ASSESSMENT OF THE FIRST-YEAR SEMINAR:
RESEARCH-BASED GUIDELINES FOR COURSE & PROGRAM EVALUATION

Joe Cuseo
Marymount College

Abstract
This manuscript attempts to provide comprehensive coverage of key issues associated specifically with evaluation of the freshman seminar and grounds them within the larger context of assessment theory and research in higher education. Four major questions about the process of freshman-seminar assessment are addressed: (a) **Why** is the assessment being conducted, i.e., What is its **purpose or objective**? (b) **What outcomes** will serve as criteria for assessing the impact of the freshman seminar? (c) **Who** should conduct the assessment project and analyze the results? and (d) **How** will the assessment be conducted, i.e., What research design or methodology will be used to assess whether the course has achieved its intended outcomes?

Eight different **research designs and methods** are described and evaluated: (a) experimental design, (b) quasi-experimental design, (c) time-series design, (d) multiple regressions analysis (a.k.a., multivariate analysis), (e) course-evaluation surveys (i.e., student ratings), (f) pre-test/post-test design, (g) analysis of students’ behavioral records (e.g., logs of student use of campus programs, trace audits, transcript analyses), and (h) qualitative research methods (e.g., focus groups, content analysis and category analysis of students' written comments).

Strategies are offered for writing the **final assessment report** in terms of (a) relating the report to institutional mission and goals, (b) tailoring content and tone of the report to specific audiences, (c) reporting results for different student subgroups, and (d) identifying how results will be acted on and put to use.

The **future** of freshman-seminar assessment is discussed in terms of what additional outcomes and alternative methods represent fertile areas for subsequent research on the freshman seminar, such as assessment of (a) specific course topics and teaching strategies that have the most impact on student outcomes; (b) how involvement in the freshman seminar influences student satisfaction with the total college experience; (c) impact of the freshman seminar on students' choice of major and their time to graduation; (d) faculty and staff perceptions of how student participation in the freshman seminar affects their behavior in the classroom and on campus; (e) impact of the freshman seminar on the transfer rate of community college students; (f) how availability of the freshman seminar at an institution can affect college marketing and student recruitment; (g) whether student performance in the freshman seminar is an effective predictor of overall student success during the first year of college; and (h) viability of the freshman seminar as a vehicle for conducting comprehensive student assessment at college entry.

The monograph concludes with a discussion of how the wide range of student outcomes targeted for freshman-seminar assessment, and the sophisticated methodologies that have been used to assess these outcomes, may serve as both a model and a stimulus for effective campus-wide assessment of other educational programs or institutional interventions designed to promote student retention and achievement.
Introduction

National survey data reveal that freshman seminars are being adopted with increasing frequency on college campuses. However, despite the increase in the number of adopted freshman seminars, the overall percentage of freshman seminars in place at American colleges and universities has remained the same (approximately 67%) since 1988, and the percentage of institutions reporting strong support for their existing freshman seminars has been declining (Barefoot & Fidler, 1996). The survey authors offer the following interpretation for these findings.

This may indicate that as freshman seminars are born, others die an untimely death for a variety of reasons which can be summarized as lack of firm institutional support . . . . Freshman seminars are generally held to higher expectations with respect to outcomes than any other course in the curriculum. The absence of such outcomes (or lack of research to demonstrate outcomes) may spell the demise of the course . . . . "Successful" seminars—those that enjoy strong broad-based institutional support and long life--are courses [that] are evaluated on a regular basis, and results of this evaluation are made available to the entire campus community (Barefoot & Fidler, 1996, pp. 5-6, 61).

In addition to increasing the likelihood of course survival, an assessment plan may also increase the likelihood of receiving institutional approval to initiate a freshman seminar program. As Kenny (1996) advises, "Assessment should be built into the proposal. Not only will this disarm opposition, but it will keep program designers focused and will enhance program credibility " (p. 71).

The freshman seminar is an educational intervention whose potential for promoting positive student outcomes qualifies it for rigorous assessment at any institution which has adopted the course, and particularly at those colleges where institutional quality is defined not simply in terms of student selectivity or national reputation, but in terms of the quality of the educational programs they provide and their demonstrated impact on student development and success.

National interest in and demand for assessment of freshman year experience programs is highlighted by John Gardner, pioneer of the freshman year experience movement and founder of the National Resource Center for the Freshman Year Experience, in a statement he made several years ago:

What I did not foresee . . . was the tremendous number of requests we would receive from educators asking for basic information on how to get started on the research they wanted to do on their own campus programs. In addition, I could not have anticipated the flood of requests we would receive from graduate students, both at the masters and the doctoral level, who would be interested in doing empirical research measuring the effectiveness of programming for freshmen (1992, p. 3).

To date, only one publication has appeared in the scholarly literature which attempts to provide systematic guidelines for assessment of freshman year programs: Primer for Research on The Freshman Year Experience by Dorothy Fidler (1992). The present monograph builds upon and
expands the recommendations offered in this earlier publication, attempts to provide comprehensive coverage of key issues associated specifically with freshman seminar evaluation, and grounds these issues within the larger context of assessment theory and research in higher education.

Key Questions and Issues Involved in Assessment of the Freshman Seminar

Assessment of the effects of the freshman seminar on student outcomes, like assessment of any other educational program, involves consideration of four major questions or issues: (a) Why is the assessment being conducted? (b) What outcomes will be assessed? (c) Who will conduct the assessment? and (d) How will the assessment be conducted? Each of these four questions will be addressed, in turn, in successive sections of this monograph.

* Why is the assessment being conducted? What is its goal or objective?

This is the first question that must be addressed in the assessment of any educational program or intervention. Its pivotal role in the assessment process is underscored by Trudy Banta, a nationally-recognized assessment scholar.

Assessment is most effective when it is based on clear and focused goals and objectives. It is from these goals that educators fashion the coherent frameworks around which they can carry out inquiry. When such frameworks are not constructed, assessment outcomes fall short of providing the direction necessary to improve programs (Banta et al., 1996, p. 22).

Two major goals of freshman seminar assessment are (a) to obtain evaluative information on the program's overall effectiveness or impact for use in bottom-line decisions about whether the program should be adopted, funded, continued, or expanded, and (b) to obtain evaluative information on the program for the purpose of improving or fine-tuning its quality.

In assessment terminology, the first purpose is referred to as "summative evaluation," i.e., assessment designed to "sum up" a program's overall value; and the second purpose is referred to as "formative evaluation," i.e., assessment that helps to form, shape, or further develop a program's effectiveness (Scriven, 1967).

* What outcomes will be selected as criteria for assessing the program's impact?

The following outcomes have been the most commonly used measures of the freshman seminar's impact.

1. Student outcomes, such as:
   - Student satisfaction with the course or student perceptions of course effectiveness—as measured by student ratings and narrative comments made on course-evaluation surveys or questionnaires administered upon course completion.
   - Student use of campus services and participation in campus activities—as measured by survey questions asking for students' reported use of campus services and their frequency of
participation in campus activities; or, by means of logs kept by student development or student service professionals that track students' use of specific campus services and participation in particular campus activities.

- Students' out-of-class interaction with faculty—as measured by the frequency and quality of such interactions reported on student surveys.
- Students' social integration into the peer community—as measured by the quantity and quality of on-campus friendships reported on student surveys.
- Student retention—as measured by (a) number of future courses/units completed, (b) student persistence to completion of the first semester of college (fall-to-spring retention), (c) student persistence to completion of the first year of college (fall-to-fall retention), and/or (d) persistence to degree/program completion.
- Students' academic achievement—as measured by (a) course content knowledge acquired in the freshman seminar, (b) cumulative grade-point average achieved at the end of the freshman year or at college graduation, (c) number of first-year students in good academic standing (versus academic probation or dismissal), (d) number of first-year courses attempted and completed (vs. dropped or failed), and (e) number of introductory courses completed with a grade of C or better.

2. Faculty/Staff outcomes, such as:

- Satisfaction with the seminar's instructor-training program, as measured by participants' survey ratings or comments.
- Changes in attitude and behavior toward students, or changes in instructional methods, which faculty report are the result of their participation in the seminar's instructor-training program.

3. Institutional outcomes, such as:

- Improved enrollment management—as measured by increased annual rates of student retention and increased institutional revenue resulting therefrom.
- Improved institutional effectiveness or efficiency—as evidenced by increased student and faculty utilization of campus resources, and reduced time taken by students to complete educational programs or degree requirements.

There are other important outcomes to which the freshman seminar may contribute which have yet to be assessed and reported in the research literature. These are discussed in a subsequent section of this monograph, titled "The Future of Freshman Seminar Assessment" (pp. ).

*Who should conduct assessment of the freshman seminar and analyze the results?*

Probably the simplest and most direct answer to this question is to have someone conduct the evaluation who has not been associated with the freshman seminar and who has no vested interest in its outcome, i.e., an "external" evaluator or "third party" assessor who would not be biased toward or against the program. This should serve to enhance the credibility of the assessment report. For example, we would have more confidence in research on the health effects of cigarette smoking that was designed, conducted, and analyzed by an external, university-affiliated research team than by researchers associated with, or employed by, the American tobacco industry.

At the University of South Carolina, evaluation of the freshman seminar is conducted by a
"highly respected researcher who has no official relationship to or responsibility for the program being studied" (Gardner, 1986, p. 271). This practice guards against evaluator bias—the tendency of the individual who designed or conducted a study, when involved in evaluating the resulting data, to unwittingly skew the findings in the direction of its intended outcome. Evaluator bias tends to occur because of (a) the "believing is seeing" trap (Weick, 1979), whereby the researcher sees what he expects to see and perceives ambiguous information in the direction of the expected or desired results (Arnoult, 1976), and (b) the "Rosenthal effect," in which unrecognized behaviors on the part of the researcher may tacitly encourage or reinforce the study's participants to respond in ways that support the intended results (Rosenthal, 1966, 1974).

It is noteworthy that ensuring the assessor be someone untied to the course or program being assessed is common practice at European colleges and universities (Adelman, 1986). In America, it is more likely for those involved in running a program or teaching a course also to be involved in its assessment. As one assessment scholar notes, "If American colleges and universities were commercial institutions, they would be in violation of the Sherman Antitrust Act for `bundling' these services" (Harris, 1986, p. 14).

One strategy for "unbundling" these two functions is to have the assessor be someone on campus who has had a history of involvement in institutional research, but who has no direct tie to the freshman seminar program. This strategy may be particularly viable today because recent national interest in the related issues of assessment and accountability has precipitated an increase in the number of institutions employing personnel with full-time or part-time responsibility for conducting on-campus assessment projects (El-Khawas, 1993).

Another option is to request the assistance of faculty. The department of education or social and behavioral sciences may be a good source of faculty who have the graduate training and professional expertise needed to conduct program assessment. Dorothy Fidler (1992) suggests that faculty from the departments of computer science, mathematics and statistics may have valuable expertise they can lend to the study's design or data analysis. As Howard Altman urges, "Make use of ‘local talent’ among the faculty . . . . There is often a great deal of faculty expertise which gets hired by other campuses by outside sources, but which is ignored on the home campus, where it is often available for free" (1988, p. 126).

Whether qualified faculty should be asked to lend their evaluation services "for free" is a sensitive and debatable issue. Given that the requested faculty member is likely to have full-time professorial responsibilities, requesting that she simply add on another major task to her existing workload might be both socially insensitive and logistically infeasible. Alternatives to this gratuitous approach could be offered to faculty in the form of extra compensation (e.g., stipends, meritorious pay), or release time from other institutional responsibilities (e.g., reduction in teaching load, publication requirements, or committee work). Rewarding faculty who are asked to conduct freshman seminar assessment should serve to increase their level of commitment and effort, and should send a clear message to the college community that the institution is fully committed to conducting a high-quality program evaluation—as evidenced by its willingness to provide evaluators with the time and/or fiscal support needed to conduct the assessment in a thorough and rigorous fashion. If such incentives cannot be provided, then at the very least, those faculty who commit their time as program evaluators should be rewarded with a formal thank-you letter. A copy of this letter should also be sent to the faculty member's department chair or academic dean, so it can be included in his retention-and-promotion file.
Another potential source of assistance in conducting assessment of the freshman seminar are students. For instance, graduate students from relevant academic departments or from student development programs could assist in program assessment. Upper-division undergraduates might also provide assistance, perhaps as research assistants on faculty-student research teams or graduate-undergraduate student research teams.

*How will the assessment be conducted, i.e., What research design or methodology will be used to assess the intended outcomes of the freshman seminar?*

The foregoing sections of this manuscript have focused on (a) the why question of assessment (i.e., its purpose or objective), (b) the what question (i.e., what types of outcome data will be collected) and (c) the who question (i.e., who will conduct the assessment). This section attempts to answer the how question (i.e., the means or methods used to collect outcome data).

The method used to be used for assessing whether or not the intended outcomes of the freshman seminar have been realized is an important decision that should be made before data are collected and analyzed (Halpern, 1987). Thus, decisions about the research design or method for evaluating the freshman seminar should be made early in the course planning and assessment process. Listed below are descriptions of a variety of potential research designs for assessing the freshman seminar, accompanied by a discussion of their relative advantages and disadvantages.

1. **Experimental Design**

   This research method involves comparing student outcomes for freshmen who are randomly assigned to either one of the following two groups: (a) an "experimental" group of students who participate in the freshman seminar, or (b) a "control" group of students who do not participate in the course.

   Historically, this method has been considered to be the scientifically ideal or "true" experimental design for evaluating educational programs because it ensures randomized assignment of students to both the experimental and control groups, thus controlling for the "volunteer effect" or "self-selection bias," i.e., the possibility that students who voluntarily decide to participate in an educational program, and then select themselves into that program, may be more intrinsically motivated and committed to college success than students who elect not to become involved in the program. In the case of the freshman seminar, any positive outcomes resulting from voluntary student participation in the course may be due to the highly motivated nature of the particular first-year students who choose to enroll themselves in the seminar, rather than to the actual effect or impact of the course itself (Fidler, 1992).

   As Pascarella and Terenzini point out in their landmark work, *How College Affects Students*,

   It has been axiomatic in the educational research community that the most valid approach for estimating the causal link between two variables and thus the net effect of one on the other is through the random assignment of individual subjects to experimental and control groups. Unfortunately, the necessary conditions for a true or randomized experiment are extremely difficult to obtain in actual field settings where self-selection rather than random assignment is the rule . . . . Perhaps the basic problem in assessing the unique influence of college on students is the issue of student self-selection or recruitment (1991, pp. 657-658).
Empirical support for the self-selection bias operating when students voluntarily enroll in the freshman seminar is provided by Schwitzer, Robbins, & McGovern (1993) who found that freshmen who enrolled voluntarily in a freshman orientation course had a better sense of goal directedness and were experiencing fewer adjustment problems than freshmen who chose not to take the course.

One procedure that has been used to circumvent this methodological problem is to solicit a larger number of students who are interested in taking the course than the actual number of course sections that can accommodate them. Half of those students who express interest in taking the course are then randomly selected to fill the available course sections (to serve as the experimental group), while the remaining half of students who had expressed interest in taking the seminar are denied access to the course (so they may serve as the control group).

To further ensure that students in both the control and experimental groups are representative of the total freshman population on campus, a "stratified" random sampling procedure may be used. In this procedure, before students are assigned to either the experimental or control group, they are subdivided into strata (subgroups or subpopulations) which represent the approximate percentage of their total population on campus (e.g., 60% female, 40% male; 25% residents, 75% commuters; 15% minority students, 85% majority students.) Students are then randomly selected from each of the designated subpopulations and assigned to both the experimental and control groups.

The major disadvantage of the experimental design is an ethical one: Its random selection of students to become course participants or non-participants (members of the control group) results in the arbitrary denial of course access to one-half of the students who want to become involved in the program and who are likely to benefit from it (Pascarella, 1986). This is akin to a common medical ethics issue involving drug research: Do you arbitrarily deny certain patients a promising drug that could significantly enhance the quality of their lives, or possibly save their lives, so they can conveniently be used as the "placebo" control group in an experiment designed to test the drug's effectiveness? Analogously, do you arbitrarily deny certain students access to a promising educational program (freshman seminar) that could significantly enhance the quality of their college experience, or enable them to survive to degree completion, so that they can serve as a control group in an experiment designed to assess the program's effectiveness?

This ethical disadvantage of the experimental design may be tempered somewhat by the argument that it is a justifiable procedure when used to conduct a "pilot study" on only one cohort of students, with the intention of gathering just enough data to serve as supporting evidence for subsequent expansion of the course, thereby ensuring the seminar's availability to all future cohorts of freshmen.

Viewed in this light, the experimental design may be seen as an ethically acceptable and methodologically rigorous research tool for marshalling initial empirical evidence to support the freshman seminar which then may be used to justify long-term, full-scale "institutionalization" of the course.

2. Quasi-Experimental Design

This research method involves comparing outcomes for freshmen who volunteer to participate in the freshman seminar (experimental group) relative to a "matched" control group, i.e., selected freshmen who have elected not to participate in the seminar but whose personal characteristics
are similar to, or "match" the experimental group on important student variables that may influence educational outcomes. For example, in previously conducted freshman-seminar assessments, students in experimental and control groups have been matched with respect to such characteristics as (a) high school grade-point average, (b) standardized college-admission test scores, (c) basic-skills placement test scores, (d) predicted GPA derived from weighted scores on college-preparation courses, (f) high school grades and SAT scores, (e) educational goals or objectives, (f) residential or commuter status, and (g) demographic characteristics such as age, gender, race or ethnicity. Matching course participants with non-participants in this fashion serves to control for, or rule out the possibility that differences in student outcomes associated with course participation could be due to the fact that course participants had personal characteristics which differed significantly from non-participants.

A major ethical advantage of this research design is that it allows all students who express interest in taking the freshman seminar to have access to the course, thus circumventing the ethical problem associated with an experimental design in which some students are arbitrarily denied course access so they can serve as a control group.

However, one methodological disadvantage of the quasi-experimental design is that students are not randomly assigned to experimental and control groups as they are in a true experimental design (hence its name, "quasi-experimental"). Consequently, this design fails to control for the volunteer effect or self-selection bias, leaving open the possibility that any positive outcomes resulting from course participation may be due to the highly motivated nature of students who elect to enroll in the freshman seminar, rather than to the effects of the course itself.

One possible strategy for addressing this limitation of the quasi-experimental design is to survey students in both the experimental group and matched-control group to assess whether they report differences in their level of college motivation. For instance, at the University of South Carolina, a short survey has been designed to assess students' motivation to stay in college and complete their degree. This survey has been administered to both freshman seminar participants and non-participants. Comparisons of survey responses provided by course participants and non-participants have revealed no differences between the two groups in their motivation to stay in college and complete their degree. Thus, the higher freshman-to-sophomore retention rate evidenced by students who participate in the university's freshman seminar is not likely to be an artifact of students selecting themselves into the course because of their higher level of motivation to persist and succeed in college (Fidler, 1991).

Similarly, the University of South Carolina has administered the Cooperative Institutional Research Program (CIRP) survey to assess the "joining" behavior of freshmen prior to college entry. These survey results indicate that there are no differences between students who eventually decide to enroll in the freshman seminar and those who do not. Hence, the higher rate of participation in co-curricular activities evinced by course participants is unlikely to be due to self-selection by students with greater motivation for extracurricular involvement (Gardner, 1994).

3. Time-Series Design

In this research design, outcomes assessed after implementation of the freshman seminar are compared with the same outcomes achieved prior to the seminar's adoption. For example, freshman-to-sophomore retention rates at the college after adoption of the freshman seminar are
compared with freshman-to-sophomore retention rates for the years preceding course adoption.

The advantage of this design is that it provides a type of "historical" control group--against which the effects of seminar participation may be compared--without having to withhold the course from a portion of entering freshmen so they can serve as a "contemporary" control group. Hence, this design circumvents the ethical drawback of an experimental design in which some students are deliberately recruited for the freshman seminar and then arbitrarily deprived course access so they can be used as a control group.

However, two caveats must be issued with respect to the time-series research design: (a) The personal characteristics of entering freshmen during years before and after implementation of the freshman seminar should be similar or matched so that any changes in student outcomes subsequent to course implementation cannot simply be due to historical changes in the entry characteristics of the freshman class (e.g., more academically qualified freshmen entering the institution during and after implementation of the freshman seminar). (b) Two or more years of outcome data should be gathered before and after institutional initiation of the freshman seminar in order to compare pre- and post-seminar outcomes--not just the year immediately before and after program implementation--because any year-to-year fluctuations in student outcomes (e.g., retention) may simply be due to random chance deviation (Pascarella, 1986). Gathering data for two or more years before and after program implementation also results in a larger sample size which can enhance the power or sensitivity of statistical tests (Cohen, 1988), such as t-tests and chi square analyses, which may be used to detect pre- to post-implementation differences in student outcomes.

Ramapo College (New Jersey) has employed this time-series design to demonstrate that the average freshman-to-sophomore retention rate for cohorts of entering students who enrolled in the freshman seminar during a five-year period immediately after the course became a requirement was significantly higher than the average retention rate for freshmen entering the college during the three-year period immediately before course adoption (Ramapo College in Barefoot, 1993).

4. Multiple Regression Analysis (a.k.a., Multivariate Analysis)

This statistical procedure, or some variant thereof, has been the favored research design of contemporary scholars interested in assessing how college experiences affect student outcomes (e.g., Astin, 1993; Pascarella & Terenzini, 1991). In short, multiple regression analysis involves computing correlations between student-outcome variables (e.g., student retention or academic performance) and two other types of variables: (a) student input variables (e.g., entering students' SAT scores) and (b) college experience variables (e.g., student participation in the freshman seminar).

For a detailed explanation of multivariate analysis, consult the appendices in Astin (1991) or Pascarella and Terenzini (1991). The following synopsis of multivariate analysis has been adapted from Astin (1993) and applied to assessment of the freshman seminar.

The first step in multiple regression analysis is to calculate correlations between all influential student-input characteristics and a single student outcome in order to obtain a "prediction score" or "estimated outcome" score for that particular outcome (e.g., the predicted or estimated first-year GPA for college freshmen based on their entering high-school GPA, SAT, and placement-test scores). This estimated outcome score, based on characteristics which students bring with
them to the institution, serves as a type of "statistical" control group or baseline against which to compare the effects of later college experiences on student outcomes. For instance, if students who participate in a freshman seminar go on to display a higher retention rate than would be expected or predicted from their college-entry characteristics, then this discrepancy (called the "residual score") suggests that participating in the seminar (a college experience variable) is having an effect on retention (a student outcome variable).

The amount of the seminar's effect can be assessed by computing the correlation between the residual score it produces and the student outcome in question. This partial correlation (called the "beta" coefficient) represents the degree to which the educational experience and the student outcome are statistically related—after all other potentially biasing student characteristics have been controlled for. In other words, it represents what the freshman seminar experience adds to the predicted student outcome--above and beyond what can be predicted by student input characteristics.

Thus, it might be said that multiple regression analysis attempts to control for confounding student variables statistically, i.e., by computing and comparing correlations between student variables and outcomes, whereas the aforementioned experimental and quasi-experimental research designs attempt to gain this control procedurally, i.e., by the procedures used to select and assign students to experimental and control groups.

Multiple regression analysis can also be adapted to assess whether the effect of the freshman seminar (or any other college-experience variable) on a student outcome is either "direct" or "indirect." A college-experience variable can be considered to have a direct effect on a student outcome if its beta coefficient remains statistically significant even after the correlations of all other college-experience variables with that student outcome have been included in the regression analysis. This suggests that a particular college-experience variable is making a unique or independent contribution to the student outcome that cannot be accounted for by other college-experience variables.

A college-experience variable may be deemed to have an indirect effect on a student outcome if its beta coefficient, which was significant after student input (entry) characteristics were controlled for, is later reduced to nonsignificance when other college-experience variables are added to the regression equation. This suggests that the effect of the particular college-experience variable is accounted for, or mediated by, other college-experience variables.

There are three key advantages associated with the use of multiple regression analysis as a research method for assessing outcomes of the freshman seminar:

(a) It circumvents the disadvantage of a "true" experimental design in which freshmen are denied access to the course so they can be used as a control group.

(b) It allows investigators to assess whether the addition of individual college-experience variables results in any incremental change in the predicted score for a student outcome (Banta, 1988). For example, multiple regression analysis could be used to answer the following question: Would student participation in the freshman seminar, plus their participation in a pre-semester orientation program, result in a higher than predicted rate of student retention than participation in the seminar alone?

(c) It allows investigators to compute the percentage of outcome variance that is attributable to a particular student-input or college-experience variable (by squaring its beta coefficient), thus providing an estimate of the variable's relative influence on the outcome under investigation.
(Pascarella & Terenzini, 1991). For example, via multiple regression analysis it would be possible to compute the approximate percentage of total variance in college students' first-year GPA that is attributable to participation in the freshman seminar, relative to the percentage of variance in first-year GPA that is attributable to students' entering SAT, high-school GPA, or placement-test scores.

Two limitations of multiple regression analysis have been cited in the assessment literature: (a) The procedure does not allow assessment of how joint or common variance between college-experience variables and student-input variables may interact to influence outcomes (Hanson, 1988). (b) It assumes that any variance in outcome that may be attributed to the joint influence of student-input and college-experience variables is attributed solely to the student-input variable, thus the influence of student-input variables on outcomes may be overestimated while the influence of college-experience variables are underestimated (Pascarella & Terenzini, 1991).

However, proponents of multiple regression analysis consider these to be relatively minor limitations which have little adverse effect on the overall validity and interpretability of the results generated by this statistical procedure (Astin, 1991; A.W. Astin, personal communication, October 21, 1992).

There is one final consideration should be kept in mind when using any statistical test to evaluate the significance of the freshman seminar's impact. If desirable student outcomes attributed to the freshman seminar are not found to be statistically significant, the seminar's effect on certain institutional outcomes may still prove to be practically significant. For instance, a freshman seminar which results in a very modest 5-10% increase in student retention may generate a gain in the college's total enrollment number that does not reach a level of statistical significance; however, the revenue gained from this modest increase in additional tuition-paying customers may contribute significantly to the institutional budget, particularly at colleges whose operational budgets are heavily tuition dependent.

This institutional outcome has been assessed at Seton Hall University. The cost/benefit ratio of its freshman studies program was evaluated by means of two statistical techniques that are commonly used in business to evaluate the economic benefits of alternative courses of action: (a) "break-even analysis" (Larimore, 1974), and (b) "elasticity coefficient" (Hoffman, 1986). Two faculty from the university's department of economics used these procedures to assess whether the total revenue generated by its freshman studies program equaled or exceeded the total costs incurred by the program. They found that the break-even point for an entering class of approximately 1,000 students who participated in Seton Hall's freshman studies program was 21 students, which represented an increased retention rate of only about two percent. This means that if implementation of the program leads to the retention of 21 additional students who would otherwise have withdrawn from the college, the program will have paid for itself. The architects of this campus-specific study concluded that Seton Hall's freshman studies program was "cost efficient [and] will more than pay for itself in economic terms alone without taking into account the quality benefits that accrue to the university and the retained students" (Ketkar & Bennet, 1989, p. 43).

These findings are consistent with early cost-effectiveness research on the freshman seminar (University 101) conducted at the University of South Carolina, whose Office of Finance reported that for every $1.00 used to support its freshman seminar, the program generated $5.36 in return (Gardner, 1981). What these campus-specific research reports strongly suggest is that
comprehensive evaluation of the freshman seminar's overall impact should involve not only its statistical effect with respect to student outcomes, but also its fiscal effect as an institutional outcome.

5. Course-Evaluation Surveys: Student Ratings

National research reveals that student ratings are the most widely used source of information for assessing teaching effectiveness in college (Seldin, 1993), and student ratings of the course or course instructor are the most commonly used strategy for assessing the freshman seminar (Barefoot & Fidler, 1996). Since student ratings represent the most frequently used strategy for evaluating teaching effectiveness in general, and for evaluating the freshman seminar in particular, an extensive discussion of this assessment strategy is provided in this section of the monograph.

Course evaluations usually take the form of student surveys or questionnaires which are designed to assess (a) students' level of course satisfaction or their perceptions of course effectiveness, and (b) self-reported student outcomes associated with the course (e.g., students' reported use of various campus services or the frequency of their interactions with key college personnel).

One major strength of student evaluations is that their reliability and validity have probably received more empirical support than any other method of course assessment; there have been over 1300 articles and books published which contain research on the topic of student ratings (Cashin, 1988). Despite perennial criticisms of student evaluations by some faculty and the publication of some isolated studies which purportedly refute the validity of student evaluations, when the results of all studies are viewed collectively and synthesized, they provide strong support for the following conclusions about student evaluations.

(a) Students' judgments correlate positively (i.e., are in agreement with) the judgments of more experienced observers (e.g., alumni, teaching assistants, faculty peers, administrators, and trained external observers (Aleamoni & Hexner, 1980; Feldman, 1988, 1989; Marsh, 1984).

(b) Students make judgments on what is taught and how it is taught. Student judgments are not unduly influenced by their own personal characteristics (e.g., student's gender or academic ability), or by characteristics extraneous to the course, such as time of day or time of year when the course is taught (Abrami, Perry, & Leventhal, 1982; Aleamoni & Hexner, 1980; Feldman, 1977; 1979; Seldin, 1993).

(c) Students' overall ratings of course quality and teaching effectiveness correlate positively with how much they actually learn in the course--as measured by their performance on standardized final exams. In other words, students rate most highly those courses in which they learn the most and those instructors from whom they learn the most (Abrami, d'Apollonia, & Rosenfield, 1997; Centra, 1977; Cohen, 1981, 1986; McCallum, 1984).

(d) Student evaluations do not depend heavily on the student's age (Centra, 1993) or level of college experience. For example, lower-division students do not provide ratings that differ systematically from upper-division students (McKeachie, 1979).

(e) Students distinguish or discriminate among specific dimensions and components of course instruction. For example, students give independent ratings to such course dimensions as course organization, instructor-student rapport, and the quality of course assignments (Marsh, 1984). As
Aleamoni (1987) illustrates, "If a teacher tells great jokes, he or she will receive high ratings in humor . . . but these ratings do not influence students' assessments of other teaching skills" (p. 27).

Bill McKeachie, who has been engaged in national research on student ratings for three decades, recently provided a succinct summary of why we should take student evaluations seriously.

Decades of research have related student ratings to measures of student learning, student motivation for further learning, instructors' own judgments of which of two classes they had taught more effectively, alumni judgments, peer and administrator evaluations, and ratings by trained observers. All of these criteria attest to the validity of student ratings well beyond that of other sources of evidence about teaching (McKeachie & Kaplan, 1996, p. 7).

Moreover, a large body of research has consistently reputed commonly held myths about student ratings. For instance, the following findings fail to support traditional criticisms of student evaluations.

(a) Students who receive higher course grades do not give higher course ratings (Theall, Franklin, & Ludlow, 1990; Howard & Maxwell, 1980, 1982).

(b) Students do not give lower ratings to difficult or challenging courses which require a heavy work load (Marsh & Dunkin, 1992; Sixbury & Cashin, 1995).

(c) Student evaluations are not unduly influenced by the instructor's personality and popularity; for example, entertaining teachers do not necessarily receive higher overall student ratings (Costin, Greenough, & Menges, 1971; McKeachie, et al., 1978; Marsh & Ware, 1982).

(d) Student ratings do not change over time or with students' post-course experiences; in contrast, there is substantial agreement between student evaluations given at the time of course completion and retrospective evaluations given by the same students one-to-five years later (Feldman, 1989; Overall & Marsh, 1980). This refutes the oft-cited argument that students are immature and only with maturity, or the passage of time, will they come to appreciate courses or instructors that were initially rated poorly.

In a comprehensive review of the research literature on students' course ratings, Cashin (1995) reaches the following conclusion.

There are probably more studies of student ratings than of all of the other data used to evaluate college teaching combined. Although one can find individual studies that support almost any conclusion, for a number of variables there are enough studies to discern trends. In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation (p. 6).

This echoes the conclusion reached by Scriven (1988): "Student ratings are not only a valid, but often the only valid way to get much of the information needed for most evaluations" (quoted in d'Apollonia & Abrami, 1997, p. 19).

These research-based conclusions are not surprising when viewed in light of the following reasons why students are in a uniquely advantageous position to evaluate the quality of a course.

(a) Students witness all elements of the course, from the syllabus to the final exam, including all
intervening in-class learning experiences and out-of-class assignments. Thus, they have a comprehensive perspective of the totality of instructional elements that comprise the course. (b) Students experience multiple courses and multiple instructors which enable them to provide course assessment from a comparative perspective. (c) Multiple students experience and evaluate a given course, thus they comprise a large sample of observers, particularly if course evaluations are amalgamated across semesters or across different course sections. It is almost an assessment axiom that such a large representative sample is a necessary precondition for drawing accurate inferences and reaching valid conclusions from empirical observation (Hays, 1973; Robinson & Foster, 1979).

Another major advantage of course-evaluation surveys or questionnaires is that they are capable of generating an extensive amount of data on a large sample of respondents in a relatively short period of time. If a student-rating survey or questionnaire is well constructed and carefully administered, it can be an effective and efficient vehicle for assessing the attitudes, perspectives, and self-reported outcomes of the institution’s most valued constituent: the learner.

Nevertheless, the degree of reliability and validity of a particular student-rating survey can be influenced by the content (items) which comprise the survey and the process by which it is administered. The following recommendations are offered as strategies for maximizing the validity, interpretability, and utility of surveys designed to solicit student evaluations of the freshman seminar. Most of the following recommendations are also relevant for improving the effectiveness of surveys designed to solicit faculty/staff evaluations of any instructor-training program that happens to be offered in conjunction with the freshman seminar.

Recommendations Regarding the Course Evaluation Instrument/Form

* Cluster individual items into logical categories that represent important course components or instructional dimensions.

For instance, include items pertaining to each of the three core components of the course experience: (a) course planning and design (e.g., questions pertaining to overall course organization and clarity of course objectives); (b) classroom instruction (e.g., items pertaining to in-class teaching, such as clarity and organization of lectures or instructional presentations); and (c) evaluation of student performance (e.g., items pertaining to the fairness of tests, assignments, grading practices, and the quality of feedback provided by the instructor). Also, a healthy balance of questions pertaining to both course content (topics and subtopics) and instructional processes (in-class and out-of-class learning activities) should be included on the evaluation form.

A major advantage of this clustering or categorizing strategy is that the categories can function as signposts or retrieval cues for the designers of the survey, ensuring that the items selected for inclusion in the instrument reflect a well-balanced sample of the major course dimensions that affect the quality of the students’ learning experience.

Another advantage of grouping items under section headings is that it can function as a cue or signal to students completing the instrument that there are distinct dimensions to the course. This may help them to discriminate among these important components of course effectiveness, increasing the likelihood that they will assess them independently.

Lastly, partitioning the instrument into separate sections that reflect separate course
dimensions should help to reduce the risk of a general "halo effect," i.e., the tendency for a student to complete the evaluation instrument by going right down the same column and filling in all "1s" or "5s" on all items, depending on whether they generally liked or disliked the course.

* Provide a rating scale that allows five-to-seven choice points or response options.

There is research evidence which suggests that fewer than five choices reduces the instrument's ability to discriminate between satisfied and dissatisfied respondents, and more than seven rating-scale options adds nothing to the instrument's discriminability (Cashin, 1990).

* If possible, do not include the neutral "don't know" or "not sure" as a response option.

This alternative could generate misleading results because it may be used as an "escape route" by students who do have strong opinions but are reluctant to offer them (Arreola, 1983).

* Include items that ask students to report their behavior.

Astin (1991) suggests a taxonomy for classifying types of data that may be collected in the assessment process that includes two broad categories: (a) psychological data reflecting students' internal states, and (b) behavioral data reflecting students' activities. Traditionally, student course evaluations have focused almost exclusively on the gathering of psychological data (student perceptions or opinions). However, given that one of the major goals of most freshman seminars is to increase students' use of campus services (Barefoot & Fidler, 1996), items which generate behavioral data pertaining to use of campus services, or frequency of participation in co-curricular activities, should also be included on the evaluation instrument.

It should be noted that commercially developed instruments are available to assess students' reported behaviors and their degree of involvement with campus services and student activities, such as (a) "The College Life Task Assessment Instrument" (Brower, 1990, 1994), (b) the "College Student Experiences Questionnaire" (CSEQ)(Pace, 1984), and (c) the "Critical Incident Techniques & Behavioral Events Analysis" (Knapp & Sharon, 1975).

This raises the larger issue of whether the college should rely exclusively on locally developed ("home grown") assessment instruments for evaluating the freshman seminar, or whether they should purchase externally constructed ("store bought") instruments from testing services or centers. There are some advantages associated with external, commercially developed instruments, namely: (a) Their reliability and validity are usually well established. (b) They are efficient, saving time that the institution would have to devote to instrument construction and scoring. (c) Norms are typically provided by the testing service that allow the institution to gain a comparative perspective by assessing its own performance relative to national averages.

However, a major disadvantage of externally developed instruments is that the questions asked and results generated may be less relevant to the institution's unique, campus-specific goals and objectives than those provided by internally constructed instruments. Some testing services attempt to minimize this disadvantage by allowing the college the option of adding a certain number of their own questions to the standardized instrument. The availability of this option should be one major factor to consider when deciding whether or not to purchase an external instrument.

Another factor to consider would be the purchase and scoring costs associated with the use of an externally developed instrument relative to the anticipated cost (in time and money) to the
college if it developed and scored its own instrument. Supporting the latter approach are the results of one college-based review of assessment methods in which it was concluded that the cost of externally developed surveys make them viable only for "one-time projects or single, annual projects, and only if campus-based expertise is unavailable" (Malaney & Weitzer, 1993, p. 126).

Peter Ewell, a nationally recognized assessment scholar, warns about another subtle disadvantage associated with what he calls "off the shelf" (externally developed) instruments.

Off-the-shelf instruments are easy. Although they cost money, it's a lot less than the effort it takes to develop your own. But buying something off-the-shelf means not really engaging the issues that we should—for example, What are we really assessing? and What assumptions are we making about the nature of learning" (Mentkowski et al., 1991, p. 7).

To improve the reliability and validity of campus-specific instruments that are internally designed, a pilot study of the instrument should be conducted on a small sample of students to assess whether the instrument's instructions are clear, the wording of each of its items is unambiguous, and the total time needed to complete the instrument is manageable. As Fidler (1992) notes, "Seasoned researchers recommend pilot studies before every research project in order to identify and quickly alter procedures, thereby alleviating problems before they can invalidate an entire study" (p. 16).

* Beneath each rating item or question, print the phrase “Reason(s) for this rating:” and leave a small space for any written remarks students would like to make with respect to that particular item.

Written comments often serve to clarify or elucidate numerical ratings, and instructors frequently report that written comments are most useful for course-improvement purposes, especially if such comments are specific (Seldin, 1992). As Jacobi (1991) points out, "The typical survey consists of a majority of closed-ended items, with limited opportunities for open-ended responses. This format does not encourage students to explore their attitudes, feelings, or experiences in depth and therefore may provide incomplete information about why students think, feel, or behave in a particular manner” (p. 196).

Allowing students to write comments with respect to each individual item, rather than restricting them to the usual "general comments" section at the very end of the evaluation form, should serve to increase the interpretability of numerical rating, as well as increase the specificity of students' written remarks—which, in turn, should increase their usefulness for course or program improvement.

* Include at least two global items on the evaluation instrument pertaining to overall course effectiveness or course impact; these items can be used for summative evaluation purposes.

The following statements illustrate global items that are useful for summative evaluation:

(a) I would rate the overall quality of this course as: (poor <-> excellent).

(b) I would rate the general usefulness of this course as: (very low <-> very high).
(c) I would recommend this course to other first-year students:
(weakly agree <- - > strongly disagree).

Responses to these global items can provide an effective and convenient snapshot of students' overall evaluation of the course which can be readily used in program assessment reports. Research has repeatedly shown that these global ratings are more predictive of student learning than student ratings given to individual survey items pertaining to specific aspects or dimensions of course instruction (Braskamp & Ory, 1994; Centra, 1993; Cohen, 1986). As Cashin (1990) puts it, global items function "like a final course grade" (p. 2).

Abrami (1989) argues further that, "it does make conceptual and empirical sense to make summative decisions about teaching using a unidimensional [global] rating. This choice then frees us to recognize that the particular characteristics of effective teaching vary across instructors" (p. 227). Thus, ratings on such unidimensional or global items may be used to make summative (overall) assessments of the course or instructor. However, it is imperative not to add up the ratings for all individual items on the questionnaire and then average them in order to obtain an overall course evaluation. This procedure not only is inefficient, it is also an ineffective index of overall course satisfaction because it gratuitously assumes that each individual item carries equal weight in shaping the students' overall evaluation of the course.

Inclusion of global items on the evaluation instrument also allows for the examination of relationships between students' overall course ratings and their ratings on individual items pertaining to specific course dimensions. Such comparisons could be helpful in answering the following question: Among those students who give the course very high global ratings versus those students who give the course very low overall ratings, on which particular items in the evaluation instrument do these two groups of students display the largest discrepancy in ratings? These particular items could reveal those specific aspects or dimensions of the course that carry the most weight in determining students' overall perceptions and their overall level of satisfaction with the freshman seminar. These specific course dimensions may represent key target areas for course improvement which could be addressed in instructional-development workshops and new-instructor orientation programs.

* Include an open-ended question asking for written comments about the course's strengths and weaknesses, and how the latter may be improved.

Such questions can often provide useful information about students' general reaction to the course as well as specific suggestions for course improvement. For example, at the end of the semester, course instructors could ask students to provide a written response to a question which asks them to "describe a major change (if any) in their approach to the college experience that resulted from their participation in the freshman seminar." Or, students could be asked, "What would you have liked to learn about being a successful student that was not addressed in the freshman seminar?" The written responses given to these questions by individual students in separate class sections could be aggregated and their content analyzed to identify recurrent themes or response categories.

* Provide some space at the end of the evaluation form so that individual instructors can add
their own questions (Seldin, 1993).
This practice enables instructors to assess specific instructional practices that are unique to their own course section. Also, this option should serve to give instructors some sense of personal control or ownership of the evaluation instrument which, in turn, may increase their motivation to use the results in a constructive fashion.

* Give students the opportunity to suggest questions which they think should be included on the evaluation form.

This opportunity could be cued by a prompt at the end of the evaluation form, such as, "Suggested Questions for Future Evaluations." This practice has three major advantages: (a) It may identify student perspectives and concerns that the evaluation form failed to address, and (b) it shows respect for student input and (c) it gives students some sense of control or ownership of the evaluation process.

Recommendations Regarding the Wording (Phrasing) of Individual Items

* Avoid vague or ambiguous descriptors which may be difficult to interpret, or whose meaning may be interpreted by different students in different ways (for example, "Pacing of class presentations was ‘appropriate’.")

* When soliciting information on the incidence or frequency of an experienced event (e.g., "How often have you seen your advisor this semester?") avoid response options that require high levels of inference on the part of the reader (e.g., "rarely"-"occasionally"-frequently"). Instead, provide options in the form of numbers or frequency counts which require less inference or interpretation by the reader (e.g., 0, 1-2, 3-4, 5 or more times).

This practice should serve to reduce the likelihood that individual students will interpret the meaning of response options in different ways.

* When asking students to rate their degree of involvement or satisfaction with a campus support service or student activity, be sure to include a zero or "not used" option. This response alternative allows a valid choice for those students who may have never experienced the service or activity in question (Astin, 1991).

* Use the singular, first-person pronoun ("I" or "me") rather than the third-person plural ("students"). For instance, "The instructor gave me effective feedback on how I could improve my academic performance" should be used rather than, "The instructor gave students effective feedback on how they could improve their performance." The rationale underlying this recommendation is that an individual student can make a valid personal judgment with respect to her own course experiences but she is not in a position to judge and report how other students, or students in general, perceive the course.

* Avoid compound sentences that ask the student to rate two different aspects of the course simultaneously (e.g., "Assignments were challenging and fair.")

This practice forces respondents to give the same rating to both aspects, even if they are more satisfied with one aspect than the other. For example, a student may feel the assignments
were very "fair," but not very "challenging."

* Include one or two negatively worded items that require students to reverse the rating scale (e.g., "I did not receive effective feedback on how I could improve my performance.").

Such items serve two purposes: (a) They encourage students to read and rate each item carefully, serving to reduce the frequency of "positive-response-set" mistakes (Arreola & Aleamoni, 1990) in which the respondent goes straight down a rating column and fills in a uniformly high rating for all items. (b) They could be used to identify evaluations forms which have not been completed carefully and may need to be deleted from the data analysis. For example, students who have responded thoughtlessly by filling in all positive or all negative ratings may be identified by their failure to reverse their response bias on the negative-worded item(s).

Recommendations for Administration of Course Evaluations

* The amount of time allotted for students to complete their evaluations should be standardized across different course sections.

Some consensus should be reached among course instructors regarding the minimal amount of time needed for students to complete evaluations in a thoughtful fashion. Further temporal standardization could be achieved if seminar instructors would agree on what time during the class period (e.g., the first or last 15 minutes of class) would be best for administering course evaluations. One argument against the common practice of having students complete their evaluations at the very end of the class period is that it could result in less carefully completed evaluations because students may be tempted to finish quickly and leave early. Another temporal factor for consideration is the time during the academic term or semester when course evaluations are administered. One option is to administer the evaluations immediately after the final exam of the course. This provides two advantages: (a) It allows students to truly assess the whole course because the final exam represents its last key component. (b) Students are not likely to be absent on the day of the final exam, so a larger and more representative sample of students would be present to complete the course evaluation than if it were administered on a routine day of class.

However, a major disadvantage of administering evaluations immediately after students complete the final exam is that students are more likely to be preoccupied, anxious, or fatigued by the just-completed exam. This may result in evaluations that are filled out more hurriedly, with fewer written comments, and less overall validity. This disadvantage is a major one and probably outweighs the advantages associated with having students evaluate the course after completing the final exam.

Perhaps the best approach is for seminar instructors to agree to administer the evaluation instrument as close to the end of the course as possible (e.g., during the last week of the term), but not immediately after the final exam. Also, this approach would better accommodate those instructors who elect not to administer a final examination in the freshman seminar.

One last consideration with respect to the issue of when course evaluations are administered is the "burn-out" or fatigue factor that may come into play when students are repeatedly required to fill out course evaluations in all their classes at the end of the term. To minimize the adverse
impact of fatigue or boredom that may accompany completion of multiple course evaluations, it might be advisable to try to administer the course-evaluation instrument to freshman seminar students at a time that does not coincide with administration of the college's standard course evaluation forms.

* **Instructions** read to students immediately before distribution of the evaluation forms should be **standardized** for all course instructors and all course sections.

  Some research has shown that student ratings can be affected by the wording of instructions that are read to students just prior to administration of the evaluation instrument (Pasen, et al., 1978). For instance, students tend to provide more favorable or lenient ratings if the instructions indicate that the evaluation results will be used for decisions about the instructor's "retention and promotion," as opposed to students being told that the results will be used for the purpose of "course improvement" or "instructional improvement" (Braskamp & Ory, 1994; Feldman, 1979). Thus, instructions read to students in different sections of the freshman seminar should be consistent (e.g., the same set of typewritten instructions read in each class).

* **Instructions** read prior to distribution of evaluation forms should effectively **prepare or prime** students for the important role they play in evaluating the freshman seminar and provide students with a **positive** "mental set".

  To increase students’ enthusiasm for course evaluation and improve the validity of the results obtained, consider including the following information in the instructions read to students prior to course evaluation.

  (a) Remind students that evaluating the course is an opportunity for them to provide meaningful input that could improve the quality of the freshman seminar for many **future generations** of first-year students.

  (b) Explain to students **why** the evaluations are being conducted (e.g., to help instructors improve their teaching and to improve the quality of the course). If items relating to specific course characteristics are to be used for instructional improvement purposes and global items for overall course-evaluation or instructor-evaluation purposes, then this distinction should be mentioned in the instructions.

  (c) Assure students that their evaluations will be **read carefully** and **taken seriously** by the program director as well as the course instructor.

  (d) Acknowledge to students that, although they may be repeatedly completing evaluations for all courses they are taking, they should still try to take the time and effort to complete the freshman-seminar evaluation form thoughtfully, because it is a non-traditional course in terms of both its content and method of instruction.

  (e) Remind students that they should avoid the temptation to give uniformly high or uniformly low ratings on every item, depending on whether they generally liked or disliked the course or the course instructor. Instead, remind them to respond to each item **independently and honestly**.

  (f) Encourage students to provide **written comments** in order to clarify or justify their numerical ratings, and emphasize that **specific comments** are especially welcome because they provide instructors with valuable feedback on course strengths and useful ideas for overcoming course weaknesses.

  (g) Inform students **what will be done with their evaluations** once they have completed them,
assuring them that their evaluations will not be seen by the instructor before grades have been
turned in (Ory, 1990), and that their hand-written comments will be converted into typewritten
form before they are returned to the instructor. (The latter assurance can alleviate student fears
that their handwriting will be recognized by the instructor; without this assurance, students may
be inhibited about writing any negative comments on the evaluation form. The need to prepare
students for their role as evaluators of course instruction, and the role of a freshman orientation
course for providing this preparatory experience, are suggested by the former president of the
Carnegie Foundation for the Advancement of Teaching, Ernest Boyer:

We urge that student evaluation be used . . . but for this to work, procedures must be well
designed and students must be well prepared. It's a mistake to ask students to fill out
a form at the end of a class without serious consideration of the process. Specifically, we urge
that a session on faculty assessment be part of freshman orientation. All incoming
students should discuss the importance of the process, and the procedures used (1991, p. 40).

* The behavior of instructors during the time when students complete their evaluations should
also be standardized.

The importance of this practice is supported by research indicating that student ratings tend
to be higher when the instructor remains in the room while students complete the course-
evaluation form (Centra, 1993; Feldman, 1989; Marsh & Dunkin, 1992). The simplest and most
direct way to eliminate this potential bias is for the instructor to be out of the room while students
complete their evaluations (Seldin, 1993). This would require someone other than the instructor
to administer the evaluations, such as a student government representative or a staff member.
Some faculty might resist this procedure, particularly if there is no precedent for it at the college.
If such resistance is extreme and widespread, even after a clear rationale has been provided, then
the following alternatives might be considered: (a) The instructor asks a student to distribute the
forms and then the instructor leaves the room while students complete their evaluations. (b) The
instructor stays in the room but does not circulate among the class while students complete their
course evaluations; instead he or she remains seated at a distance from the students (e.g., at a
desk in front of the class) until all evaluations have been completed.

Whatever the procedure used, the bottom line is that variations in how instructors behave while
students complete course evaluations should be minimized. Instructor behavior during course
evaluation is a variable which needs to be held constant so it does not unduly influence or
"contaminate" the validity of student evaluations of the freshman seminar.

Recommendations for Analyzing, Summarizing & Reporting the Results of Course Evaluations

Report both the central tendency and variability of students' course ratings.

Two key descriptive statistics can effectively summarize student ratings: (a) Mean (average)
rating per item, which summarizes the central tendency of student ratings, and (b) standard
deviation (SD) per item, which summarizes the variation or spread of student ratings for each
item. Theall and Franklin (1991) succinctly capture the meaning and significance of including
standard deviation in the analysis and summary of students’ course ratings.
The standard deviation for individual items is an index of agreement or disagreement among student raters. Perfect agreement yields a standard deviation of 0. Deviations of less than 1.0 indicate relatively good agreement in a 5-point scale. Deviations of 1.2 and higher indicate that the mean may not be a good measure of student agreement. This situation may occur when opinion in a class is strongly divided between very high and very low ratings or, possibly, is evenly dispersed across the entire response scale, resulting in a mean that does not represent a “typical” student opinion in any meaningful sense. Because students vary in their needs, a teacher [or course] may be “among the best” for some and at the same time “among the worst” for others. A mean of 3.0 or 3.5 [on a 5-point scale] cannot be construed to represent “average” performance in the sense of middle-range performance when the mean is simply an artifact of strong disagreement among students. The standard deviation is therefore an important source of information about student opinion (p. 90).

In addition to computing the means and SDs for student ratings received by individual instructors in their own course sections, these statistics can also be computed for all class sections combined, thereby allowing individual instructors to compare the mean and SD score for ratings in their own section with the composite mean and standard deviation calculated for all sections. Computing section-specific and across-section (composite) means and SDs for each item on the evaluation instrument also allow for the application of statistical tests to detect significant differences between the instructor's section-specific ratings and the average rating of all course sections combined. The results of these significance tests could provide valuable information that can be used diagnose instructional diagnosis and improvement. For instance, if an instructor's rating on an item is significantly below the collective mean for that item, it may suggest to the instructor that this is one aspect of his instruction that needs closer attention and further development. In contrast, if an instructor's mean rating on a given item is significantly above the overall mean for all course sections on that item, then this discrepancy suggests an instructional strength with respect to that particular course characteristic. What the instructor is doing to garner such a comparatively high rating might be identified and shared with other faculty who are teaching the course.

Identification and sharing of strategies for instructional improvement is an essential component of student ratings assessment, and it is a form of feedback that is frequently ignored or overlooked (Stevens, 1987; Cohen, 1990). As Stevens contends, "The instructor must learn how to design and implement alternative instructional procedures in response to feedback, which means that a coherent system of instructional resources must be easily available to the instructor. Without such a system, the instructor may be unable to gain the knowledge or support that is necessary to effect change" (1987, p. 37)

One non-threatening way to provide course instructors with specific strategies for instructional improvement is to create opportunities for instructors to share concrete teaching practices that have worked for them. Strategies could be solicited specifically for each item on the evaluation form and a compendium of item-specific strategies could then be sent to all instructors--ideally, at the same time they receive the results of their course evaluations. In this fashion, instructors are not only provided with a descriptive summary of student-evaluation results, but also with a prescriptive summary of specific strategies about what they can do to improve their instructional
performance with respect to each item on the evaluation instrument. Moreover, involving course instructors in the development and construction of these strategies serves to (a) actively engage them in the quest to improve course instruction, (b) increases their sense of ownership or control of the course, and (c) treats them like responsible agents (rather than passive pawns) in the assessment process. As Paul Dressel recommends, "Evaluation done with or for those involved in a program is psychologically more acceptable than evaluation done to them" (1976, p. 5, underlining added).

The importance of providing specific teaching-improvement feedback to course instructors is underscored by research indicating that (a) instructors prefer feedback that is specific and focused on concrete teaching behaviors (Murray, 1987; Brinko, 1993), and (b) specific feedback is more effective for helping recipients understand their evaluation results and for helping them to improve their instructional performance (Goldschmid, 1978; Brinko, 1993). As Wilson (1986) concluded following his extensive research on the effects of teaching consultation for improving instructors' course evaluations, "Items on which the greatest number of faculty showed statistically important change were those for which the suggestions were most concrete, specific and behavioral" (p. 209).

To maximize the opportunity for instructors to make instructional improvements while the course is still in progress, it is recommended that course evaluations be administered at midterm to obtain early student feedback. Cohen (1980) conducted a meta-analysis of 17 studies on the effectiveness of student-rating feedback for improving course instruction. He found that receiving feedback from student ratings during the first half of the semester was positively correlated with instructional improvement—as measured by the difference in student ratings received at midterms (before feedback was received), and ratings received at the end of the semester (after midterm feedback had been received). These findings are consistent with those reported by Murray and Smith (1989), who found that graduate teaching assistants in three different disciplines who received instructional feedback at midterms displayed higher pre-to post-test gains in student ratings than a control group of teaching assistants who did not receive midterm feedback.

Freshman seminar instructors could take advantage of this early-feedback procedure by administering student evaluations at midterms and then compare these results with those obtained at the end of the course—after some instructional change was made in response to students' midterm feedback. Thus, pre-midterm to post-midterm gain in students ratings could be attributed to the particular instructional change that was implemented during the second half of the course. This is the type of "classroom research" which has been strongly endorsed as a legitimate form of faculty scholarship (Boyer, 1991) and which integrates educational research with instructional practice (Cross & Angelo, 1988).

* To gain a reference point for interpreting student perceptions of the freshman seminar, compare student evaluations of the seminar with their evaluations of other first-semester courses.

To ensure a fair basis of comparison and a valid reference point, compare student evaluations of the freshman seminar with other courses of similar class size (e.g., a freshman course in English composition) because there is some evidence that class size can influence student ratings, with smaller classes tending to receive slightly higher average ratings than larger classes (Cashin, 1988; Feldman, 1984).
Also, depending on whether the seminar is a required or elective course, it should be compared with other first-semester courses that have the same required or elective status, because research suggests that required courses tend to receive lower student ratings than elective courses (Braskamp & Ory, 1994; Marsh & Dunkin, 1992). One institution which employed this course-comparison procedure discovered that 75% or more of their first-year students placed a higher value on the freshman seminar than they did for any other course in the college’s core curriculum (Marietta College, cited in Barefoot, 1993). As Arreola and Aleamoni advise,

> The use of comparative (normative) data in reporting results can . . . result in a more accurate and meaningful interpretation of ratings. For example, comparative data gathered on freshman-

level courses in a department allow instructors to determine how they and their courses are perceived in relation to the rest of the courses in the department. When such comparative data are not available, instructors will be interpreting and using results in a void, with very little substantiation for their conclusions and actions (1990, p. 51).

Surveys or questionnaires could also be used to obtain a different type of comparative perspective on the freshman seminar—the retrospective perceptions of alumni. Since the course emphasizes lifelong-learning and life-adjustment skills, it might be revealing to assess how alumni, looking back on the freshman seminar from the perspective of a working professional or graduate student, would respond to the following questions posed to them via phone interviews or alumni surveys: (a) Do you view the seminar differently now than you did when you were a first-year student? (b) What aspect of the seminar is most memorable or has had the most long-lasting impact on you? (c) Do you still use any ideas or skills acquired during the freshman seminar in your professional or personal life?

It is noteworthy that the only two reported studies of alumni perspectives on the freshman seminar have both revealed that former students' retrospective evaluations of the course are very positive (Hartman, et al., 1991; University of Prince Edward, cited in Barefoot, 1993).
References


the community college. In J. N. Hankin (Ed.), *The community college: Opportunity and access for America's first-year students* (pp. 59-68). The National Resource Center for The Freshman Year Experience & Students in Transition, University of South Carolina.


Marsh, H. W., & Ware, J. E., Jr. (1982). Effects of expressiveness, content coverage and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology*, 74(1), 126-134.


Addendum
Campus-specific Evidence Supporting the Positive Impact of the First-Year Seminar on Student Retention & Academic Achievement

◆ STUDENT RETENTION OUTCOMES
Based on her 10-plus years of experience reviewing research studies on the first-year seminar as Co-Director for Research & Publications at the University of South Carolina’s National Resource Center for The First-Year Experience, Betsy Barefoot (2000) recently concluded, “We are building a body of research that seems to indicate that yes, first-year seminars are positively correlated with improved student retention” (p. 1). Listed in this subsection of empirical evidence is a selected sample of campus-specific studies supporting Barefoot’s conclusion.

1. Persistence to Completion of the First Semester/Quarter of College
Research conducted at Sacramento City College revealed that students who participated in the first-year seminar persisted to completion of the first term at a rate that 50% higher than non-participants (Stupka, in Barefoot, 1993a).

2. Persistence to Completion of the First Year of College
Research conducted at Miami-Dade Community College has shown that course participants display a 67% first-year retention rate, compared to a rate of 46% for non-participants (Belcher, in Barefoot, 1993). At the University of South Carolina, a series of separate studies of first-year student cohorts enrolled in University 101 (first-year seminar) revealed that, for 16 consecutive years, students who took the seminar were more likely to persist to the sophomore year than first-year students who did not take the course. In 11 of the 16 years, these differences reached statistically significant levels, despite the fact that course participants had higher course loads and lower predicted academic potential—as measured by standardized-admissions test scores (Fidler, 1991).

At Ramapo College (New Jersey), a “time-series” research design has been employed to demonstrate that the average freshman-to-sophomore retention rate for cohorts of entering students who participated in the first-year seminar during a five-year period immediately after the course became a requirement was significantly higher than the average retention rate for freshmen who had entered the college during the three-year period immediately before course adoption (Starke, Harth, & Sirianni, 2001).
At Widener University (PA), freshman seminar participants returned for their sophomore year at a rate that was approximately 18% higher (87.3% vs. 69.6%) than their expected return rate—as predicted by entering SAT scores (Bushko, 1995).

3. Persistence to Completion of the Sophomore Year
   At the University of Maryland, an “experimental” research design was employed whereby students who were randomly selected to participate in the first-year seminar displayed significantly higher retention rates throughout their first four semesters in college compared to a randomly selected control group (Strumpf & Hunt, 1993).

4. Cumulative (Total) Number of College Units/Credits Completed
   Research conducted at Sacramento City College revealed that course participants completed 326% more units than non-participants (Stupka, in Barefoot, 1993a). At Oakton Community College (IL), course participants went on to complete 39 units, whereas non-participants completed 26 (Deutch, in Barefoot et al., 1998).

5. Persistence to Degree/Program Completion
   Research conducted at the University of Prince Edward Island in Canada, demonstrated that 49% of course participants persist to completion of the baccalaureate degree, versus 28% of non-participants (Robb, in Barefoot, 1993a). At Ohio University, 4-, 5- and 6-year graduation rates were higher for course participants than non-participants (Chapman & Kahrig, in Barefoot et al., 1998). At Dalton College (a 2-year commuter campus in the University of Georgia system), institutional researchers tracked students over a 5-year period and found that 30.8% of course participants met the 90 quarter-hour requirement for graduation, as compared to 19.4% of non-participants (Hoff, Cook, & Price, 1996).

6. Time Taken to Degree/Program Completion
   Research at Keene State College (New Hampshire) has shown that 29% of course participants graduate within four years, versus 16% of non-participants; and 52% of course participants graduate within 5-1/2 years, compared to 35% for non-participants (Backes, in Barefoot et al., 1998).

◆ ACADEMIC PERFORMANCE/ACHIEVEMENT OUTCOMES
   Evidence for the positive impact of first-year seminars on students’ academic performance is not as consistent as it is for student retention (Barefoot, 2000). Nevertheless, there have been a number of campus-specific studies that indicate that student participation in the seminar is associated with improved academic performance—as measured by different academic-achievement indicators, such as the seven highlighted in the following section.

1. Cumulative GPA Attained at the End of the First Term or First Year of College
   Research conducted at Genesee Community College (NY) revealed that course participants earned a first-term GPA of 2.87 relative to a matched control group of non-participants who earned a 2.38 GPA (Wahlstrom, in Barefoot, 1993a). At Northern Illinois University, five consecutive first-year cohorts were compared with a matched group of non-
participants, and it was found that students who took the course earned significantly higher first-term and first-year GPAs (House, in Barefoot et al., 1998).

2. Cumulative GPA Attained Beyond the First Year of College
   At Indiana University of Pennsylvania, an “experimental” research design was used in which high-risk students were randomly assigned to either register or not register for the first-year seminar. Students who successfully completed the first-year seminar achieved significantly higher GPAs (p<.01) over a 3-year period than the control group of students who did not take the course (Wilkie & Kuckuck, 1989).

3. GPA Attained vs. GPA Predicted
   At Indiana University of Pennsylvania, the predicted GPAs of randomly-selected course participants and non-participants were found to be equal at the outset of college, yet the cumulative GPAs attained by course participants at the end of their first, second, and third years of college were significantly higher than the GPAs of students who did not take the course (Wilkie & Kuckuck, 1989).

4. Total Number of First-Year Students in Good Academic Standing (i.e., Students Not Placed on Academic Probation or Academically Dismis
   At the University of Maryland, research has shown that the number of course participants who completed the first two years of college in good academic standing was significantly higher than it was for students who did not participate in the course (Strumpf & Hunt, 1993). At Northern Michigan University, significantly higher percentages of first-year seminar participants than non-participants maintained good academic standing (GPA of at least 2.0) over their first five semesters in college (Soldner, in Barefoot et al., 1998). After Averett College (a small liberal arts college in Virginia), adopted a required first-year seminar—without changing enrollment patterns or embarking on any other major retention efforts—the percentage of freshmen who ended their first year in academic difficulty dropped 20% (Vinson, 1993).

5. Total Number of First-Year Courses Passed (i.e., Not Dropped or Failed)
   Research conducted by a consortium of four community colleges in North Carolina revealed that course participants completed an average of nine more units by the end of their first year of college than did non-participants (Garret, in Barefoot, 1993a).

6. Total Number of First-Year Courses Completed with a Grade of “C” or Higher
   Research conducted at Sacramento City College revealed that course participants completed four times as many math classes, three times as many writing classes, and twice as many reading classes with a grade of “C” or higher than students who did not participate in the course (Stupka, in Barefoot, 1993a).

7. Percentage of Students Who Qualify for the Dean’s List and Honors Program
   At the University of Vermont (where the first-year seminar is taught as an introduction to the liberal arts and sciences, with an emphasis on critical/creative thinking, research skills, and oral/written communication skills), the percentage of students who made the Dean’s List and
were accepted into the school’s honor program was significantly higher among course participants than non-participants (Thomson, in Barefoot et al., 1998).

References


Columbia University, New York.


