WHAT DOES THE QUANTITATIVE RESEARCH LITERATURE REALLY SHOW ABOUT TEACHING METHODS?

William E. Becker*

For Presentation at the Scholarship of Teaching and Learning Colloquium
Indiana University – Bloomington, March 2, 2001.

Advocates and promoters of specific education methods are heard to say "the research shows that different teaching pedagogy really matters." Education specialist Ramsden (1998) asserts: "The picture of what encourages students to learn effectively at university is now almost complete."(p. 355) 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 1992-1995 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 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.[1] When advocates point to individual studies or to meta-analyses summarizing quantitative studies, little or no attention is given 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.[2]

This study advances 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 engaging students actively in the learning process. Specific criteria are advanced for both conducting and exploring the strength of discipline-specific quantitative research into

the teaching and learning process. Recent studies employing statistical inference are reviewed to identify exemplary examples of research into the teaching and learning process and to identify teaching strategies that appear to increase student learning.

Although no study may be perfect when viewed through the lens of theoretical statistics, there is relatively strong inferential evidence supporting the hypothesis that periodic use of variants of 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 other methods. This does not say, however, that alternative teaching techniques do not work. It simply says that there is no compelling statistical evidence saying that they do.

CRITERIA

A casual review of discipline-specific journals as well as general higher education journals is sufficient 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 (1991) 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.[3] 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 includes 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 endogeniety problems including simultaneous determination of variables within a system, errors in measuring explanatory variable, etc.);
8) Attention to nonrandomness, including sample selection issues and missing data problems;
9) Appropriate statistical methods of estimation, testing, and interpretation;
10) Checks on robustness of results (e.g., check on the sensitivity of results to alternative model specifications, check to ensure that it is not the model specification or estimation method itself that is producing the results); and
11) Nature and strength of claims and conclusions.

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.[4]

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 puts forward two explicit hypotheses related to student classroom participation and use of her feedback form:

H_1: Feedback Forms have no effect on student participation in class.

H_2: 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 not ideal for ease of interpretation:

H_3: 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_1: Students who write one-minute papers will perform better on a subsequent quiz than students who do not write one-minute papers.

H_{1a}: Students who write one-minute papers will perform better on a subsequent essay quiz than students who do not write one-minute papers.

H_{1b}: 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_2: 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_3: 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) ask: "Is there a significant difference in lab grades of students in cooperative learning settings versus the traditional, individual approach?"(p.186) 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) 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."(p.9) 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 poses 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 rhetoric employed by authors is not always clear on the distinction between knowledge and learning.

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 Hoshover (1998) devoted approximately a third of their thirteen 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 Cohn (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 empirical studies addressed the learning effectiveness (as measured by test scores) of the one-minute paper (endnote 3); thus, they established the need for their study.[5]

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 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.[6]

For assessing CATs and other instructional methods, 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 students in 62 introductory physics courses. He works only with mean scores for the classes; thus, his effective sample size is 62, and not 6,642. 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.[7] 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.

No matter how appealing the questions posed by the study, answering the questions depend 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)?[8]

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, ETS Advanced Placement exams. In economics, in addition to the AP exams, we have the Test of Understanding of College Economics and the Test of Economic Literacy. There are volumes written on the validity and reliability of the

SAT, ACT, GRE, and the like. I add little to this literature. Later in this paper I only comment on the appropriate use of standardized test scores assuming that anyone who has been appointed to construct a discipline-specific, nationally normed exam has at least strived 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 the advanced placement (AP) exams now have essay components. Wright, et al. (1997) report 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 prospect of contamination cannot be dismissed.

Whether multiple-choice (fixed response), essay (construct 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 observations that easy 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., only report 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 can consider the estimation of multinomial logit or probit models, Greene (2000, pp. 811-875).

More troubling for Fabry, et al. are the facts that they never define the word "enhance," and they never make clear 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 1/3 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. Thus, how do we distinguish if it is the instructors or the set of techniques that is being captured? This is a clear problem of aggregations that cannot be disentangled to get a valid answer as to 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 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.

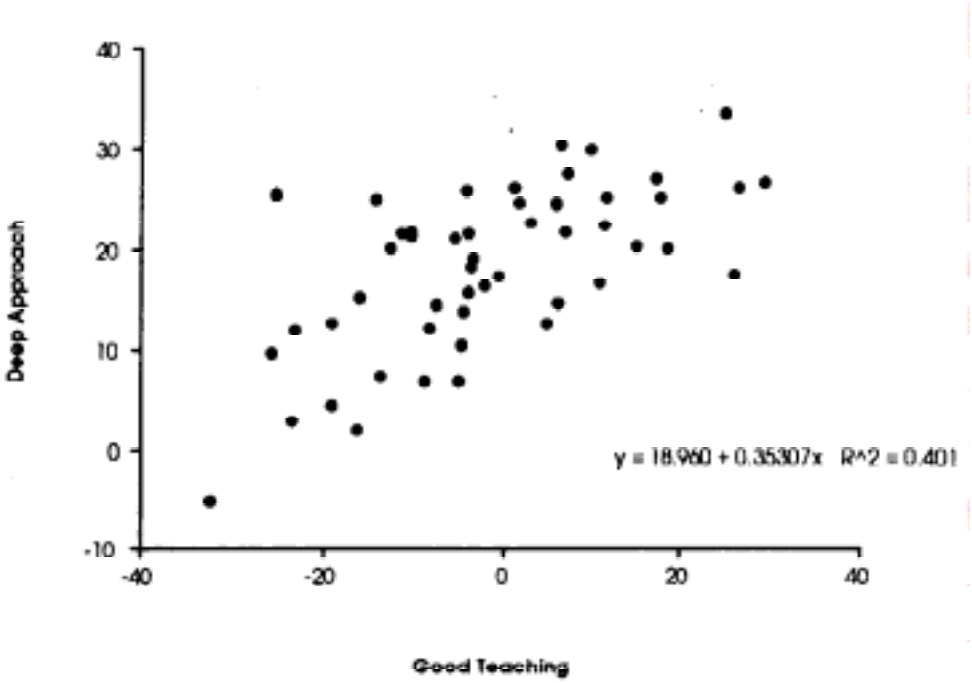
Other questions of validity arise when course grades are used as the measure of knowledge. Are individual exam grades just a measure of student knowledge at the time of the assessment? Do course grades reflect student end-of-term knowledge, learning (from beginning to end of term), rate of improvement, 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.

Indexes of performance are sometimes created to serve as explanatory variables as well as the outcome measure to be explained. Indexes 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 include

this active learning index score as one of several indexes inserted 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. [9] But what does this tell us?

Kuh, Pace, and Vesper (1997) provide multivariate analysis of students' perceived gains, including covariate indices for student background variables, 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, p. 352-54), 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.)

Figure 1: Deep Approach and Good Teaching



Source: Ramsden 1998, p. 354

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 (increases) 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.[10] Of what policy relevance is this?

SELF-REPORTED DATA

In much 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 to do so in educational research and thus acceptable 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 simultaneous from y to x, and 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 dogs education achievement equation estimates (as demonstrated mathematically in several endnotes to this paper).

Institutional policy and/or procedures adopted by registrars to over-compensate for legal privacy concerns may be the biggest obstacle to quality evaluation of CATs and other active learning methods. 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: 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 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.

OUTCOMES AND STUDY DESIGN

As already discussed, numerous measures have been proposed and used to measure cognitive– and affective–domain development of students. If truly randomized experiments could be designed and conducted in education (as discussed in Campbell and Stanley, 1963, for example), then a researcher interested in assessing cognitive– 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 alternative. 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 random sample statistic degenerates or collapses on its expected value as the sample size increases).

Unfortunately, no one has ever designed an absolutely perfect experiment; randomization is an idea better thought of in degree rather than in absolute. 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 could happen to anyone (Rosenbaum, 1999). The problem of the research in education is one of 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?

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, which has already been introduced, is the difference between a pretest (test given to students at the start of a program or course of study) and posttest (test given at the end). Similarly, for changes in attitudes the difference between a "presurvey" and "postsurvey" measure can be constructed. Calculation and use 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 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. Thus, when grades are used as the outcome measure, the research 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.

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? It also suffers from a regression to the mean problem, and if performed

with standard least squares computer programs produces test statistics based on incorrect standard errors.[11]

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. Some 30 years ago, however, Frank Ghery (1972) proposed a gap closing measure as the dependent variable for studies of educational methods where ceiling effects might be present:

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

The three test scores (*maximum score*, *posttest score,* and *pretest score*) could be defined for the individual student or an average measure 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.

Recently, Hake (1998) measured the pre- and posttest scores by the respective classroom averages on a standardized physics test. Unfortunately, the gap closing 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*.[12] 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.[13]

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 they 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 cheap hand calculator; the ability to manipulate the hand held calculator fell in value with the advance of spreadsheets, MathCAD, and statistics 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 for 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. Studies by Beron (1990), on student knowledge and the desire to take additional courses, and Vredeveld and Jeong (1990), on student-teacher goal agreement were good beginnings 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 teacher 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 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 is a linear programming technique for evaluating efficiency of decision makers when there are multiple outcomes and when meaningful aggregation is not possible. 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 doing education research on teaching and learning has yet used DEA. Johnes and Johnes (1995) and Thursby (2000) provide applications to research outputs and inputs of economics departments in the United Kingdom and United States, respectively where the research outputs are counts on department publications, citations, and numbers of Ph.Ds. awarded in a fixed time period. 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 no one has ever designed a perfect experiment: randomization is an idea better thought of in degree than in absolute. 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 could happen to anyone (Rosenbaum, 1999).

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 though the assignment of treatments was random, Almer, Jones and Moeckel recognize the need to control for differences in student ability, as well as other covariates believed to influence student performance. ANOVA was used 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 and Stanley (1963).

As introduced earlier, 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, as already introduced, 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 equations 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 clearer: to produce students who are highly knowledgeable in a subject, start with those who already have a high aptitude/ability.[14] The implications for educational research are likewise clear: unless a covariate(s) for student aptitude/abilty is included in the explanation of achievement, the results are suspect.

The input-output approach (as well as traditional ANOVA) has five problems. First, production functions are only one part of a student's decisionmaking 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 pre-post test 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. That 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. Finally, 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 in *Statistical Science* (August 1999) on design rules and methods of estimation for quasi-experiments.

ENDOGENIETY IN A STUDENT DECISIONMAKING 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.[15]

This utility maximization is subject to a time constraint, the production relationships that enable the student to acquire knowledge and consumption, and 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.

The Becker (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.[16] 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 (Forthcoming) 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 decisionmaking 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. Ignoring other issues, as discussed elsewhere in this piece, 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). Wright, et al. (1997) 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.[17] 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.[18] 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 being there to provide 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 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 decisionmaking 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 scores; and those in smaller classes are likewise more likely to persist to the final exam.[19] 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.[20] 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 does not give us any 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 form a course is related to many of the same variables that determine test scores (Anderson, et al. 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. Typically, data on students are obtained from the students themselves even though students err greatly in the data they self-report (Maxwell and Lopus, 1994). As addressed by Becker and Powers (2001), 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 descriptive statistics provided by Piccinin give some idea of the potential selection problems (Table 2). 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 to the mean as well as 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?[21] 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 choosing and persisting in courses, as well as in measuring learning.

TESTS FOR THE LEARNING EFFECT 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.[22] More recently the vote counting, meta-analysis employed by Hanushek has come under attack by educationalists, Hedges, et al., (1994a, 1994b) and economist Krueger (2000).[23] 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 in education and economics alike.

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 and the standard errors on which they are based. 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 referenced to standard errors, statistical significance and the interpretation of unstandardized magnitudes, as seen for example in Admiraal, Wubbels, and Pilot (1999), ignores what insights 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.

In closing this discussion of statistical tests, it is worth remembering that use of the Z, T, χ^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 need not be the one assumed in testing. For example, the standardized sample mean, $Z = (\overline{X} - \mu)/(\sigma_X / \sqrt{n})$, 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 σ_X is unknown. If the sample size is small and σ_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; and 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 (Forthcoming) for course grades where the units 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 non-education 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 procedure (described in Endnote 20) can produced 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 report 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): "The picture of what encourages students to learn effectively at university is now almost complete."(p. 355) But few either recognize or acknowledge the typical fallacies in using pre- and posttest, mean-different *t* tests to assess learning differences between a control and treatment group. Authors like Piccinin (1999) and Fabry, et al. (1997) acknowledge some of the shortcomings of their relatively small samples, and authors like Hake (1998) and Anaya (1999) extol the power in testing associated with large national samples. None of these studies, however, fully appreciates or attempts to adjust for the many sample selection problems in generating pre- and posttest scores.

Education is a multi-outcome endeavor. Attempting to capture these varied outcomes with an index, as in the production function approach of Kuh, Pace and Vesper (1997), will not yield easily interpretable partial effects of inputs. 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. Pre- 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.

REFERENCES

Admiraal, Wilfried, Theo Wubbels and Albert Pilot. 1999. "College Teaching in Legal Education: Teaching Method, Students' Time-on-Task, and Achievement." *Research in Higher Education*. 40:(6) pp. 687-704.

Ali, Agha Iqbal, and Lawrence Seiford. 1993. "The Mathematical Programming Approach to Efficiency Analysis." in *Measurement of Production Efficiency*. Harold Fried, C. A. Knox Lovel, and Sheldon Schmidt (Eds.) NY Oxford University Press.

Allgood, S. Forthcoming. "Grade Targets and Teaching Innovations." *Economic Education Review.*

Almer, Elizabeth Dreike, Kumen Jones and Cindy Moeckel. 1998. "The Impact of One-Minute Papers on Learning in an Introductory Accounting Course." *Issues in Accounting Education*. 13:3, pp. 485-97.

Anaya, Guadalupe. 1999. "College Impact on Student Leaning: Comparing the Use of Self-Reported Gains, Standardized Test Scores, and College Grades." *Research in Higher Education*. 40:5, pp. 499-526.

Anderson, G., D. Benjamin, and M. Fuss. 1994. "The Determinants of Success in University Introductory Economics Courses." *Journal of Economic Education*. Spring, 25:2, pp. 99-121.

Angelo, Thomas A. and Patricia K. Cross, 1993. *Classroom Assessment Techniques: A Handbook for College Teachers*. San Francisco: Josey-Bass Publishers.

Becker, William E. 1982. "The Educational Process and Student Achievement Given Uncertainty in Measurement." *American Economic Review*. March, *72*:1, pp. 229-36.

Becker, William. 2000. "Teaching Economics in the 21st Century." *Journal of Economic Perspectives*. Winter, 14:1, pp. 109-19.

Becker, William E. and William H. Greene. Forthcoming. "Teaching Statistics and Econometrics to Undergraduates." *Journal of Economic Perspectives.*

Becker, William E and Carol Johnston. 1999. "The Relationship Between Multiple Choice and Essay Response Questions in Assessing Economics Understanding." *Economic Record*. Vol. 75, December, pp. 348-357.

Becker, William E. and John Powers. 2001. "Student Performance, Attrition, and Class Size Given Missing Student Data," Forthcoming, 20(3):

Becker, William E. and William Walstad. 1990. "Data Loss from Pretest to Posttest as a Sample Selection Problem." *Review of Economics and Statistics*. February, *72*, pp. 184-88.

Beron, Kurt J. 1990. “Joint Determinants of Current Classroom Performance and additional Economics Classes: A Binary/Continuous Model” *Journal of Economic Education*. 21(3), Summer, pp. 255-264.

Campbell, D. and R. Stanley. 1963. *Experimental and Quasi-Experimental Design for Research*. Chicago : Rand McNally.

Card, David and Alan Krueger. 1996 “The Economic Return to School Quality” in William Becker and William Baumol (Eds.) Assessing Educational Practices: The Contribution of Economics. Cambridge, MA: The MIT Press. Pp. 161-182.

Chen, Yining and Leon B. Hoshower. 1998. “Assessing Student Motivation to Participate in Teaching Evaluations: An Application of Expectancy Theory.” *Issues in Accounting Education*. August, 13:3, pp. 531-49.

Chizmar, John and Anthony Ostrosky. 1999. "The One-Minute Paper: Some Empirical Findings. " *Journal of Economic Education.* Winter, 29:1, pp. 3-10.

Cottel, Philip G. and Elaine M. Harwood. 1998. “Using Classroom Assessment Techniques to Improve Student Learning in Accounting Classes.” *Issues in Accounting Education*. August, 13:3, pp. 551-64.

DeNeve, Kristina M. and Mary J. Heppner. 1997. “Role Play Simulations: The Assessment of an Active Leaning Technique and Comparisons with Traditional Lectures.” *Innovative Higher Education.* Spring, 21:3, pp. 231-46.

Douglas, Stratford and Joseph Sulock. 1995. "Estimating Educational Production Functions with Corrections for Drops." *Journal of Economic Education*. Spring, 26:2 pp. 101-13.

Fabry, Victoria J. Regina Eisenbach, Renee R. Curry and Vicki L. Golich. 1997. “Thank You for Asking: Classroom Assessment Techniques and Students’ Perceptions of Learning.” *Journal of Excellence in College Teaching.* 8:1, pp. 3-21.

Fournier, Gary, and Tim Sass. 2000. “Take May Course, Please: The Effect of the Principles Experience on Student Curriculum Choice.” *Journal of Economic Education*. 31(4), Fall, pp. 323-339.

Francisco, Joseph S. Marcella Trautmann and Gayle Nicoll. 1998. “Integrating a Study Skills Workshop and Pre-Examination to Improve Student’s Chemistry Performance.” *Journal of College Science Teaching*. February, pp. 273-78.

Friedman, Milton. 1992. “Communication: Do Old Fallacies Ever Die,” *Journal of Economic Literature*. December, 30, pp. 2129-2132.

Gleason, Joyce and William Walstad. 1988. “An Empirical Test of An Inventory Model of Student Study Time.” *Journal of Economic Education*. Fall, 19:4 pp. 315-21.

Ghery, Frank W. 1972. “Does Mathematics Matter.” Pp. 142-157. In Arthur Welsh (Ed), *Research Papers in Economic Education*. New York: Joint Council on Economic Education.

Greene, William H. 2000. *Econometric Analysis*. 4th Edition. New Jersey: Prentice Hall.

Hake, Richard R. 1998. “Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses.” *American Journal of Physics.* January, 66:1 pp. 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):364-369.

Hanushek, Eric. 1991. "When School Finance 'Reform' May Not Be a Good Policy." *Harvard Journal of Legislation*. 28, pp. 423-56.

Hanushek, Eric. 1994. “Money Might Matter Somewhat: A Response to Hedges, Lane, and Greenwald.” *Educational Researcher*, May, pp. 5-8.

Harwood, Elaine M. 1999. “Student Perceptions of the Effects of Classroom Assessment Techniques (CATs).” *Journal of Accounting Education*. 17: 4 pp. 51-70.

Harwood, Elaine M and Jeffrey R. Cohen. 1999. “Classroom Assessment: Educational and Research Opportunities.” *Issues in Accounting Education*. November,14:4 pp. 691-724.

Heckman, James. 1979. “Sample Bias as a Specific Error.” *Econometrica*. 47, pp. 153-162.

Heckman, James and Jeffrey Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*. Spring. 9:2. pp. 85-110.

Hedges, Larry; Richard Lane, and Rob Greenwald. 1994a."Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher*. April. pp. 5-14.

Hedges, Larry; Richard Lane, and Rob Greenwald. 1994b. "Money Does Matter Somewhat: A Reply to Hanushek," *Educational Researcher*. May. pp 9-10.

Johnes, Jill and Geraint Johnes. 1995. “Research Funding and Performance in UK University Departments of Economics: A Frontier Analysis.” *Economic Education Review.* 14:3, pp. 301-14.

Kennedy, Peter and John Siegfried. 1997. “Class Size and Achievement in Introductory Economics: Evidence from the TUCE III Data.” *Economics of Education Review.* 16: pp. 385-94.

Kelley, Truman. 1927. *The Interpretation of Educational Measurement*. New York: World Book.

Klionsky, Daniel J. 1998. "A Cooperative Learning Approach to Teaching Introductory Biology." *Journal of College Science Teaching*, March/April: pp. 334-38.

Krueger, Alan B. 2000. "Economic Considerations and Class Size," Princeton University Industrial Relations Section Working Paper No. 477, www.irs.princeton.edu, September.

Kuh, George D., C. Robert Pace and Nick Vesper. 1997. "The Development of Process Indicators to Estimate Student Gains Associated with Good Practices in Undergraduate Education." *Research in Higher Education.* 38:4, pp. 435-54.

Lazear, Edward. 1999. Educational Production. NBER Working Paper Series, National Bureau of Economic Research, No. 7349.

Lovell, C. A. Knox. 1993. "Production Frontiers and Productive Efficiency," in *Measurement of Production Efficiency*. Harold Fried, C. A. Knox Lovell, and Sheldon Schmidt, eds. NewYork: Oxford University Press.

Maxwell, Nan and Jane Lopus. 1994. "The Lake Wobegon Effect in Student Self-Reported Data." *American Economic Review Proceeding.* May. *84*:2, pp. 201-05.

McKeachie, Wilbert. 1997. "Student Ratings: The Validity of Use." *American Psychologist*. November, 52:11, pp. 1218-25.

Piccinin, Sergio. 1999. "How Individual Consultation Affects Teaching." *New Directions for Teaching and Learning*. Fall, pp. 71-83.

Ramsden, Paul. 1998. "Managing the Effective University." *Higher Education Research & Development*. 17(3): pp. 347-70.

Rosenbaum, Paul. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science*. August, 14:3. pp. 259-78.

Salemi, Michael and George Tauchen. 1987. "Simultaneous Nonlinear Learning Models," in *Econometric Modeling in Economic Education Research.* William E Becker and William Walstad, eds. Boston: Kluwer-Nijhoff. pp. 207-23.

Springer, Leonard, Mary Elizabeth Stanne, and Samuel Donovan. 1997. "Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis." ASHE Annual Meeting Paper. November 11.

Thursby, Jerry G. 2000. "What Do We Say about Ourselves and What Does It Mean? Yet Another Look at Economics Department Research." *Journal of Economic Literature*. June, 38: pp. 383-404.

Trautwein, Steven N, Amy Racke and Brad Hillman. 1996/1997. “Cooperative Learning in the Anatomy Laboratory.” *Journal of College Science Teaching*. December/January: pp. 183-3991.

Tversky, A.and D. Kahnemann "Belief in the law of small numbers.” In *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahnemann, P. Slovic and A. Tversky, eds., pp. 23-31. Cambridge University Press: 1982.

Utts, Jessica. 1991. "Replication and Meta-Analysis in Parapsychology." *Statistical Science*. 6(4) pp. 363-403.

Vredeveld, George, and Jin-Ho Jeong. 1990. “Market Efficiency and Student-Teacher Goal Agreement in the High School Economics Course: A Simultaneous Choice Medeling Approach.” *Journal of Economic Education*. 21(3), Summer, pp. 2317-336.

Wilson, Richard. 1986. “Improving Faculty Teaching Effectiveness: Use of Student Evaluations and Consultants.” *Journal of Higher Education*. March/April, 57:2, pp. 196-211.

Wright, John C., R. Claude Woods, Susan B. Miller, Steve A Koscuik, Debra L. Penberthy, Paul H. Williams and Bruce E. Wampold. 1997. “A Novel Comparative Assessment of Two Learning Strategies in a Freshman Chemistry Course for Science and Engineering Majors.” Wisconsin University, Madison: LEAD Center.

ENDNOTES

*William Becker is professor of economics at Indiana University – Bloomington (beckerw@indiana.edu), and adjunct professor, University of South Australia.
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 on earlier drafts was provided by Suzanne Becker and Samuel Thompson.

[1] There are exceptions, for example: Jerald Schutte ("Online Students Fare Better: Report of a Study of a Social Statistics Course," California State University, Northbridge, 1996, (http://www.csun.edu/sociology/virexp.htm) and the critique by Ed Neal (Does Using Technology in Instruction Enhance Leaning?
Or The Artless State of Comparative Research," June 1998, http://horizon.unc.edu/ts/commentary/1998-06.asp). Unfortunately, this work has not appeared in the traditional high level peer-reviewed journals typically expected by scholars. My review is restricted to published studies. In addition, Neal's critique primarily addresses only one study and lacks a general method of comparison as found in my criteria for cross-discipline comparisons.

[2] For example, Thomas Russell's Website "The No Significant Difference Phenomenon," at http://cuda.teleeducation.nb.ca/nosignificantdifference/, has 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/) has 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.

[3] Conceptually, a meta-analysis could be conducted on studies with a similar outcome measure. Anyone who has attempted to conduct a meaningful meta-analysis quickly realizes that there is no unique way to perform the aggregation.

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 2 apples plus 3 oranges equaling 5 fruit?

Second, the order in which comparisons are made may also affect results. For example, assume one research says teaching/learning method A is preferred to B, which is preferred to C. A second research say method B is preferred to C, which is preferred to A; and a third research 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?

As a fourth example of an aggregation problem in sampling, consider the question Jessica Utts (1991) posed 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: $p = 0.25$ 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"?

For those who remember how to calculate t statistics, consider a fifth example of aggregation problems in statistics as posed by A. Tversky and D. Kahnemann (1982). They 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. 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)

Tversky and Kahnemann then ask what would happen if the data of two such studies (t=2.46 and t=highest insignificant value from above) are pooled: what is the calculated value of t for the combined data? (assuming equal variances) Is there a paradox of aggregation here?

Finally, 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 criteria on the studies they consider. The quality of educational research is the focus of my work.

[4] Other discipline-based studies employ nationally normed tests to explore aspect of 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.

[5] Chizmar and Ostrosky (1999) was submitted to the *Journal of Economic Education* in 1997 before the publication of Almer, Jones and Moeckal (1998), which also addresses the effectiveness of the one-minute paper.

[6] Unlike the mean, the median reflects relative but not absolute magnitude; thus, the median may be a poor measure of change. For example, the three-item series 1, 2, 3 and the three-item series 1, 2, 300 have the same median (2) but different means (2 versus 101).

[7] 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 μ_{it}. That is, $y_{it} = \mu_{it} + \varepsilon_{it}$, where ε_{it} is the random error in testing such that its expected value is zero, $E(\varepsilon_{it}) = 0$, and variance is σ^2, $E(\varepsilon_{it}^2) = \sigma^2$, for all i and t. Let $\bar{y}_t$ be the sample mean of a test score index for the t^{th} class of n_t students. That is, $\bar{y}_t = \bar{\mu}_t + \bar{\varepsilon}_t$ and $E(\bar{\varepsilon}_t^2) = \sigma^2 / n_t$. Thus, the variance of the class mean test score index is inversely related to class size.

[8] Discussion of the reliability of an exam are traced to Kelley (1927). Kelley proposed away 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 (μ) of the group from which the test taker comes. The amount of regression to the mean is determined by the reliability (α) of the test. Kelley's equation is

$$\hat{y}_{true} = \alpha y_{observed} + (1 - \alpha)\mu$$

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. 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,

$$\lim_{\substack{l\to 0\\ h\to 1\\ q\to .5}} [qlN + 0.25(1-2q)N + qhN)] = 0.5N$$

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.
l is the proportion of correctly answered questions by the bottom scorers.

[9] 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.

[10] Ramsden completely 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 (endnote 7) and the calculation of appropriate standard errors for test statistical inference (endnote 8) 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 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)

Although I repeatedly asked for his cited data, it was never forthcoming. Aside from the murky issue of Ramsden citing his 1997 paper, which he subsequently admitted does not exist, and his not being able to provide the data on which the published 1998 paper is allegedly based, the potential problem of working with data aggregated at the university level can be demonstrated with three hypothetical university data sets. In the below spread sheet each of the three hypothetical universities show a negative relationship between *y* (deep approach) and *x* (good teaching), with slope coefficients of –0.4516, -0.0297, and –0.4664, but a regression run 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 are different from dissaggregated results.

HYPOTHETICAL DATA DEMONSTRATING SIMPSON'S PARADOX

UNIVERSITY ONE

x(4)	y(4)		
-4.11	21.8	Intercept	Slope
6.82	15.86	21.3881	-0.4516
-5.12	26.25		
17.74	14.72	Std Error	R-square
		2.8622	0.8113

UNIVERSITY MEANS

x(3means)	y(3)means		
3.833	19.658	Intercept	Slope
-6.704	17.684	18.6105	0.1848
-1.218	17.735		
		Std Error	R-square
		0.7973	0.7489

UNIVERSITY TWO

x(8)	y(8)		
-10.54	12.6	Intercept	Slope
-10.53	17.9	17.4847	-0.0297
-5.57	19		
-11.54	16.45	Std Error	R-square
-15.96	21.96	2.8341	0.0096
-2.1	17.1		
-9.64	18.61		
12.25	17.85		

UNIVERSITY THREE

x(12)	y(12)		
-23.16	27.1	Intercept	Slope
26.63	2.02	17.1663	-0.4664
5.86	16.81		
9.75	15.42	Std Error	R-square
11.19	8.84	2.4286	0.9103
-14.29	22.9		
11.51	12.77		
-0.63	17.52		
-19.21	23.2		
-4.89	22.6		
-16.16	25.9		

[11] 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 calculate 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^{th} student's pretest is

$$z_{0i} = \alpha(ability)_i + u_{0i},$$

where α is the slope coefficient to be estimated, u_{1i} 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

$$y_{1i} = \beta(ability)_i + v_{1i}$$

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:

$$y_{1i} = \beta_\alpha z_{0i} + v_{\alpha 1i}, \text{ for } \beta_\alpha = \beta/\alpha \text{ and } v_{\alpha 1i} = v_{1i} - (u_{oi}/\alpha).$$

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

$$b_\alpha = \sum_i y_{1i} z_{0i} / \sum_i z_{0i}^2 \quad \text{and} \quad \hat{y}_{1i} = b_\alpha z_{0i} = [\sum_i y_{1i} z_{0i} / \sum_i z_{0i}^2] z_{0i}$$

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:

$$(y_{1i} - \hat{y}_{0i}) = \mathbf{X}_i \tilde{\mathbf{n}} + w_{1i}$$

where **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 $\mathbf{n}$ vector contains the population slope coefficients corresponding to the variables in **X** and w_{1i} is the error term. Unfortunately, the problems with this two-stage procedure start with the first stage: b_α is a biased estimator of β_α.

$$E(b_\alpha) = E(\sum_i y_{1i} z_{0i} / \sum_i z_{0i}^2)$$
$$= \beta_\varepsilon + E\{\sum_i [v_{1i} - (u_{0i}/\alpha)]_{1i} z_{0i} / \sum_i z_{0i}^2\}$$

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, $E(b_\alpha) < \beta_\varepsilon$. As in the discussion of reliability in Endnote 8, 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 ρ estimators do not take account of the variability and degrees of freedom lost in the estimation of the residual gain score.

[12] As before 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

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

[13] 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

$$y_{0i} = \beta_0 (ability)_i + v_{0i},$$

where β_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

$$y_{1i} = \beta_1 (ability)_i + v_{1i}$$

The change or gain score model is then

$$y_{1i} - y_{0i} = (\beta_1 - \beta_0) ability + v_{1i} - v_{0i}$$

And after substituting the pretest for unobserved true ability we have

$$\Delta y_i = (\Delta\beta / \beta_0) y_{0i} + v_{1i} - v_{0i}[1 + (\Delta\beta / \beta_0)]$$

The least squares slope estimator $(\Delta b / b_0)$ has an expected value of

$$E(\Delta b / b_0) = E(\sum_i \Delta y_1 y_{0i} / \sum_i y_{0i}^2)$$
$$E(\Delta b / b_0) = (\Delta\beta / \beta_0) + E\{\sum_i [v_{1i} - v_{0i} - v_{01}(\Delta\beta / \beta_0)]_{1i} y_{0i} / \sum_i y_{0i}^2\}$$
$$E(\Delta b / b_0) \leq (\Delta\beta / \beta_0)$$

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, $E(\Delta b / b_0) \leq \Delta\beta / \beta_0$. 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. These regressions suffer from the classic errors in variables problem and regression to the mean problem associated with the use of the pretest as an explanatory index variable for unknown ability. Even those schooled in statistics continue to overlook this regression fallacy, as called to economists' attention in 1992 when Nobel laureate Milton Friedman asked "Do Old Fallacies Ever Die?"(1992, p. 2129) Forty years earlier Friedman

showed economists how to handle the regression to the mean problem associated with the estimation of a consumption function (consumption is a function of income) with the introduction of his permanent income hypothesis (1957).

[14] 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 paper, 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.

[16] 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.

[17] Wright, et al. 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.

[18] 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 Ω_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

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

where Ω_{ci} is the subset of information available from complete records and Ω_{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 braces is zero. Because willingness to complete a survey is likely not a purely random event, $E(\Delta y_i \mid \Omega_{mi}) \neq E(\Delta y_i \mid \Omega_{ci})$

[19] 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 determination over the term of a course.

[20]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 $\mathbf{T}^*$ is the vector of students' propensities to take the posttest, **H** is the matrix of observed explanatory variables including the pretest, α is the vector of corresponding slope coefficients, and ω is the vector of error terms, then the i^{th} student's propensity of take the posttest is given by

$$T_i^* = H_i\alpha + \omega_i \qquad (20.1)$$

Taking of the posttest is determined by equation (20.1) with the decision rule

$$T_i = 1, \text{ if } T_i^* > 0, \text{ and student } i \text{ takes the posttest, and} \qquad (20.2)$$
$$T_i = 0, \text{ if } T_i^* \leq 0, \text{ and student } i \text{ does not take the posttest.}$$

For estimation purposes, the error term ω_i is assumed to be a standard normal random variable that is independently and identically distributed with the other error terms in the ω 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 selection equation (20.1) and the i^{th} student's learning:

$$\Delta y_i = X_i\beta + \varepsilon_i \qquad (20.3)$$

where $\Delta y = (y_1 - y_0)$ is a vector of randomly selected change scores, **X** is the matrix of explanatory variables, and again the subscript i indicates the i^{th} student's record in the i^{th} row. β is a vector of coefficients corresponding to **X**. Each of the disturbances in vector ε are assumed to be distributed bivariate normal with the corresponding disturbance term in the ω vector of the selection equation (20.1). Thus, for the i^{th} student we have

$$(\varepsilon_i, \omega_i) \sim \text{bivariate normal}(0, 0, \sigma_\varepsilon, 1, \rho) \qquad (20.4)$$

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

$$E(\varepsilon) = E(\omega) = \mathbf{0},\ E(\varepsilon\varepsilon') = \sigma_\varepsilon^2\mathbf{I},\ E(\omega\omega') = \mathbf{I}, \text{ and } E(\varepsilon\omega') = \rho\sigma_\varepsilon\mathbf{I}. \qquad (20.5)$$

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 learning equation specification (which places the pretest on the left-hand side as opposed to having its coefficient estimated with bias as an explanatory variable on the right-hand side in a posttest regression) ensures the identification of equation (20.3), although both equations (20.1) and (20.3) are identified by the difference in functional forms. Estimates of the parameters in equation (20.3) are desired, but the i^{th} change score (Δy_i) is observed for only the subset of students for whom $T_i = 1$. The regression for this censored sample of n students is

$$E(\Delta y_i|\mathbf{X}_i, T_i = 1) = \mathbf{X}_i\beta + E(\varepsilon_i|T_i^* > 0);\ i = 1,2,\ldots,n. \qquad (20.6)$$

Similar to omitting a relevant variable from a regression, selection bias is a problem because the magnitude of $E(\varepsilon_i|T_i^* > 0)$ varies across individuals and yet is not included in the estimation of equation (20.3) for the n students. To the extent that ε_i and ω_i (and thus T_i^*) are related, estimators are biased, and this bias is present regardless of the sample size.

The learning 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

$$E(\Delta y_i|\mathbf{X}_i, T_i=1) = \mathbf{X}_i\beta + (\rho\sigma_\varepsilon)\lambda_i;\ i = 1,2,\ldots,n \qquad (20.7)$$

where $\lambda_i = f(-T^*_i)/[1-F(-T^*_i)]$, and f(.) and F(.) are the normal density and distribution functions. The inverse Mill's ratio (or hazard) λ_i is the standardized mean of the disturbance term ω_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 λ are generated from the estimated probit equation (20.1). Each student in the learning regression gets a calculated value λ_i, with the vector of these values serving as a shift variable in the learning regression. The estimates of both ρ and σ_ε and all the other coefficients in equations (20.1) and (20.3) can be obtained simultaneously using the maximum likelihood routine in statistical programs such as LIMDEP.

[21] Heckman and Smith (1995) call attention to the difficulty in constructing a counterfactual situation for an alternative instructional method when participation is voluntary or randomly assigned. 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 .

[22] 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.

[23] Hedges, Lane, and Greenwald (1994a, 1994b) use a meta-analysis involving an aggregation of p-values to cast doubt on Hanushek's assertion regarding the relevance of expenditure on instructional methods in generating test scores. A case-by-case review of their presentation of Hanushek's data, however, suggests that the focal point of much discussion in education, the teacher/pupil ratio (or class size), is irrelevant in explaining student performance when measured by 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.