed adequate controls for reliability but because the exams were oral, and the difference between the student skills emphasized in the lecture approach and in the co-operative learning approach was so severe, it is difficult to imagine that the faculty member examiners could not tell whether each student being examined was from the control or experimental group; thus, the possibility of contamination cannot be dismissed.

Whether multiple-choice (fixed response) questions, essay (constructed response) questions or oral exams measure different dimensions of knowledge is a topic that is and will continue to be hotly debated. Becker and Johnston (1999) address the simultaneity between alternative forms of testing and the lack of information that can be derived from the simple observation that essay and multiple-choice test scores are correlated. As this debate continues researchers have no choice but to use the available content tests or consider alternatives of yet more subjective forms. Self-created instruments must be treated as suspect. Fabry et al. (1997), for example, focus their analysis on student answers to the question: "Do you think CATs enhanced your learning/participation in the course?" (p. 9). Written responses were converted to a three-point scale (yes, maybe/somewhat, no). Although Fabry et al. report only the number of yes responses, converting these responses to a numerical value on a number line is meaningless. Any three ordered values could be used, but the distance between them on the number line is irrelevant. To explain unordered and discrete responses, researchers should consider the estimation of multinomial logit or probit models discussed in Greene (2003).

More troubling with respect to the study by Fabry et al. are the facts that they never define the word “enhance,” and they do not specify whether learning and participation are to be treated as synonyms, substitutes, or as an and/or statement. In addition, there is a framing problem. The scale is loaded away from the negative side. By random draw there is only a one-third chance of getting a response of “no enhanced learning/ participation.” These are classic problems found in invalid instruments; that is, a one-to-one mapping does not exist between the survey question and the responses.

Fabry et al. (1997) also aggregated student responses over four instructors who each used a different combination of CATs in the four different courses each taught. How can we distinguish if it is the effect of the instructors or the set of techniques that is being captured? This is a clear problem of aggregation that cannot be disentangled to get a valid answer about what is being measured.

Student evaluations of teaching are often used to answer questions of effectiveness, which raises another issue of validity: Do student evaluations measure teaching effectiveness? Becker (2000) argues that there is little reason to believe that student evaluations of teaching capture all the elements or the most important elements of good teaching. As measured by correlation coefficients in the neighborhood of 0.7, and often less, end-of-term student evaluation scores explain less than 50 percent of the variability in other teaching outcomes, such as test scores, scores from trained classroom observers, post-course alumni surveys, and so on.<sup>9</sup>

Other questions of validity arise when course grades are used as the measure of knowledge. Are individual exam grades just a measure of student knowledge at the time of the assessment? Do course grades reflect student end-of-term knowledge, learning (from the beginning to the end of the term), rate of improvement, or something even more subjective? Finally, grades may not be reliably assigned across instructors or over time. To establish validity of grades among a group of instructors elaborate protocols would have to be in place. Each student would have to be graded by more than one evaluator with the resulting distribution of students and their grades being roughly the same across evaluators.

Indices of performance are sometimes created to serve as explanatory variables as well as the outcome measure to be explained. Indices used to represent an aggregate can be difficult if not impossible to interpret. Kuh, Pace, and Vesper (1997), for example, create a single index they call “active learning,” from 25 items

related to student work and student personal development interests. They then use this active learning index score as one of several indexes included as independent variables in a least-squares regression aimed at explaining an index of what students' perceive they gained from attending college. Their estimated slope coefficient for females tells us that a one unit increase in the active learning index increases the predicted student's gain index by 0.30, holding all else fixed.<sup>10</sup> But what exactly does this tell us?

Kuh, Pace, and Vesper (1997) provide multivariate analysis of students' perceived gains, including covariate indices for student background variables, and institutional variables as well as the measures for good educational practices such as active learning. Their study is in the tradition of input-output or production function analysis advanced by economists in the 1960s. Ramsden (1998, pp. 352-354), on the other hand, relies on bivariate comparisons to paint his picture of what encourages university students to learn effectively. He provides a scatter plot showing a positive relationship between a y-axis index for his "deep approach" (aimed at student understanding versus "surface learning") and an x-axis index of "good teaching" (including feedback of assessed work, clear goals, etc.).

Ramsden's regression ( $y = 18.960 + 0.35307x$ ) implies that a zero on the good teaching index predicts 18.960 index units of the deep approach. A decrease (increase) in the good teaching index by one unit leads to a 0.35307 decrease (increase) in the predicted deep approach index. The predicted deep approach index does not become negative (surface approach?) until the good teaching index is well into the negative numbers (bad teaching?) at a value of  $-53.605$ .<sup>11</sup> What is the possible relevance of these numbers for the instructor? What specific information does this yield that could guide institutional policy?

---

INSERT Figure 1: Deep approach and good teaching

---

### **SELF-REPORTED DATA**

Frequently in classroom assessment work, data on students are obtained from the students themselves even though students err greatly in the data they self-report

(Maxwell and Lopus 1994) or fail to report information as requested (Becker and Powers 2001).

Kuh, Pace, and Vesper (1997) recognize the controversial nature of using self-reported data but in essence argue that it is not unusual and that in the absence of other options it has become accepted practice. When the problems of self-reported data are considered, it is the validity and reliability of the dependent variable (self-reported achievement, gain, satisfaction, etc.) that is typically addressed. Overlooked is the bias in coefficient estimators caused by measurement errors in the explanatory variables. An ordinary least-squares estimator of a slope coefficient in a regression of  $y$  on  $x$  is unbiased if the  $x$  is truly independent (which requires that causality runs from  $x$  to  $y$ , and not simultaneously from  $y$  to  $x$ , and/or that  $x$  is measured with no error). As long as the expected value of  $y$  at each value of  $x$  is equal to the true mean of  $y$  conditioned on  $x$ , measurement error in  $y$  is not a problem in regression analysis. It is the measurement error in  $x$  that leads to the classic regression to the mean phenomena that plagues education-achievement-equation estimates (as demonstrated mathematically in several endnotes to this chapter).

Institutional procedures regarding human subjects research and registrars' misconceived policies blocking access to student data in the name of the Buckley Amendment may be the biggest obstacles to quality evaluation of CATs and other active learning methods. (The Family Education Rights and Privacy Act explicitly enable teachers to have access to student information for the purpose of improving instruction, <http://www.ed.gov/offices/OM/fpco/ferpa/ferparegs.html#9931>.) Many authors report that they were prevented from getting actual individual student data from university records but were free to seek self-reported data from students. For example, Almer, Jones and Moeckel (1998) report that institutional policy precluded them from directly obtaining SAT and GPA information from student records (p. 491). They obtained 539 self-reported GPAs and 295 SAT scores from the 867 students in their final sample. Both measures of ability were found to be highly significant in explaining quiz scores. They report, however, that inclusion of either ability measure did not change the interpretation of results: the use of one-minute papers raises student performance. Because of the potential for bias resulting from missing data, both ability measures were excluded from their reported analyses. Becker and Powers (2001), on the other hand, find that the inclusion or exclusion of

potentially biased and missing data on ability measures and other standard covariates were critical to accuracy in assessing the importance of class size in learning.

There is no consistency among institutions regarding instructor's access to student data for classroom assessment studies. For example, Maxwell and Lopus (1994) were able to get access to actual individual student data as supplied by their institution's registrar in their study to show the misleading nature of student self-reported data. Chizmar and Ostrosky (1999) also obtained registrar data for their study of the one-minute paper. The complaints of educators who are unable to conduct investigations aimed at improving teaching are well documented. This is especially the case at research universities where administrators' fears of reprisals (such as the withholding of federal funds to the medical school) drive institutional policies in any and all studies involving human subjects. Legal experts are debating the extension of medical protocols in human subject research to the scholarship of teaching and learning (Gunsalus 2002). The expansion and rigid application by institutional review boards of policies intended for medical clinical trials involving human subjects to studies aimed at teaching and learning, and the erroneous application of the Family Educational Right to Privacy Act (Buckley Amendment) barring the release of student information for studies that are explicitly aimed at improving instruction are two hurdles that are causing some scholars to abandon efforts to undertake empirical study of the educational process.

### **OUTCOMES AND STUDY DESIGN**

As already discussed, numerous measures have been proposed and used to assess knowledge (cognitive domain) and attitudes (affective domain) relevant to student development. If truly randomized experiments could be designed and conducted in education, as advanced by Campbell, Stanley and Gage (1963) for example, then a researcher interested in assessing cognitive-domain or affective-domain outcomes of a given classroom treatment need only administer an achievement test or attitude survey at the end of the program to those randomly assigned to the treatment and control groups. There would be no need for pre-treatment measures. Consequences other than the treatment effect could be dismissed with reference to the law of large numbers (i.e., the distribution of a statistic to be calculated from a random sample degenerates or collapses on its expected value as the sample size increases).

Unfortunately, no one has ever designed an absolutely perfect experiment -- randomization is not an absolute; it is achieved in degrees. The best we can do in social science research is to select the treatment and control groups so that they are sharply distinct and yet represent events that could happen to anyone (Rosenbaum 1999). The difficulty is finding a believable counterfactual: If the  $i^{th}$  person is in the control (experimental) group, what would have happened if someone like this person had been in the experimental (control) group?<sup>12</sup>

Instead of talking about final achievement, researchers attempt to adjust for lack of randomness in starting positions by addressing the learning effect of one treatment versus another. The most obvious measure of the learning outcome is the difference between a pretest (test given to students at the start of a program or course of study – i.e., pre-treatment) and posttest (test given at the end – i.e., post-treatment). Similarly, for changes in attitudes an instrument to assess the difference between a “presurvey” and “postsurvey” can be constructed. The study design for assessing the treatment effect is then called “a difference in differences design” in which the difference between the pretest and posttest differences for the treatment and control groups is compared. Unfortunately, calculation and comparison of these “change scores” or “value-added measures” are fraught with problems that go beyond the psychometric issues of the validity and reliability of the instruments (as already discussed).

Researchers interested in examining the effect of a treatment occasionally use the final grade in a course. As already stated, it is never clear whether the course grade is intended to measure the student’s final position (post-knowledge) or improvement (post-knowledge minus pre-knowledge). Whether it is assigned on a relative basis (one student’s performance versus another’s) or absolute scale has implications as to what can be assessed by a comparison of grades. For many statistical procedures, a normal distribution is assumed to be generating the outcome measures. When grades are used as the outcome measure, the researcher must be concerned about the ceiling (typically 4.00) and the discrete nature of grades (as in A, B, C). The normal distribution is continuous with an infinite number of values and thus it is not appropriate when letter grades are the outcome measure.

Anaya (1999, p. 505) proposes the use of a “residual gain score” for assessing the entire undergraduate experience. She regresses end-of-undergraduate experience GRE scores on pre-undergraduate SAT scores, obtains residuals from this regression,

which she calls the “residual gain score,” and then regresses these residuals on explanatory variables of interest. Conceptually, one can easily adopt this process to individual courses using GPA or other aggregates. For example, in stage one, end-of-term numerical course grades can be regressed as a postscore on matched students’ GPAs at the start of the course as a prescore. In stage two, the residuals from this regression are regressed on a zero- or one-valued covariate (for control or experimental identification) and other explanatory variables. Unfortunately, this “residual gain score” model is inconsistent. If the residual gain score is a function of known explanatory variables, why isn’t the posttest a function of these same variables?<sup>13</sup>

Numerous transformations of test scores have been proposed as the student outcome of the teaching process. Becker (1982), for example, put forward a theoretical model of student achievement associated with the optimum allocation of time in which the appropriate outcome is a logarithmic transformation of student achievement. This model was estimated by Gleason and Walstad (1988). A log transformation has the greatest effect on extreme high values. Often, when the posttest or the posttest-minus-pretest change score is used as the dependent variable, extreme high values are not a problem because the maximum score on a test is achieved (the ceiling effect). This truncation causes a special problem in modeling achievement and learning because an achievable ceiling implies that those with high pretest scores cannot become measurably better. It also implies that test scores cannot be assumed to be normally distributed as required for most testing situations.

One modern way to handle ceiling effects is to estimate a Tobit model (named after Nobel laureate in economics James Tobin), which involves an estimate of each student achieving the ceiling and then simultaneously adjusting the regression in accordance. Before maximum likelihood computer programs were readily available to estimate Tobit models, a few psychologists and economists advanced a gap-closing measure as the dependent variable for studies of educational methods where ceiling effects might be present:

$$g = \text{gap closing} = \frac{\text{posttest score} - \text{pretest score}}{\text{maximum score} - \text{pretest score}}$$

The three test scores (*maximum score*, *posttest score*, and *pretest score*) could be defined for the individual student or as average measures for a group of students. The “average *g*” assigned to a group could be obtained from averaging the *g* calculated for

each student in the group or it could be obtained from the test score averages for the group. The resulting average  $g$  from these two methods need not be the same; that is, results may be sensitive to the average method employed.

Hake (1998), measured the pretest and posttest scores by the respective classroom averages on a standardized physics test. Unfortunately, in 1998 Hake was unaware of the literature on the gap-closing model. The outcome measure  $g$  is algebraically related to the starting position of the student as reflected in the pretest:  $g$  falls as the *pretest score* rises, for  $maximum\ score \geq posttest\ score \geq pretest\ score$ .<sup>14</sup> Any attempt to regress a posttest-minus-pretest change score, or its standardized gap-closing measure  $g$  on a pretest score yields a biased estimate of the pretest effect.<sup>15</sup>

Almost universally now researchers in education attempt to measure a treatment effect by some variant of student behavioral change over the life of the treatment. They seldom address what the value of that change score is to the student and society. Students may place little value on performing well on an exam that does not count. The market for undergraduates does not place a value on change; it values the final level of accomplishment. Employers buy graduates' contemporaneous aptitudes and skills, not the change in test scores or change in opinions. What the student knew four years ago in the first semester of the freshman year or what he or she may have learned in any given course is irrelevant to the employer, except insofar as it affects the rate of learning. Knowing the level of a test score or the difference between test scores is of little career help to a student or society without knowing the value the market places on these measures.

Knowledge of test scores may have administrative value to the classroom teacher, but that may have little relationship to the economic concept of value. Just as water has a high "value in use" but a low "value in exchange," some basic skills, such as an ability to reconcile a checkbook, may have high value in use but low value in exchange. Other skills may have a high value at one point in time and little value at another; for example, the ability to manipulate a slide rule fell in value with the availability of the inexpensive hand calculator; the ability to manipulate the hand held calculator fell in value with the advance of spreadsheets, MathCAD, and statistical computer packages. Although some skills may be viewed as essential for education, their market value is determined by demand and supply. The normative beliefs of a faculty member, department chair, curriculum committee, central administrator or university board of governance member about the importance of intellectual skills are

elusive without reference to what employers are paying for the bundle of skills embodied in graduates, and what skills they desire from the graduates. (The satisfaction derived from learning and its change score measurement modeled in Becker [1982] is ignored here in the interest of brevity).

Hansen, Kelley, and Weisbrod (1970) called attention to the problem of valuing multi-dimensional student learning and its implications for curriculum reform but few have followed their lead. As they state, who receives the benefits of instruction and how they weight those benefits will affect the valuation. In assessing benefits, researchers can explore the effect of instruction in a specific subject on the decisions of unequally endowed students to go to university and to major in that subject. The study by Beron (1990) on student knowledge and the desire to take additional courses is a good beginning in seeking answers to questions about the alternative ways in which students and teachers value discipline-specific knowledge. Fournier and Sass (2000) provide a good example of modeling student choice in course selection and persistence.

Once we move beyond looking at a single teaching outcome, the question is: using multiple outcomes and multiple inputs, are the teachers and/or students technically efficient in combining the inputs and the outcomes? A teacher and/or student is technically inefficient if, when compared to other teachers and/or students with similar levels of inputs, greater student outcomes could be achieved without increasing input use, or equivalently the same level of student outcomes could be achieved with fewer inputs. Conceptually, although this is difficult in practice as seen in the production function estimates of Anaya (1999), Kuh, Pace, and Vesper (1997), and others, regression residuals could be used to suggest inefficiencies. DEA (data envelope analysis) is a linear programming technique for evaluating the efficiency of decision-makers when there are multiple outcomes. DEA could be used to determine whether the teacher and/or student exhibits best practices or, if not, how far from the frontier of best practices the teacher and/or student lies.

Unfortunately, no one engaged in education research on teaching and learning has yet used DEA in a meaningful way for classroom teaching practices. Lehmann and Warning (2003) attempt to assess efficiency of United Kingdom universities in producing research and teaching but their teaching input measures (e.g., number of teachers and expenditures on libraries) and the teaching outputs measures (e.g., drop-out rate, employment of grads) are too broad to extract recommendations about

teaching practices. Johnes and Johnes (1995) and Thursby (2000) provide applications to research outputs and inputs of economics departments in the United Kingdom and the United States, respectively, where the research outputs are counts on department publications, citations, and numbers of Ph.D.s awarded in a fixed time period, but again no implications for teaching can be extracted. Surveys of the DEA method are provided by Lovell (1993) and Ali and Seiford (1993).

### **INPUT COVARIATES IN MULTIVARIATE ANALYSES**

If truly random experiments could be designed and conducted in education, then a multivariate analysis with covariates to control for consequences other than the treatment effects would not be needed. But, as already stated, no one has ever designed and no one likely ever will design a perfect experiment in education: recall that randomization should be thought of in degrees rather than as an absolute.

Almer, Jones and Moeckel (1998) provide a good example of a study in accounting classes that attempts to randomize applications of multi-level treatments (different combinations of types of one-minute paper and quiz types) across classrooms. But even following the random assignment of treatments, Almer, Jones, and Moeckel recognize the need to control for differences in student ability, as well as other covariates believed to influence student performance. They use analysis of variance (ANOVA) to determine the effect of different types of one-minute papers on multiple-choice and essay response quiz scores. This type of study design and method of analysis is in keeping with the laboratory science view advanced by Campbell, Stanley, and Gage (1963).

As noted early, educationalists have adopted the economist's view that learning involves a production function in which student and teacher inputs give rise to outputs. A regression function is specified for this input-output analysis. For example, Kuh, Pace, and Vesper (1997) estimate a regression in which students' perceived gains (the outputs) are produced by input indices for student background variables, institutional variables and measures for good educational practices (active learning). A traditional ANOVA table, like that found in Almer, Jones, and Moeckel (1998), can be produced as a part of any regression analysis. Unlike the traditional ANOVA analyses, however, regression modeling makes assumptions explicit, provides estimates of effect sizes directly, and extends to more complex analyses necessitated by data limitations and violations of assumptions that are algebraically

tractable. Traditional ANOVA is driven by the design of the experiment whereas production function and regression equation specifications are driven by theory.

In a typical production function (or input-output) study, a standardized multiple-choice test is used to measure each student's knowledge of the subject at the beginning (pretest) and end of a program (posttest). A change score for each student is calculated as the difference between his or her post-program score and pre-program score. The post-program scores, the change scores, or any one of several transformations of post and pre-program scores are assumed to be produced by human-specific attributes of the students (called human capital: e.g., SAT or ACT scores, initial subject knowledge, grade points, previous courses of study), utilization measures (e.g., time spent by student or teacher in given activities), and technology, environment or mode of delivery (e.g., lectures, group work, computer use). Of all the variations considered by researchers, the only consistently significant and meaningful explanatory variables of student final achievement are pre-aptitude/achievement measures such as SAT/ACT scores, GPA, class rank, etc. (As discussed in the next section, even the importance of and the manner in which time enters the learning equation is debated.) The policy implications could not be more clear: in order to produce students who are highly knowledgeable in a subject, start with those who already have a high ability.<sup>16</sup> The implications for educational research are likewise clear: unless covariates for students' aptitudes and/or prior skill levels are included in the explanation of learning, the results are suspect.<sup>17</sup>

The input-output approach (as well as traditional ANOVA) has five problems. First, production functions are only one part of a student's decision-making system. Observed inputs (covariates) are not exogenous but are determined within this system. Second, data loss and the resulting prospect for sample-selection bias in the standard pretest and posttest design are substantial, with 20 to 40 percent of those enrolled in large classes who take a pretest no longer enrolled at the time of the posttest. Third, from probability and statistical theory, we know that failure to reject the null hypothesis does not imply its acceptance. Because an experimental teaching method shows no statistically significant improvement over the lecture does not imply that it is not better. Fourth, although an aptitude/ability measure is essential in the explanation of final student achievement, how this covariate enters the system is not trivial because measurement error in explanatory variables implies bias in the coefficient estimators. Fifth, as already discussed, education is a multi-product output

that cannot be reflected in a single multiple-choice test score. These problems with the application of the production function mind-set are being addressed by econometricians and psychometricians, as seen for example in the new RAND corporation study of class size in California and the exchange between Rosenbaum (1999) and four commenters in *Statistical Science* (August 1999) on design rules and methods of estimation for quasi-experiments.

### **ENDOGENEITY IN A STUDENT DECISION-MAKING FRAMEWORK**

There are theoretical models of student behavior that provide a rationale for why researchers fail to find consistent evidence of the superiority of one teaching technique over another in the production of learning. For example, Becker (1982) constructed a model in which a student maximizes the utility (or satisfaction) of different forms of knowledge, current consumption, and expected future income.<sup>18</sup> This utility maximization is subject to a time constraint, to the production relationships that enable the student to acquire knowledge and to consume out of income, and to the manner in which the different forms of knowledge are measured and enter into future income. The “prices” of different forms of knowledge reflect opportunity costs generated by the time constraint, production functions, and uncertain future income. The desired student outcomes and time allocation decisions are endogenous or determined simultaneously within the system.

Becker’s (1982) model shows that improved teaching technology that enables students to more efficiently convert study time into knowledge in one subject need not result in any change in student desire for more of that knowledge. The time savings that result from the more efficient pedagogy in one course of study may be invested by the student in the acquisition of knowledge in other subjects or may be used for market work or leisure. The “prices” of the different forms of knowledge and the marginal utility of each form of knowledge, leisure, and future income in equilibrium determine student choices. It is not only the production function relationship that gives rise to a certain mix of inputs being combined to produce a given output. The levels of the output and inputs are simultaneously determined; the inputs do not cause the outputs in the unidirectional sense that independent variables determine the dependent variable.

Allgood (2001) modifies Becker’s (1982) model to show the lack of student effort when students are targeting grade levels in a given subject for which a new

teaching or learning technology has been introduced. These models make explicit how rewards for academic achievement in one subject affect achievement in that subject as well other subjects that jointly enter a student's decision-making framework as endogenous inputs that are simultaneously determined in the student's choices. Researchers working with the test-score data are thus wise to check if students take the tests seriously. Unfortunately, many educators continue to overlook the effect of incentives on measured student performance.

In their study of time on task, Admiraal, Wubbels, and Pilot (1999) do not recognize that observed student allocation of time and output produced are endogenous. An observation of a reduction in time students devote to a subject (following the introduction of an alternative teaching technique in the subject) without a decrease in achievement can result from the introduction of an efficient teaching/learning technique. On the other hand, observing no difference in achievement but an increase in time students devote to the subject suggests the introduction of an inefficient method. Such may be the case for the cooperative learning, group-oriented, and open-ended question approach (structured active-learning sections, SAL) versus the lecture style, and challenging quantitative, individualized homework and test questions (response learning, RL) in the work of Wright et al. (1997). They report that after an experiment at the University of Wisconsin – Madison in which SAL and RL sections in chemistry were taught at the “high end of the performance scale . . . students in both sections had performed equivalently” (p. 4), but SAL students spent 15 percent more time in out-of-class work than RL students.<sup>19</sup> Although Wright et al. report other worthwhile affective domain differences, on the basis of the oral examination results, the cooperative learning, group-oriented, and open-ended question approach of the structured active-learning approach was inefficient. Proponents of cooperative learning such as Klionsky (1998, p. 336) when confronted with their own students' negative reaction regarding time usage quip: “they have a hard time objectively judging its advantages or disadvantages.”

### **DATA LOSS AND SAMPLE SELECTION**

Becker and Powers (2001) show how studies including only those students who provide data on themselves and persist to the end of the semester are suspect in assessing the contribution of class size in student learning.<sup>20</sup> Missing data points

could be caused by students failing to report, data collectors failing to transmit the information, the researcher “cleaning the data” to remove unwanted items, or students simply not providing the data. Unlike inanimate objects or animals in a laboratory study, students as well as their instructors can self-select into and out of studies.

Well-designed studies such as that of Wright et al. (1997) address issues of self-selection into treatment groups. But few studies outside of economics consider the fact that a sizable proportion of students who enroll in introductory courses subsequently withdraw, never completing the end-of-course evaluation or final exam. A typical study of the gain or change scores from the beginning to the end of the course excludes all who do not complete a posttest. The process that determines which students quit between the pretest and the posttest is likely related to the process that determines test scores. That is, both persistence and final exam score are related in the student decision-making process (they are endogenous). Becker and Powers provide probit model estimates (which are simultaneously done with the estimation of the achievement equation via maximum likelihood routines) showing, all else equal, that individual students with higher pre-course knowledge of economics are more prone to persist with the course than those with lower pre-course knowledge; and those in smaller classes are likewise more likely to persist to the final exam.<sup>21</sup> Controlling for persistence, class size does affect student learning.

When studies ignore the class size and sample selection issues, readers should question the study’s findings regardless of the sample size or diversity in explanatory variables.<sup>22</sup> Hake (1998), for example, does not call attention to the fact that his control group, which made little or no use of interactive-engaging teaching methods, had a mean class size of 148.9 students (14 classes and 2084 students), but his experimental class size average was only 92.9 students (48 classes and 4458 students). Hake gives us no indication of beginning versus ending enrollments, which is critical information if one wants to address the consequence of attrition. Admiraal, Wubbels, and Pilot (1999) acknowledge that missing data could be a problem but have no idea of how to deal with the fact that in their two courses only 44.2 percent (349 of 790 students) and 36.6 percent (133 of 363 students) of enrolled students attended the exam and the seminar where questionnaires were administered. This is particularly troubling because one of the objectives of their study is to see how time on task, as reported on the questionnaires, affects exam performance.

The timing of withdrawal from a course is related to many of the same variables that determine test scores (Anderson, Benjamin, and Fuss 1994). For example, taking high school calculus and economics contributed greatly to a student's desire to complete the entire two-semester college economics course. However, more experienced students were more likely to drop sooner; they did not stick around if they saw "the handwriting on the wall." Consistent with these results, Douglas and Sulock (1995) conclude that prior experience with economics, accounting, and mathematics, as well as class attendance, all increase the probability of a student completing an economics course. They also show how correction for self-selection out of the course influenced the production function relationship between the standard input measures and the course grades of those who stayed, even though the course drop rate was only 12 percent. Becker and Walstad (1990) reveal yet another source of selection bias when test scores are to be explained; if test administration is voluntary, teachers who observe that their average class score is low on the pretest may not administer the posttest. This is a problem for multi-institution studies, such as that described in Hake (1998) where instructors elected to participate, administer tests and transmit data.

As already stated, missing observations on key explanatory variables can also devastate a large data set. Students and their instructors are selective in what data they provide, and those collecting and processing the data may be selective in what they report. Because there is no unique way to undo the censoring that is associated with missing data, any conclusion drawn only from students and their instructors who provide data must be viewed with skepticism regardless of the sample size. This point was lost on Piccinin (1999) in his study of how advocates of alternative teaching methods affect teaching and learning. His outcome measure was a classroom mean score from a multi-item student evaluation form. Of interest was whether any of three different levels of consultation by teaching resource center staff members with instructors had an effect on student evaluations of the instructors. (Levels of consultation: FC = interview/discussion between instructor and consultant; FCO = FC plus observation of classroom by consultant; and FCOS = FCO plus meeting between consultant and instructor's students).

Of the 165 instructors who consulted the teaching center during a seven-year period, 91 had data at the time of consulting (Pre 2) and at the end of the semester or year after consulting (Post 1), and only 80 had data three years after consultation (Post

2). Although we do not have the individual instructor data (which is needed for an analysis of selection), the discussion provided by Piccinin gives some idea of the potential selection problems. Piccinin reports that assistant professors are overrepresented in FC group. That the  $t$  statistic (based on an assumption of identical population variances) rises from - 0.3378 (for Post 1 minus Pre 2 mean changes) to a significant 2.2307 (for Post 2 minus Pre 2 mean changes) may be the result of three low-ranking faculty members being terminated. At the other extreme, the relatively low-scoring senior faculty member in the time-intensive FCOS group could be demonstrating nothing more than regression or self-selection into this group.

In the absence of perfect randomized experiments, with no entry or exit, selection problems at some point in the sampling process can always be identified. But should we care if we cannot teach a subject to the uninterested and unwilling? We are always going to be teaching to self-selected individuals, so why should our experiments not reflect the actual conditions under which we work? Why worry about what does not apply?<sup>23</sup> On the other hand, if building enrollment in our programs and departments is important, then the previously uninterested students are the ones that must be attracted. We need to understand the selection process in students choosing and persisting in courses, as well as in measuring their learning.

### **TESTS FOR THE EFFECTS OF INSTRUCTIONAL VARIABLES**

Hanushek (1991; 1994) and others writing in the economics of education literature in the 1980s and early 1990s advanced the notion that instructional variables (class size, teacher qualifications, and expenditures on the like) are unimportant in explaining student learning.<sup>24</sup> More recently the vote counting, meta-analysis employed by Hanushek has come under attack by educationalists, Hedges, Lane and Greenwald (1994a; 1994b) and economist Krueger (2000).<sup>25</sup> Regardless of the merits of the Hedges et al. and Krueger challenges, Hanushek's or any other researcher's conclusion that certain instructional variables are insignificant in explaining student test scores (and thus acceptance of the null hypothesis of no average effect in the populations is confirmed) is wrong.

Statisticians cringe at the idea of "accepting the null hypothesis." The null hypothesis of no learning effect can never be accepted for there is always another hypothesized value, in the direction of the alternative hypothesis, that cannot be rejected with the same sample data and level of significance. The Type II error

inherent in accepting the null hypothesis is well known but largely ignored by researchers.

The power of the test (one minus the probability of not rejecting the null hypothesis when the null is false) can always be raised by increasing the sample size. Thus, if statistical significance is the criterion for a successful instructional method, then ever-larger sample sizes will “deliver the goods.” Statistical significance of an instructional method might be demonstrated with a sufficiently large sample, but the difference in change scores will likely be trivial on multiple-choice tests with 25 to 40 items (the number of questions typically required to demonstrate an internally reliable test that able students can complete in a 50 to 75 minute period). Differences of only a few correct answers in pretest and posttest comparisons of control and experimental group results are the rule, not the exception, even after adjusting for sample selection.

Similar to small changes in test scores producing statistically significant difference of no practical importance, student evaluations of instructors can produce statistically significant differences with no real difference in teacher performance. For instance, Piccinin (1999, pp. 77-78) reports that the 0.28 point increase in mean aggregated student evaluation scores from 3.77 to 4.05, for those consulting with a teaching specialist, is statistically significant, but the 0.16 point decrease from 4.01 to 3.85, for those also observed in the classroom, is not. What is the practical meaning of the 0.16 difference? As psychologist McKeachie (1997, p. 1223), a long-time provider of college teaching tips, puts it: “Presentation of numerical means or medians (often to two decimal places) leads to making decisions based on small numerical differences – differences that are unlikely to distinguish between competent and incompetent teachers.”

That “practical importance” is more relevant than “statistical significance” does not tell us to ignore p-values, the standard errors on which they are based, or other measures of dispersion. Klionsky’s (1998) failure to report standard errors or any other descriptive statistics related to variability makes it impossible to assess the sensitivity of the estimate to random sampling error. Recent emphasis on reporting “effect sizes” without reference to standard errors, statistical significance and the interpretation of unstandardized magnitudes, as seen for example in Admiraal, Wubbels, and Pilot (1999), ignores the insights that can be gained from this information. The point is not whether descriptive statistics (means, standard errors,

etc.) of actual magnitudes should be reported – they should. The point is that researchers cannot blindly use the sharp edge of critical values in hypotheses testing.

As already suggested, one of the myths of educational research is that students in a classroom can be treated as if they were randomly assigned rats in a laboratory experiment where descriptive statistics are sufficient for analysis. This is seen in much of the discussions involving the reporting of “effect size” (e.g., a confidence interval: estimator  $\pm$  margin of error). These discussions typically focus on which one of the dozens of measures of effect size is best for meta-analysis (Elmore and Rotou 2001; Thompson 2002) with only passing reference to estimator bias resulting from sampling problems and model misspecifications.

Conceptually, results from multiple studies employing the same outcomes and descriptive statistics could be aggregated to form a meta-analysis; however, any statistician who has attempted to conduct a meaningful meta-analysis must be bothered by the fact that there is no unique way to perform the aggregation. The fact that there are numerous articles advocating one or another method of aggregation should lead even the naive researcher to be suspicious. When considering doing meta-analysis or relying on the results of a meta-analysis there are at least five issues to consider.

First, there may be no way to interpret combined results from studies employing diverse models and estimation methods. For example, what is the meaning of two apples plus three oranges equaling five fruit?

Second, the order in which comparisons are made may affect the results. For example, assume one researcher says teaching/learning method A is preferred to B, which is preferred to C. A second researcher says method B is preferred to C, which is preferred to A; and a third researcher says method C is preferred to A, which is preferred to B. What is the preferred teaching/learning method across these three researchers if we first assess whether A is preferred to B, with the winning A or B method then compared to C? Instead, what is the preferred teaching/learning method across these three researchers if we first ask if B is preferred to C, with the winning B or C method then compared with A? Nobel laureate in economics Kenneth Arrow (1951) recognized this and related paradoxes of voting behavior and aggregation schema with his “Impossibility Theorem.”

A third example of an aggregation problem in sampling was provided by Jessica Utts (1991) at a History of Philosophy of Science seminar at the University of California at Davis:

Professors A and B each plans to run a fixed number of Bernoulli trials to test

$$H_0: p = 0.25 \text{ versus } H_A: p > 0.25$$

Professor A has access to large numbers of students each semester to use as subjects. In his first experiment, he runs 100 subjects, and there are 33 successes (p-value = 0.04, one-tailed). Knowing the importance of replication, Professor A runs an additional 100 subjects as a second experiment. He finds 36 successes (p-value = 0.009, one-tailed).

Professor B teaches only small classes. Each quarter, she runs an experiment on her students to test her theory. She carries out ten studies this way, with the following results.

Attempted Replications by Professor B

| n  | Number of successes | One-tailed p-value |
|----|---------------------|--------------------|
| 10 | 4                   | 0.22               |
| 15 | 6                   | 0.15               |
| 17 | 6                   | 0.23               |
| 25 | 8                   | 0.17               |
| 30 | 10                  | 0.20               |
| 40 | 13                  | 0.18               |
| 18 | 7                   | 0.14               |
| 10 | 5                   | 0.08               |
| 15 | 5                   | 0.31               |
| 20 | 7                   | 0.21               |

Which professor's results are "most impressive"?

(In addition to looking at the p-values, count up the relative number of successes of each professor.)

Nobel laureate in economics Daniel Kahneman and long-time collaborator Amos Tversky provided a fourth problem involving aggregation when they demonstrated the importance of power and sample size in defining successful replications. Tversky and Kahneman (1982) distributed a questionnaire at a meeting of psychologists, with the following inquiry:

An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value,  $t = 2.46$  (one-tail p-value = 0.0275). Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of  $t$  with the same number of subjects. The direction was the same in both sets of data. You are reviewing the literature. What is

the highest value of  $t$  in the second set of data that you would describe as a failure to replicate? ( p. 28)

They reported the following results:

The majority of our respondents regarded  $t = 1.7$  as a failure to replicate. If the data of two such studies ( $t = 2.46$  and  $t = 1.7$ ) are pooled, the value of  $t$  for the combined data is about 3.00 (assuming equal variances).<sup>26</sup> Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study. (p. 28)

Fifth, a meta-analysis requires that the studies underlying the results do not have material faults; yet, those doing meta-analysis, like that of Springer, Stanne, and Donovan (1997) on the effect of learning in small groups, make no attempt to impose a quality criterion on the studies they consider. The quality of educational research is the focus of this chapter and its importance cannot be overestimated.

In closing this discussion of statistical tests, it is worth remembering that the use of the  $Z, T, \chi^2, F$  or any other probability distribution requires that the sampling situation fits the underlying model assumed to be generating the data when critical values are determined or p-values are calculated. Every estimator has a distribution but it may not be the one assumed in testing. For example, the standardized sample mean,  $Z = (\bar{X} - \mu) / (s_x / \sqrt{n})$ , is normally distributed if  $X$  is normally distributed or if the sample size  $n$  is large. The asymptotic theory underpinning the Central Limit Theorem also shows that for a sufficiently large sample size  $Z$  is normal even if the population standard deviation  $s_x$  is unknown. If the sample size is small and  $s_x$  is unknown, then  $Z$  is not normal, but it may be distributed as Gosset's  $T$  if  $X$  is itself normally distributed. (Similar conditions hold for the other common distributions used in parametric hypotheses testing.) When mean sample scores are 4.01 and 4.05, on a 5-point scale, with sample standard deviations in the 0.37 and 0.38 range for the small sample shown in Piccinin's (1999) Table 3, the normality assumption is untenable. The ceiling of 5 must be treated as a readily reachable truncation point in the population distribution; thus, the population cannot be assumed to be normal.

Violations of distribution assumptions require more complex modeling or a move to nonparametric statistics, as demonstrated by Becker and Greene in this volume for estimating the likelihood that a student will increase his/her grade in a

second course where course grades are discrete (A, B, C, D, or F) and bounded (A is an achievable maximum and F is an achievable minimum). Biometricians, psychometricians, econometricians, and like specialists in other disciplines are doing this in noneducation-based research. Outside economics, there is little indication that such is being done in the scholarship of teaching and learning research.

### **ROBUSTNESS OF RESULTS**

In the move to more complex modeling, it is always possible that the modeling and method of estimation, and not the data, are producing the results. For instance, labor economists are well aware that parametric sample-selection adjustment procedures (described in endnote 22) can produce spurious results. An option is to report results under alternative sets of assumptions or with different (parametric or nonparametric) estimators.

There are several examples of researchers reporting the consequence of alternative measures of the outcome measures. There are also a few examples of authors reporting results with and without key explanatory variables that are measured with error. As mentioned earlier, Almer, Jones, and Moeckel (1998) discuss the effect of the one-minute paper on student learning, with and without the use of student reported GPA as a covariate.

Outside of economic education I could find no examples of education researchers checking alternative regression model specifications. Examples of such checking within economics can be seen in Chizmar and Ostrosky (1999) in their analysis of the one-minute paper reporting regression results for the posttest on the pretest, and other explanatory variables, and the change score on the other explanatory variables. Becker and Powers (2001) consider regressions of posttest on the pretest, and other explanatory variables, and the change score on the other explanatory variables, with and without the use of self-reported GPA, and with and without adjustment for sample selection.

### **CONCLUSION**

In drawing conclusions from their empirical work, few authors are as blatant as Ramsden (1998, p. 355): “The picture of what encourages students to learn effectively at university is now almost complete.” Yet, few either recognize or acknowledge the typical fallacies that result from using pre- and posttest, mean-

different  $t$  tests to assess learning differences between a control and treatment group. Fewer still appreciate the complexity of specifying and estimating an appropriate population model that is believed to be generating the data; they neither address nor attempt to adjust for the many sample-selection problems associated with the testing of students in real educational settings.

Education is a multi-outcome endeavor. Researchers attempting to capture these varied outcomes with a single index will always be subject to aggregation problems. The untried DEA approach to multi-outcome production may be an alternative that does not require aggregation of outcomes and may provide easily interpretable measures of technical efficiency in teaching and learning. As authors acknowledge (but then proceed regardless), the use of the educational production functions with test scores as the only output measure is too narrow. Pretest and posttest, single-equation specifications, with potentially endogenous regressors, simply may not be able to capture the differences that we are trying to produce with diverse teaching methods. Adjustments for sample-selection problems are needed but even after these adjustments with large samples, failure to reject the null hypothesis of no instructional effect may point more to deficiencies in the multiple-choice test outcome measure or application of the classical experimental design than to the failure of the alternative instructional method under scrutiny.

The state of quantitative research has changed greatly in the past 30 years primarily through discipline-based scholarship in the social sciences at the major research universities. The movement primarily at lower tier universities to assess teaching and learning has ignored these developments in quantitative methods. Here I have provided only some of the mathematical and statistical shortcomings that education researchers are overlooking by not working from alternative and explicitly specified population models that may be generating the sample data on individuals and classes of those individuals. The next step for readers who want to learn more of the current state of the science for model specification, estimation and testing of education treatment effect might be to read a series of short articles on evaluating treatment effects by Manski (2001), Heckman and Vytlačil (2001), Smith and Todd (2001) and Ichimura and Taber (2001). We will not learn much, however, from continuing to apply a research paradigm intended for randomized laboratory experiments when our study designs are far from random.

## REFERENCES

- Admiraal, W., T. Wubbels, and A. Pilot. 1999. College teaching in legal education: Teaching method, students' time-on-task, and achievement. *Research in Higher Education* 40 (6): 687-704.
- Ali, A. I., and L. Seiford. 1993. The mathematical programming approach to efficiency analysis. In H. Fried, C. A. K. Lovell, and S. Schmidt, eds., *Measurement of production efficiency*. New York: Oxford University Press.
- Allgood, S. 2001. Grade targets and teaching innovations. *Economics of Education Review* 20 (October): 485-94.
- Almer, E. D., K. Jones, and C. Moeckel. 1998. The impact of one-minute papers on learning in an introductory accounting course. *Issues in Accounting Education* 13 (3): 485-97.
- Anaya, G. 1999. College impact on student learning: Comparing the use of self-reported gains, standardized test scores, and college grades. *Research in Higher Education* 40 (5): 499-526.
- Anderson, G., D. Benjamin, and M. Fuss. 1994. The determinants of success in university introductory economics courses. *Journal of Economic Education* 25 (Spring): 99-121.
- Angelo, T. A., and P. K. Cross. 1993. *Classroom assessment techniques: A handbook for college teachers*. San Francisco: Jossey-Bass.
- Arrow, K. 1951. *Social choice and individual values*. Monograph No. 12. Cowles Commission for Research in Economics. New York: John Wiley and Sons.
- Becker, W. E. 1982. The educational process and student achievement given uncertainty in measurement. *American Economic Review* 72 (March): 229-36.
- \_\_\_\_\_ 2000. Teaching economics in the 21st century. *Journal of Economic Perspectives* 14 (Winter): 109-19.
- Becker, W. E., and W. H. Greene. 2001. Teaching statistics and econometrics to undergraduates. *Journal of Economic Perspectives* 15 (Fall): 169-82.
- Becker, W. E., and C. Johnston. 1999. The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record* 75 (December): 348-57.
- Becker, W. E., and J. Powers. 2001. Student performance, attrition, and class size given missing student data. *Economics of Education Review* 20 (August): 377-88.
- Becker, W. E., and M. Salemi. 1979. The learning and cost effectiveness of AVT supplemented instruction: Specification of learning models. *Journal of Economic Education* 8 (Spring): 77-92.

Becker, W. E., and W. Walstad. 1990. Data loss from pretest to posttest as a sample selection problem. *Review of Economics and Statistics* 72 (February): 184-88.

Beron, K. J. 1990. Joint determinants of current classroom performance and additional economics classes: A binary/continuous model. *Journal of Economic Education* 21 (Summer): 255-64.

Campbell, D., and D. Kenny. 1999. *A primer on regression artifacts*. New York: The Guilford Press.

Campbell, D., J. Stanley, and N. Gage. 1963. *Experimental and quasi-experimental design for research*. Boston: Houghton Mifflin.

Card, D., and A. Krueger. 1996. The economic return to school quality. In W. Becker and W. Baumol, eds., *Assessing educational practices: The contribution of economics*, pp. 161-82. Cambridge, MA: MIT Press.

Chen, Y., and L. B. Hoshower. 1998. Assessing student motivation to participate in teaching evaluations: An application of expectancy theory. *Issues in Accounting Education* 13 (August): 531-49.

Chickering, A. W., and Z. Gamson. 1987. Seven principles for good practice in undergraduate education. *AAHE Bulletin* 39 (7): 3-7.

Chizmar, J., and A. Ostrosky. 1999. The one-minute paper: Some empirical findings. *Journal of Economic Education* 29 (Winter): 3-10.

Cottel, P. G., and E. M. Harwood. 1998. Using classroom assessment techniques to improve student learning in accounting classes. *Issues in Accounting Education* 13 (August): 551-64.

DeNeve, K. M., and M. J. Heppner. 1997. Role play simulations: The assessment of an active learning technique and comparisons with traditional lectures. *Innovative Higher Education* 21 (Spring): 231-46.

Douglas, S., and J. Sulock. 1995. Estimating educational production functions with corrections for drops. *Journal of Economic Education* 26 (Spring): 101-13.

Elmore, P., and O. Rotou. 2001. A primer on basic effect size concepts. Paper presented at the April Annual Meeting of the American Educational Research Association, Seattle, WA.

Fabry, V. J., R. Eisenbach, R. R. Curry, and V. L. Golich. 1997. Thank you for asking: Classroom assessment techniques and students' perceptions of learning. *Journal of Excellence in College Teaching* 8 (1): 3-21.

Fleisher, B., M. Hashimoto, and B. Weinberg. 2002. Foreign GTAs can be effective teachers of economics. *Journal of Economic Education* 33 (Fall): 299-326.

- Finn, J., and C. M. Achilles. 1990. Answers and questions about class size: A statewide experiment. *American Educational Research Journal* 27 (Fall): 557-77.
- Fournier, G., and T. Sass. 2000. Take my course, please: The effect of the principles experience on student curriculum choice. *Journal of Economic Education* 31 (Fall): 323-39.
- Francisco, J. S., M. Trautmann, and G. Nicoll. 1998. Integrating a study skills workshop and pre-examination to improve student's chemistry performance. *Journal of College Science Teaching* 28 (February): 273-78.
- Friedman, M. 1992. Communication: Do old fallacies ever die? *Journal of Economic Literature* 30 (December): 2129-32.
- Gleason, J., and W. Walstad. 1988. An empirical test of an inventory model of student study time. *Journal of Economic Education* 19 (Fall): 315-21.
- Goldberger, A. S. 1991. *A course in econometrics*. Cambridge: Harvard University Press.
- Greene, W. H. 2003. *Econometric analysis*. 5th ed. New Jersey: Prentice Hall.
- Gunsalus, C. K. 2002. Rethinking protections for human subjects. *Chronicle of Higher Education* 49 (November 15): B24.
- Hake, R. R. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics* 66 (January): 64-74.
- Hansen, W. L., A. Kelley, and B. Weisbrod. 1970. Economic efficiency and the distribution of benefits from college instruction. *American Economic Review Proceedings* 60 (May): 364-69.
- Hanushek, E. 1991. When school finance 'reform' may not be a good policy. *Harvard Journal of Legislation* 28: 423-56.
- \_\_\_\_\_ 1994. Money might matter somewhat: A response to Hedges, Lane, and Greenwald. *Educational Researcher* 23 (May): 5-8.
- Harwood, E. M. 1999. Student perceptions of the effects of classroom assessment techniques (CATs). *Journal of Accounting Education* 17 (4): 51-70.
- Harwood, E. M., and J. R. Cohen. 1999. Classroom assessment: Educational and research opportunities. *Issues in Accounting Education* 14 (November): 691-724.
- Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153-62.
- Heckman, J., and J. Smith. 1995. Assessing the case for social experiments. *Journal of Economic Perspectives* 9 (Spring): 85-110.

- Heckman, J., and E. Vytlačil. 2001. Policy-relevant treatment effects. *American Economic Review Proceedings* 91 (May): 108-11.
- Hedges, L., R. Lane, and R. Greenwald. 1994a. Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher* 23 (April): 5-14.
- \_\_\_\_\_. 1994b. Money does matter somewhat: A reply to Hanushek. *Educational Researcher* 23 (May): 9-10.
- Ichimura, H., and C. Taber. 2001. Propensity-score matching with instrumental variables. *American Economic Review Proceedings* 91 (May): 119-24.
- Johnes, J., and G. Johnes. 1995. Research funding and performance in UK university departments of economics: A frontier analysis. *Economics of Education Review* 14 (3): 301-14.
- Kennedy, P., and J. Siegfried. 1997. Class size and achievement in introductory economics: Evidence from the TUCE III data. *Economics of Education Review* 16 (August): 385-94.
- Kelley, T. 1927. *The interpretation of educational measurement*. New York: World Book.
- Klionsky, D. J. 1998. A cooperative learning approach to teaching introductory biology. *Journal of College Science Teaching* 28 (March/April): 334-38.
- Krueger, A. B. 2000. Economic considerations and class size. Princeton University Industrial Relations Section Working Paper No. 477, [www.irs.princeton.edu](http://www.irs.princeton.edu), September.
- Krueger, A. B., and D. M. Whitmore. 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *Economic Journal* 111 (January): 1-28.
- Kuh, G. D., C. R. Pace, and N. Vesper. 1997. The development of process indicators to estimate student gains associated with good practices in undergraduate education. *Research in Higher Education* 38 (4): 435-54.
- Kvam, P. 2000. The effect of active learning methods on student retention in engineering statistics. *American Statistician* 54 (2): 136-40.
- Lazear, E. 1999. Educational production. NBER Working Paper Series, National Bureau of Economic Research, No. 7349.
- Lehmann, E., and S. Warning. 2003. Teaching or research? What affects the efficiency of universities. Working Paper, Department of Economics, University of Konstanz, Germany.

Lovell, C. A. K. 1993. Production frontiers and productive efficiency. In H. Fried, C. A. K. Lovell, and S. Schmidt, eds., *Measurement of production efficiency*. New York: Oxford University Press.

Manski, C. 2001. Designing programs for heterogeneous populations: The value of covariate information. *American Economic Review Proceedings* 91 (May): 103-06.

Maxwell, N., and J. Lopus. 1994. The Lake Wobegon effect in student self-reported data. *American Economic Review Proceedings* 84 (May): 201-05.

McKeachie, W. 1997. Student ratings: The validity of use. *American Psychologist* 52 (November): 1218-25.

Moulton, B. R. 1986. Random group effects and the precision of regression estimators. *Journal of Econometrics* 32 (August): 385-97.

Piccinin, S. 1999. How individual consultation affects teaching. In C. Knapper, ed., *Using consultants to improve teaching. New Directions for Teaching and Learning* 29 (Fall): 71-83.

Ramsden, P. 1998. Managing the effective university. *Higher Education Research & Development* 17 (3): 347-70.

Rosenbaum, P. 1999. Choice as an alternative to control in observational studies. *Statistical Science* 14 (August): 259-78.

Salemi, M., and G. Tauchen. 1987. Simultaneous nonlinear learning models. In W. E. Becker and W. Walstad, eds., *Econometric modeling in economic education research*, pp. 207-23. Boston: Kluwer-Nijhoff.

Smith, J., and P. Todd. 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review Proceedings* 91 (May): 112-18.

Springer, L., M. E. Stanne, and S. Donovan. 1997. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. ASHE Annual Meeting Paper. November 11.

Thompson, B. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31 (April): 25-32.

Thursby, J. G. 2000. What do we say about ourselves and what does it mean? Yet another look at economics department research. *Journal of Economic Literature* 38 (June): 383-404.

Trautwein, S. N., A. Racke, and B. Hillman. 1996/1997. Cooperative learning in the anatomy laboratory. *Journal of College Science Teaching* 26 (December/January): 183-89.

- Tversky, A., and D. Kahneman. 1982. Belief in the law of small numbers. In D. Kahneman, P. Slovic and A. Tversky, eds., *Judgment under uncertainty: Heuristics and biases*, pp. 23-31. Cambridge and New York: Cambridge University Press.
- Utts, J. 1991. Replication and meta-analysis in parapsychology. *Statistical Science* 6 (4): 363-403.
- Wilkinson, L. 1999. Statistical methods in psychology journals: Guideline and explanations. *American Psychologist* 54 (August): 594-604.
- Wainer, H. 2000. Kelley's paradox. *Chance* 13 (1) Winter: 47-48.
- Wilson, R. 1986. Improving faculty teaching effectiveness: Use of student evaluations and consultants. *Journal of Higher Education* 57 (March/April): 196-211.
- Wright, J. C., R. C. Woods, S. B. Miller, S. A. Koscuik, D. L. Penberthy, P. H. Williams, and B. E. Wampold. 1997. A novel comparative assessment of two learning strategies in a freshman chemistry course for science and engineering majors. University of Wisconsin, Madison: LEAD Center.

---

Figure 1: Deep approach and good teaching  
Source: Ramsden 1998, p. 354

---

## ACKNOWLEDGMENT

Financial support was provided by an Indiana University, Bloomington, Scholarship of Teaching and Learning Research Grant, and by the University of South Australia, School of International Business, Faculty of Business and Management. Constructive criticism was provided by Suzanne Becker, George Kuh, and Samuel Thompson.

## NOTES

---

<sup>1</sup> For example, before he closed his Web sites, Thomas Russell “The No Significant Difference Phenomenon,” <http://cuda.teleeducation.nb.ca/nosignificantdifference/>, had brief quotes from over 355 research reports, summaries and papers on the use of technology for distance education. His site at “Significant Difference” (<http://cuda.teleeducation.nb.ca/significantdifference/>) had few. My review will not include studies of computer technology or distance learning. DeNeve and Heppner (1997, p. 232) report that in seven of the 12 studies they identified in their ERIC search “active learning techniques appear to have some benefits.” Although they do not calculate it, there is a 0.387 probability of getting at least 7 successes in 12 trials in random draws, with a 0.5 probability of success on each independent and identical trial. A p-value of 0.387 is hardly sufficient to reject the chance hypothesis. My review is restricted to selections from the traditionally published studies.

<sup>2</sup> There have been numerous attempts to force standards on journal editors in education and psychology in regard to what is and is not acceptable use of statistics. For example, although there was and continues to be a push to abandon statistical significance testing among educational researchers in psychology (Thompson 2002), the American Psychological Association Task Force on Statistical Inference did not endorse the banning or strict requirement for the use of any statistics: “Always provide some effect-size estimate when reporting a p value.” (Wilkinson 1999, p. 599)

<sup>3</sup> Other discipline-based studies employ nationally normed tests to explore various aspects of the teaching – learning environment. For example, in economics the three editions of the Test of Understanding of College Economics have been used to assess the learning effect of class size on student learning, native language of the teacher, student and instructor gender, and the lasting effect of a course in economics. Typically, these studies are institution-specific.

<sup>4</sup> Chizmar and Ostrosky (1999) was submitted to the *Journal of Economic Education* in 1997 before the publication of Almer, Jones, and Moeckel (1998), which also addresses the effectiveness of the one-minute paper.

<sup>5</sup> Unlike the mean, the median reflects relative but not absolute magnitude; thus, the median may be a poor measure of change. For example, the series 1, 2, 3 and the series 1, 2, 300 have the same median (2) but different means (2 versus 101).

<sup>6</sup> Let  $y_{it}$  be the observed test score index of the  $i^{th}$  student in the  $t^{th}$  class, who has an expected test score index value of  $m_t$ . That is,  $y_{it} = m_t + e_{it}$ , where  $e_{it}$  is the random error in testing such that its expected value is zero,  $E(e_{it}) = 0$ , and variance is  $s^2$ ,

---

$E(\mathbf{e}_{it}^2) = \mathbf{s}^2$ , for all  $i$  and  $t$ . Let  $\bar{y}_t$  be the sample mean of a test score index for the  $t^{\text{th}}$  class of  $n_t$  students. That is,  $\bar{y}_t = \bar{\mathbf{m}} + \bar{\mathbf{e}}_t$  and  $E(\bar{\mathbf{e}}_t^2) = \mathbf{s}^2/n_t$ . Thus, the variance of the class mean test score index is inversely related to class size.

<sup>7</sup> As in Fleisher, Hashimoto, and Weinberg (2002), let  $y_{gi}$  be the performance measure of the  $i^{\text{th}}$  student in a class taught by instructor  $g$ , let  $F_g$  be a dummy variable reflecting a characteristics of the instructor (e.g., nonnative English speaker), let  $x_{gi}$  be a  $(1 \times n)$  vector of the student's observable attributes, and let the random error associated with the  $i^{\text{th}}$  student taught by the  $g^{\text{th}}$  instructor be  $\mathbf{e}_{gi}$ . The performance of the  $i^{\text{th}}$  student is then generated by

$$y_{gi} = F_g \mathbf{g} + x_{gi} \mathbf{b} + \mathbf{e}_{gi}$$

where  $\mathbf{g}$  and  $\mathbf{b}$  are parameters to be estimated. The error term, however, has two components: one unique to the  $i^{\text{th}}$  student in the  $g^{\text{th}}$  instructor's class ( $u_{gi}$ ) and one that is shared by all students in this class ( $\mathbf{x}_g$ ):  $\mathbf{e}_{gi} = \mathbf{x}_g + u_{gi}$ . It is the presence of the shared error  $\mathbf{x}_g$  for which an adjustment in standard errors is required. The ordinary least squares routines employed by the standard computer programs are based on a model in which the variance-covariance matrix of error terms is diagonal, with element  $\mathbf{s}_u^2$ . The presence of the  $\mathbf{x}_g$  terms makes this matrix block diagonal, where each student in the  $g^{\text{th}}$  instructor's class has an off-diagonal element  $\mathbf{s}_x^2$ .

<sup>8</sup> Discussions of the reliability of an exam are traced to Kelley (1927). Kelley proposed a way to visualize a test taker's "true score" as a function of his or her observed score in a single equation that relates the estimated true score ( $\hat{y}_{true}$ ) to the observed score ( $y_{observed}$ ). The best estimate comes from regressing the observed score in the direction of the mean score ( $\mathbf{m}$ ) of the group from which the test taker comes. The amount of regression to the mean is determined by the reliability ( $\mathbf{a}$ ) of the test. Kelley's equation is

$$\hat{y}_{true} = \mathbf{a}y_{observed} + (1 - \mathbf{a})\mathbf{m}$$

If a test is completely unreliable (alpha is zero) the best predictor of a test taker's true score is the group mean. That is, the observed score is a random outcome that only deviated from the group mean by chance. If alpha is one (the test is perfectly reliable) then there is no regression effect and the true score is the same as the observed. Alpha between zero and one gives rise to the "errors in variables" problem discussed in later endnotes. Unfortunately, alpha is unknown and, as discussed in later endnotes, attempts to estimate it from observed test scores is tricky to say the least.

Reliability is often built into a test by placing questions on it that those scoring high on the test tend to get correct and those scoring low tend to get wrong. Through repetitive trial testing (called "test norming"), questions that contribute to

differentiating students are sought in the construction of highly reliable tests. In the extreme case, this type of test construction can be expected to yield test scores that are close to 50 percent correct regardless of the number of alternatives provided on each of many multiple-choice questions.

For instance, if each question on an  $N$  question multiple-choice test has four alternatives, then the expected chance score is  $0.25N$  items correct. But if some test takers are better guessers than others, or know more about the subject, then the test developer may experiment with repeated testing and place questions on the sequential exams that the  $q$  percent with the highest overall test score tend to get correct, and that the bottom  $q$  percent get wrong. As the identification of differentiating questions approaches perfection and as  $q$  approaches 0.5, the expected number of correct answers approaches  $0.5N$ . That is,

$$\lim_{\substack{l \rightarrow 0 \\ h \rightarrow 1 \\ q \rightarrow 0.5}} [qlN + 0.25(1 - 2q)N + qhN] = 0.5N$$

where  $N$  is the number of multiple-choice questions, each with 4 alternative answers,  $q$  is the proportion of top and bottom scored exams used for question selection,  $h$  is the proportion of correctly answered questions by the top scorers, and  $l$  is the proportion of correctly answered questions by the bottom scorers.

<sup>9</sup> A reviewer of this chapter argued, “an R-square of 0.5 is pretty good for social science work.” But this statement does not recognize the distinction between the use of  $R^2$  as a descriptive measure of a sample relationship (as it is used to describe the association between end-of-semester student evaluations and other outcomes) and a measure of goodness-of-fit in regression model building (where  $R^2 \leq 0.5$  in cross-section studies are not unusual). R-square is relatively irrelevant in the latter context, as Goldberger (1991, p. 177) makes clear:

From our perspective,  $R^2$  has a very modest role in regression analysis . . . Nothing in the CR (classical regression) model requires that  $R^2$  be high . . . the most important thing about  $R^2$  is that it is not important in the CR model. The CR model is concerned with parameters in the population, not with goodness of fit in the sample.

<sup>10</sup> Kuh, Pace, and Vesper (1997) tell us that the probability of getting this coefficient estimate is significantly different from zero at the 0.0005 Type I error level ( in a one- or two-tail test is not clear). They do not tell us how the standard errors were calculated to reflect the fact that their explanatory variables indices are themselves estimates. It appears, however, that they are treating the active learning index (as well as the other regressor indices) as if it represents only one thing whereas in fact it represents an estimate from 25 things. That is, when an estimated summary measure for many variables is used as a covariate in another regression, which is estimated with the same data set, more than one degree of freedom is lost in that regression. If the summary measure is obtained from outside the data set for which the regression of interest is estimated, then the weights used to form the summary measure must be treated as constraints to be tested.

---

<sup>11</sup> Ramsden (1998) ignores the fact that each of his 50 data points represent a type of institutional average that is based on multiple inputs; thus, questions of heteroscedasticity and the calculation of appropriate standard errors for test statistical inference are relevant. In addition, because Ramsden reports working only with the aggregate data from each university, it is possible that within each university the relationship between good teaching ( $x$ ) and the deep approach ( $y$ ) could be negative but yet appear positive in the aggregate.

When I contacted Ramsden to get a copy of his data and his coauthored “Paper presented at the Annual Conference of the Australian Association for Research in Education, Brisbane (December 1997),” which was listed as the source for his regression of the deep approach index on the good teaching index in his 1998 published article, he replied:

It could take a little time to get the information to you since I no longer have any access to research assistance and I will have to spend some time unearthing the data. The conference paper mentioned did not get written; another instance of the triumph of hope over experience. Mike Prosser may be able to provide a quick route to the raw data and definitions. (email correspondence 9/22/00)

Aside from the murky issue of Ramsden citing his 1997 paper, which he subsequently admitted does not exist, and his not providing the data on which the published 1998 paper is allegedly based, a potential problem of working with data aggregated at the university level can be seen with three hypothetical data sets. The three regressions for each of the following hypothetical universities show a negative relationship for  $y$  (deep approach) and  $x$  (good teaching), with slope coefficients of  $-0.4516$ ,  $-0.0297$ , and  $-0.4664$ , but a regression on the university means shows a positive relationship, with slope coefficient of  $+0.1848$ . This is a demonstration of “Simpson’s paradox,” where aggregate results differ from disaggregated results.

#### University One

$\hat{y}(1) = 21.3881 - 0.4516x(1)$     Std. Error = 2.8622     $R^2 = 0.81$      $n = 4$   
 $y(1): 21.8 \ 15.86 \ 26.25 \ 14.72$   
 $x(1): -4.11 \ 6.82 \ -5.12 \ 17.74$

#### University Two

$\hat{y}(2) = 17.4847 - 0.0297x(2)$     Std. Error = 2.8341     $R^2 = 0.01$      $n = 8$   
 $y(2): 12.60 \ 17.90 \ 19.00 \ 16.45 \ 21.96 \ 17.1 \ 18.61 \ 17.85$   
 $x(2): -10.54 \ -10.53 \ -5.57 \ -11.54 \ -15.96 \ -2.1 \ -9.64 \ 12.25$

#### University Three

$\hat{y}(3) = 17.1663 - 0.4664x(3)$     Std. Error = 2.4286     $R^2 = 0.91$      $n = 12$   
 $y(3): 27.10 \ 2.02 \ 16.81 \ 15.42 \ 8.84 \ 22.90 \ 12.77 \ 17.52 \ 23.20 \ 22.60 \ 25.90$   
 $x(3): -23.16 \ 26.63 \ 5.86 \ 9.75 \ 11.19 \ -14.29 \ 11.51 \ -0.63 \ -19.21 \ -4.89 \ -16.16$

---

University Means

$$\hat{y}(\text{means}) = 18.6105 + 0.1848x(\text{means}) \quad \text{Std. Error} = 0.7973 \quad R^2 = 0.75 \quad n = 3$$

$$y(\text{means}): 19.658 \quad 17.684 \quad 17.735$$

$$x(\text{means}): 3.833 \quad -6.704 \quad -1.218$$

<sup>12</sup> Although attempts at truly random selection are rare in educational research, an exception can be seen in the student/teacher achievement ratio (STAR), which was a four-year longitudinal class-size study funded by the Tennessee General Assembly and conducted by the State Department of Education. Over 7,000 students in 79 schools were randomly assigned into one of three interventions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide). Classroom teachers were also randomly assigned to the classes they would teach. Numerous research articles (e.g., Finn and Achilles 1990; Krueger and Whitmore 2001) have used the data to demonstrate the multiple outcomes advantages of smaller classes over larger size classes. I know of no like attempt at a truly randomized design at the tertiary level.

<sup>13</sup> According to Anaya (1999, p. 505), the first stage in assessing the contribution of teaching to learning, when the same instrument is not available as a pre- and posttest, is the calculation of a "residual gain score." This only requires the ability to regress some posttest score ( $y_1$ ) on some pretest score ( $z_0$ ) to obtain residuals. Implicit in this regression is a model that says both test scores are each driven by the same unobserved ability, although to differing degrees depending on the treatment experienced between the pretest and posttest, and other things that are ignored for the moment. The model of the  $i^{\text{th}}$  student's pretest is

$$z_{0i} = \mathbf{a}(\text{ability})_i + u_{0i},$$

where  $\mathbf{a}$  is the slope coefficient to be estimated,  $u_{0i}$  is the population error in predicting the  $i^{\text{th}}$  student's pretest score with ability, and all variables are measured as deviations from their means. The  $i^{\text{th}}$  student's posttest is similarly defined by

$$y_{1i} = \mathbf{b}(\text{ability})_i + v_{1i}$$

Because *ability* is not observable, but appears in both equations, it can be removed from the system by substitution. Anaya's regression is estimating the reduced form:

$$y_{1i} = \mathbf{b}_a z_{0i} + v_{a1i}, \text{ for } \mathbf{b}_a = \mathbf{b} / \mathbf{a} \text{ and } v_{a1i} = v_{1i} - (u_{0i} / \mathbf{a}).$$

Her least squares slope estimator and predicted posttest score for the  $i^{\text{th}}$  student is

$$\mathbf{b}_a = \sum_i y_{1i} z_{0i} / \sum_i z_{0i}^2 \quad \text{and} \quad \hat{y}_{1i} = \mathbf{b}_a z_{0i} = \left[ \sum_i y_{1i} z_{0i} / \sum_i z_{0i}^2 \right] z_{0i}$$

The  $i^{th}$  student's "residual gain score," is  $(y_{1i} - \hat{y}_{1i})$ . In Anaya's second stage, this residual gain score is regressed on explanatory variables of interest:

$$(y_{1i} - \hat{y}_{1i}) = \mathbf{X}_i \boldsymbol{\beta} + w_{1i}$$

where  $\mathbf{X}$  is the matrix of explanatory variables and here the subscript  $i$  indicates the  $i^{th}$  student's record in the  $i^{th}$  row. The  $\boldsymbol{\beta}$  vector contains the population slope coefficients corresponding to the variables in  $\mathbf{X}$  and  $w_{1i}$  is the error term.

Unfortunately, the problems with this two-stage procedure start with the first stage:  $b_a$  is a biased estimator of  $\mathbf{b}_a$ .

$$\begin{aligned} E(b_a) &= E\left(\frac{\sum_i y_{1i} z_{0i}}{\sum_i z_{0i}^2}\right) \\ &= \mathbf{b}_a + E\left\{\frac{\sum_i [v_{1i} - (u_{0i} / \mathbf{a})] z_{0i}}{\sum_i z_{0i}^2}\right\} \end{aligned}$$

Although  $v_{1i}$  and  $z_{0i}$  are unrelated,  $E(v_{1i} z_{0i}) = 0$ ,  $u_{0i}$  and  $z_{0i}$  are positively related,  $E(u_{0i} z_{0i}) > 0$ ; thus,  $E(b_a) < \mathbf{b}_a$ . As in the discussion of reliability in other endnotes, this is yet another example of the classic regression to the mean outcome caused by measurement error in the regressor. Notice also that the standard errors of the ordinary least squares  $\boldsymbol{\beta}$  vector estimator does not take account of the variability and degrees of freedom lost in the estimation of the residual gain score.

<sup>14</sup> Let the change or gain score be  $\Delta y = [y_1 - y_0]$ , which is the posttest score minus the pretest score, and let the maximum change score be  $\Delta y_{\max} = [y_{\max} - y_0]$ , then

$$\frac{\partial(\Delta y / \Delta y_{\max})}{\partial y_0} = \frac{-(y_{\max} - y_1)}{(y_{\max} - y_0)^2} \leq 0, \text{ for } y_{\max} \geq y_1 \geq y_0$$

<sup>15</sup> Let the posttest score ( $y_1$ ) and pretest score ( $y_0$ ) be defined on the same scale, then the model of the  $i^{th}$  student's pretest is

$$y_{0i} = \mathbf{b}_0(\text{ability})_i + v_{0i},$$

where  $\mathbf{b}_0$  is the slope coefficient to be estimated,  $v_{0i}$  is the population error in predicting the  $i^{th}$  student's pretest score with ability, and all variables are measured as deviations from their means. The  $i^{th}$  student's posttest is similarly defined by

$$y_{1i} = \mathbf{b}_1(\text{ability})_i + v_{1i}$$

The change or gain score model is then

$$y_{1i} - y_{0i} = (\mathbf{b}_1 - \mathbf{b}_0)\text{ability} + v_{1i} - v_{0i}$$

And after substituting the pretest for unobserved true ability we have

$$\Delta y_i = (\Delta \mathbf{b} / \mathbf{b}_0) y_{0i} + v_{1i} - v_{0i} [1 + (\Delta \mathbf{b} / \mathbf{b}_0)]$$

The least squares slope estimator  $(\Delta \mathbf{b} / \mathbf{b}_0)$  has an expected value of

$$E(\Delta \mathbf{b} / \mathbf{b}_0) = E\left(\frac{\sum_i \Delta y_i y_{0i}}{\sum_i y_{0i}^2}\right)$$

$$E(\Delta \mathbf{b} / \mathbf{b}_0) = (\Delta \mathbf{b} / \mathbf{b}_0) + E\left\{\frac{\sum_i [v_{1i} - v_{0i} - v_{0i}(\Delta \mathbf{b} / \mathbf{b}_0)] y_{0i}}{\sum_i y_{0i}^2}\right\}$$

$$E(\Delta \mathbf{b} / \mathbf{b}_0) \leq (\Delta \mathbf{b} / \mathbf{b}_0)$$

Although  $v_{1i}$  and  $y_{0i}$  are unrelated,  $E(v_{1i} y_{0i}) = 0$ ,  $v_{0i}$  and  $y_{0i}$  are positively related,  $E(v_{0i} y_{0i}) > 0$ ; thus,  $E(\Delta \mathbf{b} / \mathbf{b}_0) \leq \Delta \mathbf{b} / \mathbf{b}_0$ . Becker and Salemi (1979) suggest an instrumental variable technique to address this source of bias and Salemi and Tauchen (1987) suggest a modeling of the error term structure.

Hake (1998) makes no reference to this bias when he discusses his regressions and correlation of average normalized gain, average gain score and posttest score on the average pretest score. In <http://www.consecol.org/vol5/iss2/art28/>, he continued to be unaware of, unable or unwilling to specify the mathematics of the population model from which student data are believed to be generated and the method of parameter estimation employed. As the algebra of this endnote suggests, if a negative relationship is expected between the gap closing measure

$$g = (\text{posttest} - \text{pretest}) / (\text{maxscore} - \text{pretest})$$

and the pretest, but a least-squares estimator does not yield a significant negative relationship for sample data, then there is evidence that something is peculiar. It is the lack of independence between the pretest and the population error term (caused, for example, by measurement error in the pretest, simultaneity between  $g$  and the pretest, or possible missing but relevant variables) that is the problem. Hotelling receives credit for recognizing this endogenous regressor problem (in the 1930s) and the resulting regression to the mean phenomenon. Milton Friedman received a Nobel prize in economics for coming up with an instrumental variable technique (for estimation of consumption functions in the 1950s) to remove the resulting bias inherent in least-squares estimators when measurement error in a regressor is suspected. Later Friedman (1992, p. 2131) concluded: "I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data ..." Similarly, psychologists Campbell and Kenny (1999, p. xiii) state: "Regression toward the mean is an artifact that as easily fools statistical experts as lay people." But unlike Friedman, Campbell and Kenny do not recognize the instrumental variable method for addressing the problem.

In an otherwise innovative study, Paul Kvam (2000) correctly concluded that there was insufficient statistical evidence to conclude that active-learning methods (primarily through integrating students' projects into lectures) resulted in better retention of quantitative skills than traditional methods, but then went out on a limb by concluding from a scatter plot of individual student pretest and posttest scores that students who fared worse on the first exam retain concepts better if they were taught

using active-learning methods. Kvan never addressed the measurement error problem inherent in using the pretest as an explanatory variable. Wainer (2000) calls attention to others who fail to take measurement error into account in labeling students as “strivers” because their observed test scores exceed values predicted by a regression equation.

<sup>16</sup> Given the importance of pre-course aptitude measures, and the need to tailor instruction to the individual student, it is curious that faculty members at many colleges and universities have allowed registrars to block their access to student records for instructional purposes. As Maxwell and Lopus (1994) report, students are less than accurate in providing information about their backgrounds. Thus, as discussed in this chapter, using student self-reported data in regressions will always involve problems of errors in variables. Salemi and Tauchen (1987) discuss other forms of errors in variables problems encountered in the estimation of standard single-equation learning models.

<sup>17</sup> The effect of omitting a relevant explanatory variable, such as aptitude, from a model of learning, as measured by the difference between the posttest score ( $y_1$ ) and pretest score ( $y_0$ ), depends on the relationship between the included and excluded variables. To see this assume the true linear model of the  $i^{th}$  student’s learning,  $\Delta y_i = (y_{1i} - y_{0i})$ , is a function of only two explanatory variables ( $x_{1i}$  and  $x_{2i}$ ):

$$\Delta y_i = \mathbf{b}_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \mathbf{e}_i$$

where the critical error term assumptions are  $E(\mathbf{e}_i | x_{1i}, x_{2i}) = 0$  and  $E(\mathbf{e}_i x_{ji}) = 0$ . But if there is a suspected linear relationship between  $x_{1i}$  and  $x_{2i}$ , and if  $x_{2i}$  is omitted from the learning equation ( $\Delta y_i = \mathbf{b}_0 + \beta_1 x_{1i} + \mathbf{e}_i^r$ ), then the expected value of the  $x_{1i}$  coefficient estimator  $b_1^r$  is

$$E(b_1^r) = \mathbf{b}_1 + \mathbf{d}_1 \mathbf{b}_2$$

where  $x_{2i} = \mathbf{d}_0 + \mathbf{d}_1 x_{1i} + \mathbf{h}_i$ , the  $\mathbf{d}$ ’s are parameters, and  $\mathbf{h}_i$  is the well-behaved error term for which  $E(\mathbf{h}_i | x_{1i}) = 0$ ,  $E(\mathbf{h}_i^2 | x_{1i}) = \mathbf{s}_h^2$ , and  $E(x_{1i} \mathbf{h}_i) = 0$ . Only if the  $x_{1i}$  and  $x_{2i}$  are not related ( $\mathbf{d}_1 = 0$ ), will  $b_1^r$  be an unbiased estimator of the  $x_{1i}$  coefficient when  $x_{2i}$  is omitted. The building of models based on data-mining routines such as stepwise regression are doomed by the omitted variable problem. If a relevant variable is omitted from a regression in an early step, and if it is related to the included variables then the contribution of the included variables is estimated with bias. It does not matter with which of the related explanatory variables the model builder starts; the contribution of the included variables will always be biased by the excluded.

<sup>18</sup> The word “knowledge” is used here to represent a stock measure of student achievement; it can be replaced with any educational outcome produced by the student with various forms of study time and technology, as measured at a single point in time.

---

<sup>19</sup> Wright et al. (1997) report that 20 percent of the students in the SAL section continued to work independently (p. 4). Assuming that these students invested the same average time in out-of-class work as the RL students implies that those who truly worked together in the SAL section spent 18.75 percent more time on course work than those who worked independently.

<sup>20</sup> To assess the consequence of the missing student data on estimators in the matched pre- and posttest learning models, for example, consider the expected value of the change score, as calculated from a regression of the difference in posttest score ( $y_1$ ) and pretest score ( $y_0$ ) on the set of full information for each student. Let  $\Omega_i$  be the full information set that should be used to predict the  $i^{th}$  student's change score  $\Delta y_i = (y_{1i} - y_{0i})$ . Let  $P(m_i = 1)$  be the probability that some of this explanatory information is missing. The desired expected value for the  $i^{th}$  student's learning is then

$$E(\Delta y_i | \Omega_i) = E(\Delta y_i | \Omega_{ci}) + P(m_i = 1)[E(\Delta y_i | \Omega_{mi}) - E(\Delta y_i | \Omega_{ci})]$$

where  $\Omega_{ci}$  is the subset of information available from complete records and  $\Omega_{mi}$  is the subset of incomplete records. The expected value of the change score on the left-hand side of this equation is desired but only the first major term on the right-hand side can be estimated. They are equal only if  $P(m_i = 1)$  is zero or its multiplicative factor within the brackets is zero. Because willingness to complete a survey is likely not a purely random event,  $E(\Delta y_i | \Omega_{mi}) \neq E(\Delta y_i | \Omega_{ci})$ .

<sup>21</sup> Lazear (1999) argues that optimal class size varies directly with the quality of students. Because the negative congestion effect of disruptive students is lower for better students, the better the students, the bigger the optimal class size and the less that class size appears to matter: “. . . in equilibrium, class size matters very little. To the extent that class size matters, it is more likely to matter at lower grade levels than upper grade levels where class size is smaller.” (p. 40) However, Lazear does not address how class size is to be measured or the influence of class size on attrition. Nor does his analysis address the dynamics of class size varying over the term of a course.

<sup>22</sup> Why the  $i^{th}$  student does ( $T_i = 1$ ) or does not ( $T_i = 0$ ) take the posttest is unknown, but assume there is an unobservable continuous dependent variable  $T_i^*$  driving the student's decision; that is,  $T_i^*$  is an unobservable measure of the students propensity to take the posttest. As in Becker and Powers (2001), if  $T_i^*$  is positive, the student feels good about taking the posttest and takes it; if  $T_i^*$  is negative, the student is apprehensive and does not take it. More formally, if  $T^*$  is the vector of students' propensities to take the posttest,  $H$  is the matrix of observed explanatory variables including the pretest,  $\mathbf{a}$  is the vector of corresponding slope coefficients, and  $\omega$  is the vector of error terms, then the  $i^{th}$  student's propensity of take the posttest is given by

$$T_i^* = H_i \mathbf{a} + \omega_i$$

Taking of the posttest is determined by this selection equation with the decision rule

---

$T_i = 1$ , if  $T_i^* > 0$ , and student  $i$  takes the posttest, and  
 $T_i = 0$ , if  $T_i^* \leq 0$ , and student  $i$  does not take the posttest.

For estimation purposes, the error term  $\omega_i$  is assumed to be a standard normal random variable that is independently and identically distributed with the other error terms in the  $\omega$  vector.

The effect of student attrition on measured student learning from pretest to posttest and an adjustment for the resulting bias caused by ignoring students who do not complete the course can be summarized with a two-equation model formed by the above selection equation and the  $i^{\text{th}}$  student's learning:

$$Dy_i = X_i \mathbf{b} + \mathbf{e}_i$$

where  $Dy = (y_1 - y_0)$  is a vector of change scores,  $X$  is the matrix of explanatory variables, and again the subscript  $i$  indicates the  $i^{\text{th}}$  student's record in the  $i^{\text{th}}$  row.  $\beta$  is a vector of coefficients corresponding to  $X$ . Each of the disturbances in vector  $\epsilon$  are assumed to be distributed bivariate normal with the corresponding disturbance term in the  $\omega$  vector of the selection equation. Thus, for the  $i^{\text{th}}$  student we have

$$(\mathbf{e}_i, \mathbf{w}_i) \sim \text{bivariate normal}(0, 0, \mathbf{S}_e, I, \mathbf{r})$$

and for all perturbations in the two-equation system we have

$$E(\epsilon) = E(\omega) = \mathbf{0}, E(\epsilon\epsilon') = \sigma_\epsilon^2 I, E(\omega\omega') = I, \text{ and } E(\epsilon\omega') = \rho\sigma_\epsilon I.$$

That is, the disturbances have zero means, unit variance, and no covariance among students, but there is covariance between selection and the posttest score for a student.

The linear specification of the learning equation specification and nonlinear specification of the selection equation ensures the identification of each equation. Estimates of the parameters in the learning equation are desired, but the  $i^{\text{th}}$  change score ( $\Delta y_i$ ) is observed for only the subset of  $n$  of  $N$  students for whom  $T_i = 1$ . The regression for this censored sample of  $n$  students is

$$E(\Delta y_i | X_i, T_i = 1) = X_i \beta + E(\mathbf{e}_i | T_i^* > 0); i = 1, 2, \dots, n < N.$$

Similar to omitting a relevant variable from a regression, selection bias is a problem because the magnitude of  $E(\mathbf{e}_i | T_i^* > 0)$  varies across individuals and yet is not included in the estimation of the learning equation for the  $n$  students. To the extent that  $\mathbf{e}_i$  and  $\mathbf{w}_i$  (and thus  $T_i^*$ ) are related, estimators are biased, and this bias is present regardless of the sample size.

The learning equation regression involving matched pretest and posttest scores can be adjusted for student attrition during the course in several ways. An early Heckman-type solution to the sample selection problem is to rewrite the omitted variable component of the regression so that the equation to be estimated is

---


$$E(\Delta y_i | X_i, T_i=1) = X_i\beta + (\rho\sigma_\varepsilon)\lambda_i; i = 1, 2, \dots, n$$

where  $\lambda_i = f(-T_i^*)/[1 - F(-T_i^*)]$ , and  $f(\cdot)$  and  $F(\cdot)$  are the normal density and distribution functions. The inverse Mill's ratio (or hazard)  $\lambda_i$  is the standardized mean of the disturbance term  $w_i$ , for the  $i^{th}$  student who took the posttest; it is close to zero only for those well above the  $T = 1$  threshold. The values of  $\lambda$  are generated from the estimated probit selection equation. Each student in the learning regression gets a calculated value  $\lambda_i$ , with the vector of these values serving as a shift variable in the learning regression. The estimates of both  $\rho$  and  $\sigma_\varepsilon$  and all the other coefficients in selection and learning equations can be obtained simultaneously using the maximum likelihood routines in statistical programs such as LIMDEP or STATA.

<sup>23</sup> Heckman and Smith (1995) show the difficulty in constructing a counterfactual situation for alternative instructional methods when participation is voluntary or random. Without a counterfactual situation (i.e., what would have happened if these same people were in the control group), it is impossible to do assessment.

<sup>24</sup> Card and Krueger (1996) report a consistency across studies showing the importance of school quality on a student's subsequent earnings. They recognize that tests can be administered easily at any time in the education process and thus provide a cheap tool for monitoring programs. In recognition of the time lag for measuring earnings effects, they recommend the use of drop-out rates as an alternative to test scores for immediate and ongoing program assessment. After all, unless students finish their programs, they cannot enjoy the potential economic benefits.

<sup>25</sup> Hedges, Lane, and Greenwald (1994a; 1994b) use a meta-analysis involving an aggregation of p-values to cast doubt on Hanushek's assertion about the relevance of expenditures on instructional methods in generating test scores. Krueger (2000) reviews the Hanushek and Hedges et al. debate and contributes the observation that it is the peculiar weighting employed by Hanushek that is producing his vote-counting results. As demonstrated in this chapter, there is no unique way to do a meta-analysis.

<sup>26</sup> The first  $t$ ,  $t_1 = 2.46$ , and second,  $t_2 = 1.7$ , can be solved for  $\bar{x}_1$  and  $\bar{x}_2$  and then  $\bar{x}_3$  can be created as  $\bar{x}_3 = 0.5(\bar{x}_1 + \bar{x}_2)$  and  $t_3$  can be determined as, where the population mean is set equal to zero under the null hypothesis:

$$t_3 = \frac{(0.5) \left[ 2.46 \left( \frac{s}{\sqrt{15}} \right) + m + 1.7 \left( \frac{s}{\sqrt{15}} \right) + m \right] - m}{\frac{s}{\sqrt{15}}} = 2.94 \cong 3$$