Program quality in higher education — what it is, how to identify it, and how to foster it — has been an enduring concern among educators since the founding of Harvard College in 1636. Today, a variety of social forces has made program quality perhaps the single most important issue in higher education. Public disenchantment with educational institutions, pressures for accountability, and declining resources are but three factors that have precipitated a renewed interest in quality. State systems of higher education often must decide which programs to retain in the face of declining enrollments and limited resources. Many private liberal arts colleges must demonstrate for their very survival that their high costs are linked to a quality education. Local tax revolts are forcing many community colleges to decide which programs to eliminate. The federal government and private foundations need to identify quality programs before investing their limited resources.

Much recent literature has focused on the assessment of program quality. Aside from program evaluations at individual institutions, which are not reviewed here, program evaluation research has been aimed primarily at assessing and ranking departments, professional schools, and sometimes entire colleges and universities. In these evaluation studies, quality has been assessed either through a "reputational" approach or through an approach based on "objective" indicators of quality. More recently, there has emerged a third approach to research on program quality. Instead of evaluating program quality, this line of research has focused on the identification of quantifiable program characteristics associated with programs considered to be of high quality.

The major purpose of this article is to review and critique these lines of research and to examine the implications of each for future research. The paper is
divided into four major sections. In order to provide a context for examining
the three areas of research, the first section looks at the multiple meanings of
quality. The second section examines the program evaluation research that is
based on a "reputational" approach; the third reviews the program evaluation
research based on "objective" indicators; and the fourth looks at research on
the quantitative correlates of program quality. In each section the pertinent lit-
erature and major findings are reviewed, the major criteria and methodological
procedures examined, and a critique of the limitations and strengths of the ap-
proach provides the foundation for some recommendations for future
research.

THE MEANINGS OF QUALITY

Quality . . . you know what it is, yet you don't know what it is. But that's self-
contradictory. But some things are better than others. That is they have more quality.
But when you try to say what that quality is, apart from the things that have it, it all
goes poof! There's nothing to talk about. But if you can't say what quality is, how do
you know that it even exists? If no one knows what it is, then for all practical pur-
poses it doesn't exist at all. But for all practical purposes it does exist. What else are the
grades based upon? Why else would people pay fortunes for some things and throw
others in the trash pile? Obviously, some things are better than others . . . but what's
the "betterness"? . . . So round and round you go, spinning mental wheels and no-
where finding any place to get traction. What the hell is Quality? What is it? [Pirsig,
1974, p. 179]

As Pirsig concluded, the concept of quality is as elusive as it is pervasive. The
difficulty of definition makes it not unlike obtaining general consensus on what
is pornographic. When Justice Potter Stewart remarked, "I can't define it, but
I know it when I see it," he in part was admitting that it would be impossible to
obtain universal agreement on the concept.

While universal agreement on the concept of quality is impossible, we can il-
luminate the concept even while acknowledging that quality has multiple mean-
ings and that no single definition will be acceptable to all people. It is particu-
larly useful to examine the ways the term is used in different cultures and in dif-
ferent contexts. Such an examination brings to the fore the often unstated as-
sumptions about the meaning of quality that underpin studies of program qual-
ity in American higher education.¹

As a point of departure, we can say that quality seems to have a national
(cultural) dimension to it. While there are many cultural similarities between
our country and other nations, we tend to set the term within the context of our
own cultural growth and development. Our belief in democracy contains within
it the goal of equality, as Bonham (1977, p. 6) has noted:

We still tend to believe that we can be equal and superior [highest quality] at the same
time, that the potential for human growth is bounded only by the given opportunity
and of course access thereto. . . . We cannot all hit home runs, but we all deserve a
chance at bat. . . . When we seek excellence [quality] it must be in the context of a
concern for all.
The American concern for excellence (highest quality) is also associated with
the process of achievement, the effort the individual makes to accomplish
goals. Furthermore, the scope of excellence is large. Rogers (1981) puts it this
way:

There is the excellence of a beautifully crafted silver bas-relief on a delicate drinking
vessel — and there is the excellence of a well crafted critical essay on the structure and
meaning of a great work of literature.

There is the excellence of a master training a wild horse and eventually riding in pre-
cise dressage. And there is the excellence of mastery of rules of evidence by which an
intelligent and interesting argument can be made both truthful and convincing.

There is the excellence of a finely tuned instrument played with both technical compe-
tence and deep feeling. And there is the excellence of an elegantly constructed research
design in mathematics or the sciences.

In every case there are those hard-won elements of discipline, refinement, quest for
constant improvement, extension to the limit — the final blending of attention to
every detail and to the attainment of beauty in the whole.

McClelland (1961, p. 144) also places excellence in a national context:
Our national problem is that we have tended to focus increasingly on encouraging one
type of excellence, and a practical, measurable action oriented type of excellence at
that. Other types of human excellence exist, particularly those involving character,
and inner life, ... the imagination and human sensitivity.

As these quotations suggest, the concept of quality in this country clearly has
multiple dimensions and can be seen in many contexts. As we develop this dia-
lectic on “quality” and focus more sharply on its meaning in higher education,
it will be well to remember that even within this limited sphere the term is likely
to have different meanings within college and university subcultures. Program
quality is not likely to be the same in a highly rated research university as it is in
a state college or a selective liberal arts college or an open-admissions four-year
institution or an urban community college.

Before illuminating the meanings of quality by contrasting it with excellence
— and other concepts such as goal accomplishment, efficiency, accountability,
and effectiveness — it is useful to examine some of the major elements or ingre-
dients of quality as it is commonly defined in our colleges and universities. By
itself, no single dimension will be a sufficient condition of quality, and none
will even be a necessary condition in all people’s eyes and in all contexts. Each,
however, appears at one time or another in someone’s meaning of quality.

Quality programs are almost always related to characteristics of the faculty
responsible for the implementation of the curriculum. In fact, as will be seen
below, studies of program quality frequently seem to be little more than efforts
to ascertain *faculty* quality. Others, including ourselves — especially as one
moves from highly rated research universities to consider other types of colleges
and universities — consider facilities, support, the curriculum, and student
attributes as dimensions of quality. A program of high quality is presumed to
have the facilities necessary for its success such as well-equipped laboratories,
appropriate library holdings, computers, and all the material things needed for the desired learning to take place. A quality program is one with adequate support — the resources to maintain the operation, provide for faculty travel, and attract and retain outstanding faculty. The curriculum of a high-quality program is one that has all of the essential ingredients — such as variety and depth of courses. And last, the students exist in sufficient numbers so that there is an adequate mix to foster students' learning from one another and yet not so many students that individualized attention is lost.

In addition to these frequently appearing elements of quality, some individuals identify other, less quantifiable attributes of a quality program: leadership, esprit, morale, clarity of purpose, and a healthy organizational climate. For many people a quality program cannot be reduced to a set of quantitative indicators, and is more like a living organism than a collection of component parts.

While individuals will nod their heads affirmatively when such words as "esprit" are introduced into a discussion of program quality, it is usually not clear that all participants attach the same meaning to the terms. Seldom are such concepts operationalized, indicators for them agreed upon, and evidence collected so as to ascertain the extent of their existence and their degree of relationship to the more common ingredients of quality. This is not to say that such components as morale and leadership are not a part of quality; rather, one finds that as yet they have not been especially helpful in attempts to determine the relative quality of programs.

Another set of terms sometimes advanced in discussions of quality are similar to the above terms in their nebulousness, but they are somewhat different in kind. They tend to address more personal dimensions, like achievement, persistence, purpose, worth, beauty, meritoriousness, and character. These terms are implied in the introductory quotations. They carry distinct values related to American ideals. No doubt they are a part of many people's notion of quality. That they also introduce complications when it comes to their assessment does not mean that they are not important. In fact, it is because of this richness of the concept of quality that Kuh (1981) argues for taking a holistic approach to the assessment of quality. He believes that breaking out components that are measurable vitiates the essence of the concept itself.

The meaning of quality can also be illuminated by comparing it with other concepts that are used in discussions of programs. Four frequently employed terms are accountability, efficiency, effectiveness, and excellence. "Accountable" implies that a program is at least adequate: a program that meets some minimum set of standards and achieves its goals is said to be an accountable one. "Quality," on the other hand, suggests that essentially complete goal fulfillment is attained, not just an acceptable amount. Thus quality encompasses accountability. In addition, demonstrating the accountability of a program frequently implies legitimizing its existence for others: if a program can do so, then it has justified itself.
“Efficiency,” like “accountability,” carries an economic overtone. Once goals to be achieved are established, alternative programs for attaining them are advanced. The superior program is the one which is most efficient. The efficient program is one which will accomplish the same ends as another program but with less cost or in less time, or both. A quality program will more likely be efficient than inefficient. In fact, parsimony is an attribute of high value for those who strive for quality. Persons advocating quality will place efficiency farther down on their list of critical elements for judging what is best. They expect efficiency, but they ask for more than that when they insist on quality.

As Cameron and Whetten (1983) make clear, investigators of organizational effectiveness do not agree as to the meaning of the term in the large body of studies that have been conducted under its name. The reasons for the differences are many, just as they are for the meaning of quality. And because they are multiple, it is all but impossible to contrast each of the meanings of effectiveness with each of the meanings of quality. But some contrasts can be made. In judging program effectiveness in contrast to its quality, the former is likely to concentrate on how well the outcomes are achieved. Those assessing relative quality, while paying attention to outcomes, will look beyond the structure of a program; they will also be examining process. For example, one would look at the environment and the amount and kind of student/faculty interaction. Pace (1980) describes the day-to-day operation of an academic program along a number of scales called “quality of student effort” and means by this phrase something more than effectiveness. Effectiveness assumes efficiency but does not include all that is meant by quality.

When it comes to the concepts of excellence and quality, the line of demarcation becomes more blurred. While John Gardner (1961) was not the first person to address excellence, his book set a tone that is still felt today. Our opening quotations found the two terms essentially interchangeable. Like quality, excellence carries with it a dimension of style, not just an outstanding accomplishment but the manner in which the accomplishment is achieved. Both excellence and quality imply the highest standards and an unwillingness to settle for anything less than that which could be achieved. “Excellence” may even be the more egalitarian term, for frequently — mistakenly, we believe — “quality” is identified with a snobbish elitism, especially in higher education. (Writers in the field invariably talk about the top-ranked graduate schools as the elite, the institutions with “high quality” programs, as if quality does not exist in other settings.)

Nonetheless, a minor distinction can be noted between the two terms. In current usage “excellence” is used more frequently in connection with the acts of an individual, whereas “quality” is the label that is placed on the collectivity, such as an organization or a program. Still, the ingredients for both are essentially identical. We have opted for the term “quality” because it pervades the literature reviewed below. In addition, despite its ambiguities, the term is not likely to fall from usage. “Quality” carries too great an emotional charge or
subjective overtone to be grounded by our dialectical analysis. That said, we proceed with a review of the literature on program quality in higher education.

**REPUTATIONAL STUDIES**

Reputational studies have long dominated research on program quality in higher education. Built upon a foundation of peer evaluation in which experts in the field are considered to be the best judges of program quality, these studies have been criticized as severely by some as they have been advanced by others. Notwithstanding the controversy surrounding the approach, much of the best-known research on program quality has been in this tradition.

The most salient characteristic of reputational studies is the emphasis placed on peer evaluation, in which judgments of program quality are made by faculty raters or by other experts such as department chairpersons and deans. In general, reputational studies follow a similar procedure: the researcher selects one or more criteria (such as faculty quality) to serve as a basis for evaluation, employs a panel of experts to rate individual programs in terms of those criteria, and then combines individual panelists' responses in order to generate a ranking of programs by institution. Since most reputational studies have evaluated programs at the doctoral level, we turn first to a review of studies at the graduate level before examining the literature on professional schools and undergraduate programs.

The first major reputational study of graduate programs was conducted by Hughes in 1924. Using panels of scholars in each discipline to identify top scholars in the field, Hughes (1925) ranked 38 universities in 20 graduate fields according to the number of top scholars they employed. A second Hughes (1934) study rated 59 universities in 35 fields according to judgments of staff and facilities made by faculty experts. The two studies by Hughes established important precedents for later studies of program quality: a focus on the graduate (doctoral) level, reliance on ratings by academics rather than outside observers, and an emphasis on the nation's most visible institutions (Lawrence and Green, 1980, p. 4). No less important, Hughes's studies set the precedent that the major criterion for evaluating quality should be the quality of the faculty.

A quarter of a century after the second Hughes study, Keniston (1959) selected department chairmen from institutional members of the American Association of Universities to rank 24 graduate programs at the 25 institutions leading in doctorate production. In addition to ranking programs by institution, Keniston rank-ordered the top 20 institutions. His study, as well as the two earlier Hughes studies, was severely criticized by researchers on methodological grounds. These criticisms, which included complaints about geographical and rater bias, led to subsequent reputational studies aimed at updating and improving ratings of graduate programs.

The most methodologically sophisticated of the reputational studies were designed under the auspices of the American Council of Education (ACE), which
sponsored ratings of the doctoral programs in the arts and sciences in 1964 and 1969 (Cartter, 1966; Roose and Anderson, 1970). The Cartter study yielded institutional rankings in 29 disciplines at a total of 106 universities, while the Roose-Anderson study ranked programs in 36 disciplines at 130 universities.

In both ACE surveys, ratings within each discipline were made by department chairpersons, distinguished senior scholars, and selected junior scholars. Each respondent rated doctoral programs in his or her own field in terms of two criteria: quality of the graduate faculty and effectiveness of the graduate program. The individual ratings were then averaged, and the leading programs in each field were rank-ordered in two separate lists, one based on faculty quality and the other on program effectiveness. The two sets of rankings in both the Cartter and Roose-Anderson studies corresponded closely.

Several scholars have rightly noted that the methodological procedures in the ACE studies gave careful attention to the statistical validity and reliability of the ratings (Clark, 1976; Lawrence and Green, 1980). Cartter (1966) used detailed analyses to show that the rankings were similar whether respondents were separated by rank, institution of employment, degree, or geographical region. Moreover, he showed consistency between the ACE ratings and ratings by smaller and more select panels in several fields.

Since 1970, three major reputational studies of doctoral program quality have been published. One study, conducted under the joint sponsorship of the Council of Graduate Schools (CGS) and the Educational Testing Service (ETS), used faculty raters to assess and rank graduate programs (Hartnett, Clark, and Baird, 1978). As in the ACE ratings, programs were rated in terms of faculty quality and program effectiveness. Another study, by Ladd and Lipset, asked more than 4,000 faculty members at four-year colleges and universities to name the five departments nationally in their disciplines that had the most distinguished faculties (Scully, 1979). Ladd and Lipset then ranked the departments that were rated among the top five by at least 10% of the respondents in each of 19 fields.

The most recent reputational study was sponsored by the Conference Board of Associated Research Councils (1982) and compared the quality of research-doctorate programs in 32 disciplines from 228 universities. The multidimensional approach used to assess graduate programs employed 16 separate measures "related to quality," including a reputational survey in which faculty raters were asked to rate faculty scholarly competence, program effectiveness, and program improvement over the last five years. The results were presented in raw scores and standardized scores showing how far above or below the mean each program was ranked; no single composite measure of the quality of graduate programs was developed. While there were few surprises in most of the ratings, the ratings of programs that had improved in quality over the past five years identified institutions that had not been ranked in previous reputational studies.

Comparison of the rankings of graduate programs across all of the studies
discussed above reveals a consistent pattern of findings: although there are minor variations in the absolute rankings of programs, reputational assessments of graduate program quality have consistently identified the same graduate programs at the top of the rankings. In terms of departments with middle and lower rankings, the absolute rankings vary considerably but still show a high degree of stability across broad groups of departments.

Although most reputational researchers discourage the aggregation of departmental rankings into institutional rankings, many have felt compelled to make such aggregations (American Council on Education, 1971; Magoun, 1966; Morgan, Kearney, and Regents, 1976; National Science Board, 1969; Petrowski, Brown, and Duffy, 1973). When universities with high overall standings in the Ladd-Lipset and ACE surveys are compared with the overall rankings of Hughes (1925) and Keniston (1959), the stability of university reputations is clearly seen: all but one of the top 20 institutions are still at the top, and less than a handful of new ones has been added. The remarkable stability in the rankings is supported by the fact that six institutions (California-Berkeley, Chicago, Harvard, Michigan, Princeton, and Yale) have been ranked among the top ten institutions in all six of the major reputational studies of graduate education (Cartter, 1966; Conference Board of Associated Research Councils, 1982; Hughes, 1925; Keniston, 1959; Ladd and Lipset in Scully, 1979; Roose and Anderson, 1970).

In summary, the reputational method has served as the major way of evaluating program quality at the graduate level. Reputational studies, which place major emphasis on faculty quality as judged primarily by faculty raters, have yielded highly stable ratings of top-ranked graduate programs and institutions.

While most studies of program quality have examined graduate departments in the arts and sciences, recent interest in professional programs has led a few researchers to apply the reputational method to assessing professional program quality. Two fairly recent studies evaluated programs in at least three professional fields, and several other studies (conducted both by academics and by the professions themselves) have focused on a single field.

In a widely publicized study of professional education, Margulies and Blau (1973) rated programs in 17 professional fields. Deans of professional schools served as raters and were asked to identify the top five programs in their field. The responses were then aggregated and programs were ranked, by institution, for all 17 fields. Criticized for the low overall response rate to their survey, Blau and Margulies (1974-1975) subsequently completed a replication study using the same fields plus one more. Even though the response rate increased dramatically in their follow-up study, the rankings remained about the same in all fields.

Another study of professional programs rated programs in law, education, and business (Cartter and Solmon, 1977). Unlike the earlier study by Blau and Margulies, Carter and Solmon used faculty members as well as deans as raters, had raters evaluate faculty and program attractiveness on Likert-type
scales rather than simply select the top five programs, and provided raters with
a list of institutions to evaluate rather than asking them to name, by recall, top
programs in their field. Despite these methodological differences, comparisons
between the Blau-Margulies and Cartter-Solmon rankings in law, education,
and business reveal many similarities. There are, however, some notable differ­
ences, which may be due primarily to differences in sample size and selection
and in the survey instruments used (Munson and Nelson, 1977).

Other reputational studies of professional education have ranked programs
in a single field. For example, Cole and Lipton (1977) ranked top medical
schools; Carpenter and Carpenter (1970) ranked library science programs;
the top 15 business programs in two reputational studies; and Juris Doctor
magazine ("The Popular Vote," 1976) ranked the top 20 law schools.

Two important observations can be made about reputational studies of pro­
fessional programs. First, while they have been influenced by earlier reputa­
tional studies at the graduate level (particularly the methodology of the 1966
Cartter study), they often have utilized more diverse groups of raters and more
diverse rating criteria. Second, as Lawrence and Green (1980, p. 22) have re­
ported, most reputational rankings have consistently identified the same pro­
fessional schools at the top, though there is some variance in the rankings of in­
stitutions across studies. Three rankings of law schools (Blau and Margulies,
1974–1975; Cartter and Solmon, 1977; "The Popular Vote," 1976) have in
common seven institutions among those at the top of the rankings; three rank­
ings of business schools (Blau and Margulies, 1974–1975; Cartter and Solmon,
1977; "The Top 15," 1975) share six institutions at the top; and two rankings of
medical schools (Blau and Margulies, 1974–1975; Cole and Lipton, 1977) have
eight institutions in common among those at the top.

In the last few years, there have been only two reputational studies of pro­
gram quality at the undergraduate level. One study, sponsored by Change
magazine, identified ten "leading" institutions in terms of national influence
(Johnson, 1978). The list of ten institutions is very similar to the top-rated in­
stitutions as identified in the aforementioned reputational studies of graduate
programs (see Petrowski, Brown, and Duffy, 1973).

A second, more recent study (Solmon and Astin, 1981; Astin and Solmon,
1981) also rated undergraduate programs using a reputational approach. Sol­
mon and Astin asked faculty raters to rank departments in their field of spe­
cialization (from a list of about 100 institutions) on the basis of six rating cri­
teria. After faculty in seven departmental specializations had ranked
departments according to each of the six rating criteria, the researchers
identified the ten top-ranked institutions on each list. The final list of top­
ranked undergraduate departments in each of the seven fields included those
departments which had been ranked in the top ten according to at least one
criterion.

While the Solmon and Astin lists of top-ranked institutions included most of
the institutions identified in earlier reputational studies of graduate education, the researchers identified a fairly large number of institutions that had not appeared in those studies. This finding may be especially significant, for it suggests that broadening the rating criteria to include dimensions such as "faculty commitment to undergraduate teaching" leads to the identification of programs of excellence which are otherwise ignored when the traditional criterion of faculty quality is the main criterion for evaluation.

Astin and Solmon (1981) admit that their reputational study is only preliminary, given the relatively small samples of departments and institutions. Still, they have extended the assessment of quality beyond faculty quality and found that diversifying rating criteria can lead to the identification of "quality" programs that would otherwise be overlooked.

Critique

Given the visibility of most of the reputational studies, it is hardly surprising that they have received a stream of criticism. Most of this criticism has been on methodological rather than substantive issues, and some of it has been unfounded. Our intent here is to examine the most valid criticisms and major strengths of reputational studies, as well as to make some suggestions regarding future research. We begin by discussing two methodological weaknesses of reputational studies.

First, reputational studies can be justly criticized for rater bias, which can take several forms. For example, the overall reputation of an institution may influence raters' assessments of a particular department, especially when raters are not well informed. Such "halo effects" have been found by researchers at both the undergraduate and graduate level (see Solmon and Astin, 1981, p. 27). Also, departmental or institutional reputation may lag several years behind current quality, thereby introducing another source of rater bias (Hughes, 1925; Cox and Catt, 1977). In addition, scholars from the highly ranked universities often are overrepresented as raters and may tend to rate higher those departments with similar emphases as their own (Webster, 1981, p. 21). Other sources of rater bias — such as "alumni effects" (the tendency of raters to highly rate their alma maters) and the influence of an institution's age or size on raters' perceptions of quality — have been noted in the literature (Lawrence and Green, 1980, p. 10). However, there is insufficient evidence to conclude that these sources affect raters' assessments of program quality.

A second, and related, methodological criticism of reputational studies is that many raters are not sufficiently well informed to make judgments about the quality of programs at other institutions, particularly when the evaluative criteria extend beyond faculty quality. Since most reputational studies have ranked programs presumably on the criterion of faculty scholarly achievement, this criticism often has been rejected on the grounds that who but faculty are most competent to judge their peers (Blackburn and Lingenfelter, 1973, p. 24). While this rejoinder has some merit, it overlooks the fact that many faculty are
relatively uninformed about faculty at other institutions, particularly at less
visible universities. Moreover, if reputational studies are to evaluate something
more than faculty quality, raters should be selected who are well informed
about programs in terms of all the criteria serving as the basis for evaluation.

At least five other criticisms can be directed against most reputational
studies. First, because the primary emphasis is on ranking programs rather
than evaluating them against a normative standard of quality, reputational
studies do not inform us whether a particular program is of high, medium, or
low quality; they tell us only its relative quality vis-à-vis those programs to
which it is compared.

Second, the criteria used to assess quality in most reputational studies are
severely limited. Most studies at the graduate level have relied on a single yard-
stick to assess quality: faculty quality. As a consequence, most reputational
studies are peer judgments of departmental faculty quality, presumably based
almost entirely on scholarly publication. Given the use of a single criterion, it is
hardly surprising that the same institutions are consistently ranked at the top.
By reifying the criterion of faculty quality, most reputational studies have ig-
nored a number of other potential evaluative criteria, most notably teaching
and program effectiveness. As stated in one review of reputational studies at
the graduate level: “They say little or nothing about the quality of instruction,
the degree of civility or humaneness, the degree to which scholarly excitement is
nurtured by student-faculty interactions, and so on” (Hartnett, Clark, and
Baird, 1978, p. 1311). Faculty quality alone is an insufficient criterion for as-
sessing overall program effectiveness — whether at the graduate, professional,
or undergraduate level.

Third, because most reputational studies have used the department as the
unit of analysis, the institutional environment outside of the department is rare-
ly considered in the evaluation. Especially at the undergraduate level, where
students are exposed to a variety of fields and extracurricular influences, quali-
ty assessment should take the institutional environment into account. Indeed,
the argument has been forcefully made in terms of the graduate level: “Quality
is a property of the total institution . . . [and] a component of high quality
can be expected to flourish best within an environment of high quality” (Na-
tional Science Board, 1969, p. 50).

Fourth, since they have focused exclusively on ranking the “top” 20 to 150
institutions, all of the major reputational studies have failed to consider pro-
grams at most of the postsecondary institutions in the nation. Graduate, pro-
fessional, and undergraduate programs in state and regional colleges and uni-
versities have been ignored.

Fifth, while reputational studies at the graduate level have received consid-
erable attention, relatively little attention has been given to professional schools
and undergraduate programs. To be sure, there have been a fair number of
studies of professional schools in recent years, but with only two exceptions
(Blau and Margulies, 1974–1975; Cartter and Solmon, 1977) all of these
studies have been limited to a single field. At the undergraduate level, there have been only two major studies of program quality.

Notwithstanding these criticisms, reputational studies have contributed to the understanding of program quality in American higher education. Reputational studies at the graduate level have provided defensible ratings of "leading" departments in terms of faculty scholarship (Blackburn and Lingenfelter, 1973, p. 25). At the professional school level, reputational studies have consistently identified most of the same "top" programs while using multiple criteria for evaluation. And at the undergraduate level, some progress has been made in assessing program quality based on multiple criteria.

Over a decade ago, the controversy over the last of the ACE-sponsored graduate-level studies (Roose and Anderson, 1970) prompted the American Council on Education to discontinue sole sponsorship of any further reputational studies. Since then, with the exception of the recent study sponsored by the Conference Board of Associated Research Councils (1982), most reputational studies have been conducted at the professional school and undergraduate level. In our view, controversy over reputational studies should not deter researchers from conducting such studies in the future. If reputational studies are designed to respond to the criticisms raised here, they can make a more important contribution to evaluating quality in higher education. Especially if methodological refinements are made, if quality is evaluated through multiple criteria, if normative standards are used, and if quality is evaluated not just at "leading" schools, then future reputational studies may help to shed further light on program quality at the undergraduate and professional as well as the graduate level.

STUDIES BASED ON OBJECTIVE INDICATORS OF QUALITY

Although reputational studies have dominated research on program quality, criticism of the reputational approach has prompted some researchers to explore more "objective," quantifiable procedures for assessing program quality in higher education. These researchers have sought to "quantify" quality in order to measure it empirically.

Quantitative studies of program quality generally follow a straightforward format: the evaluator selects criteria (such as faculty scholarship) along with objective indicators (such as average number of articles per faculty member for each of the last five years) on an a priori basis, develops an index of those indicators, and then rates programs on the basis of that index. With regard to the identification of evaluative criteria and objective indicators, a few researchers have designed systematic studies to identify criteria and indicators by surveying selected respondents (Barak, 1976; Clark, Hartnett, and Baird, 1976; Clewell, 1980; Fotheringham, 1978; Longanecker, 1978; Lynd, 1976). Most researchers, however, seem to have relied on their own insight and judgment to develop quality indicators or have borrowed indicators from accreditation or fugitive institutional or statewide reports on program quality.
Owing to the lack of agreement among researchers regarding appropriate indicators, a wide range of criteria and indicators have been used to assess and rank graduate programs and undergraduate colleges and, in a few cases, provide composite institutional rankings (Adams and Krislov, 1978; Bowker, 1964; Calvert, Pitts, and Dorion, 1971; Clemente and Sturgis, 1974; Conference Board of Associated Research Councils, 1982; Cox and Catt, 1977; Eells, 1960; Glenn and Villemez, 1970; Hurlbert, 1976, Knudsen and Vaughan, 1969; Krause and Krause, 1970; Lewis, 1968; Siebring, 1969; Somit and Tanenhaus, 1964). In some studies only one or two indicators have been used to assess quality, but in the majority, multiple criteria and multiple indicators have been used.

One of the first major quantitative studies of program quality at the graduate level, conducted by Bowker (1964) two decades ago, illustrates the general procedure followed by researchers using this approach to examining program quality. In order to develop a ranking of U.S. graduate schools in the social sciences—humanities and the sciences, Bowker developed a quantitative index which included objective indicators in each of the two major fields. In the social sciences—humanities, for example, he used four quantifiable indicators of quality: number of former Woodrow Wilson fellows on the faculty, number of Woodrow Wilson fellows choosing to attend the institution, number of American Council of Learned Societies Award winners on the faculty, and the number of Guggenheim Memorial fellows on the faculty.

Since Bowker's study, most of the research based on objective indicators of quality has focused on the evaluation and ranking of graduate programs in particular disciplines, with graduate programs in sociology and psychology receiving the most attention. Numerous rankings of graduate departmental quality have been published, but many of these rank different programs and thus are not comparable to one another. In those cases where rankings can be compared, there is a pattern: although the absolute ranks of graduate programs vary (some considerably, others hardly at all) from one study to the next, objective assessments have consistently identified the same graduate programs at the top of the rankings. To be sure, the high degree of consensus may be attributed to heavy reliance on the criterion of faculty research productivity and research-oriented indicators, and the fact that researchers have been concerned only with the ranking of a relatively small number of institutions (usually 20 to 80). Still, the relatively high agreement among the rankings is noteworthy, since most of the studies have been based on more than one criterion and multiple, frequently different, indicators of quality.

The debate over rankings of graduate programs and institutions has prompted some scholars to compare reputational rankings with rankings based on objective indicators of quality. After ranking graduate programs and institutions based on objective measures, these researchers have compared their rankings with the ACE-sponsored rankings by Cartter (1966) and Roose-Anderson (1970). A consistent pattern emerges across these comparisons: there is a close
correspondence between objective and reputational rankings, although the absolute ranks of particular programs usually vary modestly (Adams and Krislov, 1978; Clemente and Sturgis, 1974; Cox and Catt, 1977; Glenn and Villemez, 1970; Hurlbert, 1976; Knudsen and Vaughan, 1969; Lewis, 1968; Siebring, 1969). In other words, the research so far suggests that for graduate programs the same institutions will be ranked at the top whether or not departments are evaluated “objectively” through quantifiable indicators or “subjectively” through peer ratings.

With the exception of a study which ranked law schools on the basis of a “resource index” (Kelso, 1975), there has been little published research evaluating and ranking professional schools on the basis of objective indicators of quality. However, the evaluation and the ranking of undergraduate colleges have received a modest amount of attention in the last several decades.

Studies of undergraduate colleges tend to utilize one criterion — student achievement — as the basis for rating the institutions. Using this criterion, Bowker (1964) ranked the top 20 institutions on the basis of the number of baccalaureate recipients who received Woodrow Wilson fellowships and the number of baccalaureate recipients who eventually earned a doctorate; Krause and Krause (1970) ranked colleges according to the number of their baccalaureate graduates who contributed articles to *Scientific American*; Dube (1974) ranked 100 undergraduate colleges according to the number of their alumni who entered medical school; Tidball and Kristiakowski (1976) ranked institutions according to the proportions of their baccalaureate graduates who went on to earn doctorates; and Astin and Solmon (1979) ranked institutions according to a selectivity index based on an estimate of the average academic ability of an institution’s entering freshmen (Astin and Henson, 1977).

Several other studies have used multiple criteria and multiple indicators as a basis for ranking undergraduate colleges (Brown, 1967; Gourman, 1967; Jordan, 1963). By way of illustration, Brown (1967) grouped colleges on the basis of eight indicators: (1) proportion of faculty with a doctorate, (2) average faculty compensation (salary and fringe benefits), (3) proportion of students going on to graduate school, (4) proportion of graduate students, (5) number of library volumes per full-time student, (6) total number of full-time faculty, (7) student-faculty ratio, and (8) total current income per student.

Comparisons of the published rankings of undergraduate colleges, whether the rankings are based on single or multiple criteria and indicators, reveal substantial disparities. Although many of the same colleges appear in all or most of the rankings, studies using objective assessments have not consistently identified the same colleges at the top of the rankings. Moreover, even in those studies where many of the same institutions have been ranked, there is wide variation in the absolute ranks of colleges. In short, research aimed at ranking undergraduate institutions has not resulted in the high degree of agreement across studies that has characterized research on graduate programs.
Critique
Since studies based on objective indicators have been viewed more as a check on reputational studies than as an independent method of rating programs, there has been relatively little scholarly criticism of this line of research, and most criticism has been confined to objections about specific indicators. Nevertheless, several overall criticisms — both methodological and substantive — should be raised about evaluation studies based on objective indicators.

Most important, serious methodological questions can be asked about the appropriateness of many of the indicators or measures that have been employed in these studies. Instead of establishing the appropriateness of their measures, either through conducting independent surveys of professionals or by selecting indicators from the literature, most researchers apparently have selected indicators on the basis of their personal preferences and, presumably, data availability. As a consequence many, if not most, of the indices used to assess various evaluative criteria may not be good indicators of program quality.

As a result of the rather haphazard approach to selecting indicators, there is little consensus in the literature concerning the most appropriate measures for evaluating program quality. For example, there is little agreement among researchers regarding the best measures of undergraduate student quality. Is it the proportion of graduates going on to graduate school? Proportion of graduates who earn doctorates? Academic ability of freshmen as measured by standardized tests? Of course, differences over the appropriateness of any measure or index are to be expected in any area of investigation. But when there is such little agreement among researchers regarding the best measures, it must be concluded that more attention needs to be given to establishing appropriate indicators.

In the search for measures of program quality, researchers may be well advised to seek quality indicators that assess adequacy as well as frequency or volume. Most of the existing research is based on indicators that assess easily quantifiable program characteristics (such as number of volumes in the library) rather than adequacy (such as comprehensiveness and accessibility of the library). To be sure, some researchers have attempted to measure the adequacy of certain program characteristics. In the area of scholarly productivity, for example, some investigators have sought to separate quantity of publication from quality of publication. Instead of simply measuring scholarly publication using frequency tabulations, they have constructed "citation indexes" as a measure of the quality of scholarship (Smith and Fiedler, 1971). While citation indexes represent an honest attempt to improve quality measurement, they have — like other indices that attempt to measure quality — come under attack from various scholars (Webster, 1981, pp. 22-23). The debate over most indicators of quality serves to underscore the more telling point: the search for appropriate indicators of program quality is only in its infancy. Put another way, researchers need to improve upon their "objective" assessments of the valuative, subjective concept of quality.
Questions about the appropriateness of indicators are linked to a second major criticism that can be made about most studies based on objective indicators, namely, that they are in fact highly subjective. Since research is infrequently based on "objective" measures that have been established through research, most studies reflect the researchers' biases in the selection of indicators. Ironically, much of the research in this tradition can be fairly criticized for its failure both to acknowledge and reduce the "subjectivism" that is intended to be reduced or eliminated in this approach to evaluation research.

Third, studies at the graduate level have been heavily based on the single criterion of faculty research productivity. Although a range of research-oriented indicators have been used to measure that criterion, the reliance on a single criterion has doubtless contributed to the marked similarity in rankings of programs across comparable studies. Most studies at the graduate level have neglected other evaluative criteria, especially teaching and program effectiveness.

Fourth, most of the indicators used to assess quality are useful only for ranking schools at the very top (Webster, 1981, p. 22). For example, the number of Guggenheim Memorial fellows on the faculty or the number of students contributing articles to *Scientific American* may be good indicators of faculty and student quality at leading institutions. But for the vast majority of postsecondary institutions, their utility, relevance, and appropriateness are highly doubtful.

Fifth, none of the studies using objective indicators has assessed programs in terms of contributions to bringing about changes in students — presumably the major purpose of education at the graduate and professional levels as well as the undergraduate level. This "value added" concept has been advanced by many scholars in the higher education literature (Astin, 1977; Bowen, 1977; Blackburn and Lingenfelter, 1973; Clark, Hartnett, and Baird, 1976; Lawrence and Green, 1980), but such an approach has not been used in any large-scale study of program quality in higher education.

Finally, several criticisms made earlier about reputational studies also can be applied to research based on objective indicators. First, most studies have focused on the department as the unit of analysis and thus have failed to use indicators of quality that take the institutional environment into account. Second, because almost all studies have been concerned with ranking programs, they have identified a rank order of excellence rather than evaluated programs in terms of a normative standard of quality. Third, by making the questionable assumption that programs have identical goals, objective indicator studies have assessed all programs against the same yardstick. Fourth, they have consistently focused on ranking only the "top" institutions, thereby ignoring programs in the vast majority of postsecondary institutions.

Despite these criticisms, studies based on objective indicators have already made a contribution to evaluating program quality in higher education. In terms of methodological sophistication, many studies have gone beyond reputational studies not only in using a more "objective" approach but, perhaps
more significantly, in employing a multidimensional approach to evaluating programs. Especially at the undergraduate level, a range of criteria — and multiple indicators of those criteria — have been used to assess program quality. Moreover, studies at the graduate level have yielded defensible rankings of "leading" departments in terms of faculty scholarly achievement — rankings which generally have a close correspondence to those in the most recent ACE-sponsored reputational studies (Cartter, 1966; Roose and Anderson, 1970). At the undergraduate level, "top" programs have been identified — though the wide difference in rankings across studies indicates considerable disagreement over the top-ranked programs.

In summary, this overall approach to program evaluation holds considerable promise. If researchers can more effectively respond to the methodological and substantive criticisms raised here, and above all seek to anchor their criteria in theory and relate them to program goals, then research in this tradition can truly become more "objective."

STUDIES OF THE QUANTITATIVE CORRELATES OF QUALITY
In the last few years, a growing number of researchers have searched for quantitative correlates of quality. These researchers have attempted to identify objective characteristics, or "correlates of quality," associated with programs that are taken to be of high quality.

Most of the studies on the quantitative correlates of quality have followed a similar format. The researcher first selects one or more fields that have been evaluated and ranked in a reputational study (in some studies, more than one field is examined so that comparisons can be made across fields). Then, by either comparing simple (bivariate) correlations of one or more variables, or by using multiple regression or multiple discriminate function analysis to estimate the relationship between groups of variables and those programs having high ratings, the researcher distills a cluster of objective traits associated with program quality.

Most studies in this tradition have examined the correlates of program quality as measured in the reputational studies by Cartter (1966) and Roose-Anderson (1970). Since the two ACE studies focused exclusively on evaluating graduate program quality, most researchers have examined the correlates of program quality solely at the graduate level. However, one major study has been completed at the undergraduate level (Astin and Solmon, 1981; Solmon and Astin, 1981), which will be discussed after a review of the literature on the correlates of graduate program quality.

Cartter (1966, pp. 112-114) was the first to find strong relationships between several objective correlates — including faculty compensation and library resources — and institutional rankings. Shortly thereafter, a number of other researchers followed his lead, examining the relationship between a range of potential correlates of quality and the Cartter and Roose-Anderson rankings in specific disciplines.
Many early studies focused on graduate programs in sociology. For example, Lewis (1968) found some correspondence between measures of faculty and student publication and the Cartter rankings of sociology departments. Other investigators found relationships between the Cartter rankings of sociology departments and various indicators of quality, including individual prestige (Lightfield, 1971), research productivity (Knudsen and Vaughan, 1969; Glenn and Villemez, 1970), student-faculty ratios (Janes, 1969; Lavendar, Mathers, and Pease, 1971), and hiring patterns (Gross, 1970; Schichor, 1970). Studies conducted later, both in sociology and other single disciplines, identified such correlates as size (Elton and Rogers, 1971; Elton and Rose, 1972) and research productivity (Drew, 1975; Guba and Clark, 1978) as well as a number of other correlates (Abbott, 1972; Solmon, 1972; Solmon and Walters, 1975).

In the last decade, several other researchers have worked with subsets of the ACE-rated departments. Hagstrom (1971), for example, sampled 125 departments in four scientific fields: physics, chemistry, mathematics, and biology. He found large, significant correlations between departmental prestige (Cartter ratings of the quality of graduate faculty) and department size, research production, research opportunities, faculty background, student characteristics (including undergraduate selectivity), and faculty awards and offices. Subsequent studies which sampled at least four major fields also found relationships between the ACE ratings and other quality indicators: a National Science Board (1969) study found 14 correlates of quality; Beyer and Snipper (1974) identified fourteen independent variables; Morgan, Kearney, and Regens (1976) found four predictor variables; and Drew and Karpf (1981) identified a powerful single predictor variable — departmental rate of publication in highly cited journals.

While the search for empirical correlates of graduate program quality has yielded some impressive findings, several caveats are in order. First, many of the correlates that researchers have identified are highly interrelated with one another, and the independent effects of specific correlates have not always been isolated. Second, the magnitude of correlations varies somewhat across comparable studies, owing at least in part to the particular correlates and measures of those correlates that researchers have used. Third, some researchers have found some noteworthy differences among the relative importance of some key correlates across different fields (Hagstrom, 1971; Beyer and Snipper, 1974).

Although the magnitude of correlations varies somewhat across studies and disciplines, the correlations have been generally high enough to conclude that researchers have been quite successful in identifying quantitative correlates of departmental quality at the graduate level (as indicated in the ACE reputational studies). As Hagstrom (1971, p. 389) stated a decade ago: "Unless it is possible to devise some systematic causal theory, it will make little sense to seek additional predictors of departmental prestige or productivity. More refined measurements will permit only slight improvements in the proportion of variance explained."
While research on the quantitative correlates of quality has focused almost exclusively on the graduate level, a recent study by Astin and Solmon (1981) examined objective correlates of reputational quality ratings in seven undergraduate fields. In brief, Astin and Solmon wanted to know if it was possible to estimate undergraduate quality by examining objective, quantifiable information about an institution.

While the researchers found some differences across fields, they identified a number of factors that correlated quite highly with their reputational ratings of “overall quality of undergraduate education” (Solmon and Astin, 1981). The major correlates they identified include size, prestige, selectivity, financial expenditures per student (such as expenditures for educational and general purposes), and measures of curricular concentration (for example, higher quality was associated with relatively large concentration of bachelor’s degrees in the natural sciences and engineering). They found that a combination of prestige (weighted positively) and size (weighted negatively) provided a highly accurate estimate of overall quality ratings in each of the seven fields.

The study by Astin and Solmon is an important piece of research for a number of substantive and methodological reasons, not least of which is that it focuses attention on quality at the undergraduate level. Yet as Astin and Solmon themselves suggest, more research is needed on the correlates of quality at the undergraduate level. In particular, research is needed which is based on assessments of quality other than the Solmon and Astin (1981) ratings and which uses quality ratings from a larger sample of institutions and undergraduate fields.

Critique

Based on this review of the research on quantitative correlates, it is tempting to conclude that this approach has been highly successful — at least in the graduate domain. After all, researchers seem to have reached considerable agreement on the major correlates of quality at the graduate level, and some scholars have concluded that further research at the graduate level is unlikely to make much of a contribution. In our view, however, this conclusion is premature, and we begin our critique with a discussion of the major limitation of research in this tradition.

First, almost all of the research has examined correlates of quality as measured by the ACE-sponsored reputational studies of graduate programs at leading institutions (Cartter, 1966; Roose and Anderson, 1970), now more than a decade old. Significantly, researchers have not searched for correlates of program quality as measured by objective indicators or through such studies as accreditation reports and statewide reviews of program quality. As a consequence of being completely dependent on the ACE ratings of program quality, research on the quantitative correlates of quality is, in turn, subject to all of the limitations associated with these reputational studies of program quality.

Because of the dependency of correlational studies on the ACE-sponsored
reputational studies, the main limitations of those studies bear repeating: they have methodological limitations (such as rater bias), they focus exclusively on the graduate (doctoral) level, they measure program quality based largely on faculty reputation, and they focus only on highly visible institutions. Even if the methodological limitations are minimized, the singular fact remains that quality has been identified only at the “top” institutions, at one level (doctoral), using one major criterion (faculty quality).

Further research is needed which overcomes the limitations associated with exclusive reliance on reputational ratings to identify program quality. More specifically, researchers should identify high-quality programs at institutions which have been rated on criteria in addition to faculty quality and which are not limited to the “top-ranked” ACE institutions. Either by conducting their own studies or by relying on various independent assessments of program quality, researchers should use ratings of program quality that are not as limited in scope and generalizability as the ACE studies.

Second, most studies have used an “ atheoretical” approach in identifying quantitative correlates of quality. Instead of identifying potential correlates on the basis of a theory of quality (which, in turn, would suggest specific variables that may be associated with quality), many investigators seem to have rummaged through their data in search of any factors that might conceivably be linked empirically to program quality (exceptions include Drew and Karpf, 1981, and Solmon and Astin, 1981). Thus, even though research has yielded considerable agreement on the major correlates of program quality at the graduate level, there are few explanations as to why particular correlates make theoretical sense. If we are to move toward the development of theory about quality, the selection of correlates must be linked to some emerging theory of quality. Otherwise, we are left with only a series of empirical correlates that lack any theoretical foundation.

Third, many of the correlates of quality that have been identified are highly interrelated with one another (for example, size and prestige), and many researchers have failed to isolate the independent effects of individual correlates. In some cases, this omission is due to limitations associated with the research design. For example, studies that rely on bivariate correlations cannot, by definition, isolate the independent effects of individual correlates. In other cases, researchers using multivariate approaches (such as regression analysis) have not examined the interrelationships of variables.

Finally, studies of the objective correlates of quality have focused almost exclusively on the graduate domain. Aside from the recent study discussed above (Solmon and Astin, 1981), little attention has been given to the undergraduate level, and no studies have been conducted using professional schools.

Despite these criticisms, there have been some impressive findings regarding the quantitative correlates of graduate program quality. Although researchers may be justly criticized for often failing to link their findings to any theoretical scaffolding, the fact remains that a number of correlates have been identified
that seem to have a strong association with program quality. Thus this approach to studying program quality has rich possibilities. At the undergraduate and professional levels, as well as the graduate level, further research is needed which addresses the methodological criticisms raised above and, equally important, is based on evaluations of program quality that are not encumbered by the limitations of the ACE reputational studies. The ACE studies have been a convenient measure of program quality, but other assessments are needed if the major correlates identified are to be linked with a more multifaceted conception of quality.

SUGGESTIONS FOR FUTURE RESEARCH
There is a pressing need to assess program quality not just in the nation's most visible universities but also in programs that are scattered across nearly 3,000 other institutions. Our search of the literature uncovered few such studies, and those efforts, for the most part, were not successful. In short, there are practical and theoretical needs to expand the level of program assessment, for what has been done so far is limited primarily to the Ph.D. and professional-degree levels.

Our own efforts to move beyond the elite institutions are beginning to bear some fruit (Conrad and Blackburn, 1985). In brief, we have collected extensive data on all public four-year colleges and universities in two states. The data include institutional self-studies and reports, public documents from the states, HEGIS reports, data collected from site visits, vita on all faculty in departments of biology, chemistry, education, history, and mathematics, and, most important of all, the evaluation reports of outside teams of peers who reviewed the programs. None of the institutions are in the Carnegie Research University-I classification; most are regional colleges. The departments studied were principally at the master's level.

One of our most important findings has been that faculty scholarship accounts for only 28% of the variance in departmental quality in these colleges and universities, not the 50% to 80% that is usually found when leading Ph.D. programs are examined. Hagstrom concluded (1971, p. 385) that there is little need to extend the number of measures when several will account for over 70% of the variation. But we have had to introduce other variables, conceptually as well as operationally, in order to predict program quality in regional colleges and universities.

In addition to collecting information on faculty attributes, we systematically obtained data on program characteristics (curriculum), facilities, students, and support. Future research is needed along the dimensions just indicated, and on other aspects as well. For example, Pace (1980) has instruments for estimating the quality of student effort in a program, a series of scales that rest on the assumption that the quality of a student's education depends in part on how well she or he makes use of what the institution has available. Pace's findings show variation across institutions, but as yet there have been no systematic studies
which link his scales to an independent assessment of program quality.

In a similar vein, Clark, Hartnett, and Baird (1976) developed for the Educational Testing Service a set of instruments which contain a variety of dimensions expected to be related to program quality. (These exist at both the master's and doctoral level.) In addition to the typical faculty indicators, Clark and her colleagues have included student assessment of instruction and general satisfaction with a program. Her scales also include data collected from alumni. As yet, however, there have been no large-scale studies of program quality against an independent measure. They, too, should be on an agenda of future research.

In our own studies in progress, we noted the mention of leadership in connection with program quality in more than one peer assessment report. A department often would be considered "weak," i.e., of lower quality, when the department chair was judged ineffective or if there was not imaginative leadership at the dean or vice-presidential level, and conversely.

In addition to leadership, many of the peer assessment reports in our study mentioned esprit, morale, clarity of purpose, and a healthy organizational climate as factors contributing to program quality. While there are few readily available quantifiable indicators for these variables, that does not mean that they cannot be created. A profitable area of research would be to develop ways of estimating such variables and then relating them to program quality.

Finally, as attempts are made to assess the quality of programs at other levels — for example, at the bachelor's level and in remedial English programs in community colleges — we believe ever increasing attention must be paid to the environment in which a program resides. Indicators of the environment need to be included to assess quality. For example, the quality of a biology department may be highly dependent on the quality of the chemistry department.

In conclusion, there will doubtless continue to be controversy surrounding any approach to studying program quality. For no matter how it is approached, "quality" is ultimately based on subjective judgments. Nevertheless, this review and critique of the literature suggests that researchers can improve their designs for studying quality. Given the concern over quality in the nation's colleges and universities by the public and the profession, we invite researchers to address with alacrity the research agenda on program quality.

NOTES

1. We are indebted to A. Richardson Love for many of our reflections on the dimensions of quality. His analysis exists in an unpublished paper (Ph.D. preliminary examination), "The Assessment of Quality in a Selected Liberal Arts College," Center for the Study of Higher Education, University of Michigan, 1982.

2. For example, Young's (1983, p. 461) editorial in Science states: "Yokagawa-Hewlett-Packard (YHP) was honored last year with the Deming Prize, Japan's highest prize for overall quality. The award recognized a 5-year program that reduced the production costs by one-third and inventory by two-thirds, and warranty failure rates by
more than half. During the 5-year period, YHP almost tripled its market penetration" (emphasis added). The use of the term “quality” here makes it essentially equivalent to effectiveness.

3. Although the CGS-ETS study does not identify institutions, the authors report that their rankings are very similar to the 1966 and 1970 ACE lists (Hartnett, Clark, and Baird, 1978).

4. For example, one argument advanced against reputational studies is that they are intrinsically “subjective.” This criticism is unfair, for it ignores the fact that reputational studies are purposively based on “subjective” peer evaluations. Moreover, subjectivity cannot be completely avoided in evaluation studies, regardless of the evaluation technique used. As Cartter noted, even so-called objective measures (such as number of Nobel laureates on the faculty) are, for the most part, “‘subjective’ measures once removed” (1966, p. 4).

5. While several of the reputational studies have used more than one criterion, faculty quality has been the major criterion in all studies at the graduate level. For example, though the Cartter (1966) study of doctoral programs used “effectiveness of the doctoral program” as well as “quality of the faculty,” the high correlation between the two sets of rankings in the study suggests that raters probably chose to emphasize what they knew best (i.e., quality of the faculty) when evaluating “effectiveness” (Blackburn and Lingenfelter, 1973, p. 24). In effect, faculty quality was the major criterion.

6. It should be noted that expanding evaluative criteria beyond faculty quality may lead to different rankings of programs. Several evaluation studies, one conducted at the graduate level, found that when independent assessments of commitment to teaching/program effectiveness and faculty quality are made, programs are ranked quite differently (Astin and Solmon, 1981; Hartnett, Clark, and Baird, 1978).

REFERENCES


American Council on Education. Survey of 37 academic fields and 130 schools in U.S. finds graduate facilities have improved substantially between '64 and '69 at most schools. New York Times, January 3, 1971, p. 54.


Rogers, W. R. Inaugural address of the president of Guilford College, Greensboro, N.C., July 31, 1981.


