In writings on the theory of valuing, many take the position that impacts on the relevant outcome dimensions should be aggregated to arrive at one summary assessment of program merit. A contrary position is taken here, specifying that the impacts should be kept separate and unweighted and expressed only in their own original measurement scales. All impacts, however, should be portrayed, including those for which no rigorous data analysis has been carried out. It is argued that aggregating, even by the individual stakeholder, is both futile and misleading. An extensive evaluation of the effects of research grants on the university is included as a full-scale illustration of the method. Financial impacts are considered, as well as impacts on faculty and student quality, on university prestige, and on the quality of instruction.

THE IMPACT PROFILE APPROACH TO POLICY MERIT

The Case of Research Grants and the University

LAWRENCE B. MOHR
University of Michigan

Tribe, in 1972, wrote a penetrating article that, for the most part, was a critique of what we would now call *quantitative policy analysis*. He found several theoretical problems that might serve to invalidate the whole idea of policy analysis as commonly practiced, even in its own internal terms. His root problem, although he does not discuss it as such, is with policy analysis as a means of deciding on the overall merit or value of a multiple-outcome policy. Many now refer to this area as the *theory of valuing*. Some of the difficulties he raised apply quite generally to almost all policy analysis. For example, analyses tend to be distorted by the practice of accepting uncritically the stakeholders' declarations as to which outcomes are relevant, as well as by the use of "detached and emotionless vocabulary" that masks moral realities

AUTHOR'S NOTE: *Previous versions were delivered at the Fourth National Public Management Research Conference, University of Georgia, Athens, Georgia, October 30 through November 1, 1997, and at the meetings of the American Evaluation Association, Atlanta, Georgia, November 6 through 10, 1996. The research reported here was supported by the Office of the Vice President for Research, the University of Michigan. I gratefully acknowledge the assistance of Charis Kaskiris as well as that of Mark Long and Jahna Otterbacher.*

EVALUATION REVIEW, Vol. 23 No. 2, April 1999 212-249

(Tribe 1972, 107). Two of his concerns, however, have to do with the technical problem of aggregating impacts on several outcomes into one summary judgment. Here, terms such as *discontinuities* and *interaction* crop up in his treatment. Tribe himself does not take the view that such difficulties make reasonably constructive policy analysis infeasible but rather that they should prompt us to recognize that the objectivity and rationality purported to be inherent in quantitative policy analysis generally are not achieved and that there are steps that might be taken to keep these various problems from invalidating the products of the discipline.

Tribe (1972) is not alone in harboring reservations about the validity of a kind of policy analysis that depends on aggregating the relevant costs and benefits. In economics, in particular, concern has been expressed about such problems as the determination and inclusion of the value of nonmarketed goods in such analyses (e.g., Kopp, Pommerehne, and Schwarz 1997) and the nonuse value of the simple existence of an entity (e.g., Rosenthal and Nelson 1992). Still, there has been a fairly general willingness to accept revealed preference, or actual market behavior, as a measure of value when it is applicable, without emphasizing the complexity of market decisions in the same spirit as this emphasis is applied to nonmarket decisions—for example, to verbal declarations of value in response to survey questions (Kopp, Pommerehne, and Schwarz 1997). Tribe's (1972) concerns cut to the heart of the analysis of values placed on goods and services no matter how they are revealed, whether the governing discipline be economics, certain branches of psychology, law, or the general practice of policy analysis by other social scientists. This article is meant to reinforce these basic concerns and to continue the quest for alternative methods that do not suffer the same liabilities.

The position taken here will be that the sorts of technical problems raised by Tribe 1972 and by others are so severe as to counsel strongly against aggregating or synthesizing impacts to form a summary judgment of value. This does not invalidate policy analysis but rather suggests that it be carried out on a different basis—one that omits any final step of arriving at a single statement of value covering a multiple-outcome program or policy as a whole and replaces the summary judgment with another kind of information. The argument does not deal specifically with problems in economics such as the valuation of nonmarketed goods, but it does have implications for methods of dealing with such problems that I hope will be clear. In addition, the theoretical discussion will be bolstered by an extended research example that is more true-to-life in its complexity than the hypothetical illustrations often used in this context.

Valuing will be taken to mean assessing the merit or worth of a program or policy. For example, one might seek to determine that a policy has been satisfactory, marginal, or unsatisfactory or that conducting a certain program was better than doing nothing at all. The point of view defended here is that placing a value of this sort on a multiple-outcome program cannot be done rigorously and, in light of this and other considerations, should not be attempted by practitioners or scholars as part of policy analysis. Our responsibilities as researchers end at properly assessing the multiple impacts of the program or policy. Going beyond that to assessing its overall worth is something we do only in the role of and primarily with the methods of concerned private citizens. On the other hand, valuing is and must be carried out by individual stakeholders, and the policy analyst has a definite role in providing the best information possible to support that purpose. Thus, it is proposed in this article that the aggregating of impacts, as in benefit-cost analysis, multiple-attribute utility technology (MAUT) (Edwards and Newman 1982), or the qualitative weight and sum method (Scriven 1994), be replaced consistently by the presentation of a profile of impacts, as will be explained and illustrated below.

THE THEORY OF VALUING

The position on the theory of valuing to be proposed below is at odds with that of Scriven and consistent in many respects with that of Cronbach. Both of these perspectives have been reviewed admirably by Shadish, Cook, and Leviton (1991). Scriven (1994, 1995) does not quarrel with the limiting view held by many policy analysts that impact assessment marks the outer edge of their role, but he works in this debate from the perspective of the discipline of program evaluation, and he feels that those who hold this view should not call themselves “evaluators.” It makes little sense to use that descriptor, he feels, when one does no evaluation (i.e., valuing) whatsoever. Scriven himself advocates going beyond impact assessment to reach a conclusion on merit, either absolute or relative, and he argues in favor of the validity of various methods available to accomplish the task. It is that particular debate that is joined in this article. Scriven hews to the principle that those who would call themselves “evaluators” should not stop short of valuing. However, it is not necessary and is perhaps ill-advised to be so controlled by labels. It remains true that the central and salient focus of what good evaluators do is, above all, to *contribute* to valuing. Beyond that, if one considers it a technical misnomer to apply the term *evaluators* to policy analysts who stop short of assessing overall merit, then perhaps it is better to remain somewhat misnamed than, being strictly ruled by the name, to emphasize and commit a lot of energy to an activity that cannot be implemented with sound and valid methods.

To go as far toward the judgment of merit as the policy analyst properly should, it has been suggested that the analysis process should have five essential components, namely, finding, limiting, impact assessment, common scaling, and weighting (Mohr 1995, 275). Most will readily agree that the third of these, impact assessment, is a necessary component of an evaluation process that will end in judgments of merit or worth by the stakeholders. Of the remaining four components, it will be argued below that the first two, finding and limiting, are crucial for researchers to carry out in order to permit stakeholders to do a good job of valuing on their own. However, as for the last two, common scaling and weighting, there are so many pitfalls and inadequacies in attempting to provide such information, the costs of the attempt are so great, modes of partial implementation that might be good enough for practical purposes are so difficult to find, and the results can be so misleading, that these components, it is concluded, should be avoided by the analyst.

Finding. *Finding* means determining those outcome dimensions on which the subject program's potential impacts are of interest. The term *of interest* here means that certain appropriate information on these dimensions would be useful to stakeholders in assessing the overall value of the program. This is an impossible task to perform in a definitive way—that is, so that one can say, "I have now done a perfect job of finding or identifying the relevant outcomes." Some of the impacts of the program, although they eventually will be recognized as important, might be so far in the future or so obscured by complexity or ignorance that no one can possibly think of them when the finding job needs to be done for a current evaluation. Thus, we are confronted immediately with the consequence that valuing cannot be a strictly logical, rigorous affair. At the same time, policy analysis would be seriously undermined if, on this account, we gave up completely the goal of providing adequate information to permit the assessment of value. The job generally must be done and because of the inherent inadequacy of outcome-identification methodology, it cannot be done algorithmically. One must conclude that it is the responsibility of evaluation theorists and practitioners to work out satisfactory, although not definitive, solutions to this problem. In doing so, the finding component in large measure becomes one of identifying outcomes in such a way as to keep partiality out of evaluation as much as is possible.

It follows that the principle on which the component of finding should be executed is the principle of inclusiveness. The finding process should consist of making contact with all possible stakeholders, trying one's best to discover all of the outcomes that might affect their assessment of worth, and including all of those in the evaluation in some form (cf. Scriven 1994 on goal-free evaluation and avoiding bias, Shadish, Cook, and Leviton 1991:79-81; and

Cronbach 1982 on inclusiveness of outcome-finding to generate evaluations in the public interest, Shadish Cook, and Leviton 1991:354-355). Again, failure is almost certain. If one can point afterward to a single stakeholder whose outcomes of interest have been omitted, then the finding component has not been perfectly executed. This result is to be expected on occasion no matter how "professional" the intentions of the analyst. In this matter, we carry out our professional responsibility when we have done our best—that is, when we have made an honest and energetic effort to achieve inclusiveness and have plainly documented the process for review by others.

Limiting. Similarly, investigating all of the outcomes that have been discovered through all of the stakeholders is a greater task than any evaluator is likely to have the time and other resources to carry out. Some of the outcomes will become the object of a lot of research design and data collection, but some will no doubt have to be left to speculative impact assessment. Again, this is an area for the evaluator's professional judgment (keeping in mind that without the discipline of evaluation, almost all impact assessment would be speculative), and these judgments make up the component of limiting. For one, rigorous impact assessment on some outcomes will be impossible or too expensive or time consuming. Beyond that, however, choices frequently may have to be made governing on which outcomes to spend time and money based solely on one's estimates of the potential importance of the outcomes in the assessment of worth by stakeholders (similar to Cronbach's [1982, 240] leverage). Some probably will be of negligible importance, and those should be left for speculative assessment.

But there is a very important principle to be followed here in the area of program evaluation, and that is to articulate all identified outcomes and not to lose sight of any of them in the reports of the analysis, regardless of whether they become the objects of research design and data collection. There are two critical reasons for this. One is to fulfill the obligation of providing the best information one can to permit assessments of merit by all stakeholders. If some potential impacts are omitted from mention entirely, certain stakeholders may forget about them or never think of them and so have a distorted view of the overall impact of the program from their preferred perspective. True, one will not be in a position to give them a rigorous impact assessment on some of the outcomes, but from the stakeholder's perspective in valuing, knowing where one does not have desirable information is far better than never imagining certain information to be desirable at all. The second reason for keeping all outcomes of potential interest in focus is to provide a check on the possible bias of the evaluator. The choices of what to get data on should be defensible and therefore public, even if one had to resort to the flip of a coin to make some decisions in the limiting component. In other words, no outcome of potential interest should be swept under the rug.

Common scaling and weighting. In the end, the value of the program in the judgment of the evaluator or any stakeholder and any action preferred, such as discontinuing the program, increasing the funding, and so forth, must be based on a consideration of the levels of impact on all of the various outcomes. This sets up what Scriven (1995) refers to as the "performance synthesis" problem. To know how much we value the overall performance of a program, it is necessary to know what all of the important individual impacts have been. Then, to arrive at the total or net worth, it commonly is considered that one must put all of the impacts on the same measurement scale: One cannot meaningfully add and subtract quantities in different units. For example, one cannot meaningfully subtract efficiency losses expressed in units of output per input dollar from morale gains expressed as points on an attitude scale. Here is where we begin to see the futility of valuing as a responsibility of the professional evaluator.

There are several prominent approaches to the common-scaling problem. One, covered in multiattribute utility technology (Edwards and Newman 1982) and sketched by Scriven (1994) as the quantitative weight and sum approach, proceeds as follows: For each outcome (such as efficiency and morale), identify points on the original measurement scale that one can label as *minimum* and *maximum plausible* (or *tolerable*). Rescale these points as 0 and 100, respectively. Then, convert quantities in the original scale to this new scale so that they become, for example, 28% of the way from minimum to maximum plausible, or 35%, and so forth. The problem with this method is that identification of the anchor points of the new scale depends on judgments, sometimes difficult ones, and often can be arbitrary or can require guesswork. There is no algorithm, even for one individual valuing alone. When the final figures are obtained, especially if one has some lack of faith in the overall verdict on merit or some degree of discomfort or surprise, one quite naturally wonders what the result would have been if these anchoring scale points had been different. If there are many outcomes and therefore many anchoring points to play with individually and in combinations, the complexity becomes imposing. Furthermore, when looking back at the wisdom of particular anchor points and considering changes once the data are known, it becomes difficult not to be influenced by the potential results of any revised specifications in terms of the verdict on merit. Objectivity is lost. One is pulled toward manipulating the figures to comport with one's gut feeling about composite value. On the other hand, it will not help to apply some rule such as, "Once you have set the anchor points and seen the data, you cannot go back and change your mind." Given the situation, that would only mean that the outcome is not to be trusted whether one changes one's mind or not.

The other prominent method of common scaling, one that is integral to benefit-cost analysis as generally practiced, is to convert everything into monetary units, such as dollars. The problem here is that this exercise cannot be carried out without executing the weighting component at the same time. To equate so many dollars with a score of 53 on the Morale scale, one has to decide how much a point on the Morale scale is worth. What is functionally the same kind of thing must eventually be done in multiattribute utility technology as well—that is, one must assign weights that tell us that 28% of the way on the Morale scale is worth half as much or twice as much, and so forth, as 28% of the way on the Efficiency scale. Weights similarly must be assigned in the weight and sum method described by Scriven (1994, 374-378). Benefit-cost analysis thus carries us automatically into the weighting component, but all methods of valuing also must include this component in some fashion. Here, however, is where the professional approach to valuing finally must break down altogether, even if all of the other problems could be managed in some reasonably practical if less-than-rigorous way. Remembering the difficulty that arises with the need for common scaling, the weighting problem introduces two additional, serious obstacles.

The first is that even if weighting can be carried out successfully by a given individual, the weighting result—which criteria or dimensions of value (e.g., morale, efficiency) shall be more important and which less and how much so—depends on the values of that individual. Therefore, the weighting scheme of the evaluator has no better claim to validity than that of any of the stakeholders, so that when the professional comes to the stage of valuing, he or she cannot reach into the evaluation specialist's tool kit and find a justification for one particular set of weights.

The second of the serious obstacles introduced by the weighting problem is that weighting is an arbitrary and futile exercise even for one individual (Tribe 1972). As with the anchor points in common scaling, this means that one will always wonder whether one's first weighting scheme was correct, or was what one "really believed." Tinkering afterward with the relative weights because of lack of confidence in this "correctness" can be influenced by the overall outcomes implied for the value of the program. Moreover, weighting is devilishly difficult, even if people do not think so at first glance. It is commonly pointed out that value or importance is not linear (Shadish, Cook, and Leviton 1991), and this is true in three noteworthy respects. First is curvilinearity: How important morale is—what sort of weight one should put on it—may well depend on how high or low morale happens to be. If we succeed in increasing it some with our program, for example, one then might attach less importance to increasing it further. Second is interaction: How important morale is depends on how much of various other things we have—and not only things connected with the particular program that is supposed to improve morale but also other things, such as whether there is a national or organizational emergency, whether particular leaders quit, and so forth. These things change in the short-term as well as the long. Third, in a similar concern for curvilinearity and interaction, Tribe (1972) and Scriven (1994) both emphasize threshold effects: Perhaps a minimum level of attainment on Criterion A must be reached in order for the program to be considered satisfactory. If it is not, the lack cannot be compensated by any degree of performance on any number of lower weighted criteria; but if it is, the weight of further gains on Criterion A is only moderate. The potential for such curvilinearity, interaction, and threshold effects is not only infinite but also probably very much with us in any complex policy analysis. The implication is that frequently, reliable importance weights, even for a single individual, are not a possibility.

Furthermore, the importance an individual assigns to a dimension is not an inherent characteristic of that individual. It is merely a symptom of how the individual is feeling about things at the time he or she is asked or the time at which he or she reveals a supposed preference by spending money on certain goods in the marketplace (Blackorby 1990; Mohr 1996). Weights easily can change with time—even a short time—in response to new configurations of information and emotion. This is clear in the case of dimensions that are notoriously difficult to quantify monetarily, such as the degradation of a natural vista or the satisfaction of families with the progress of the patient, and it is in fact true of the weighting of anything and everything. If one stakeholder is to accept another's (or the evaluator's) weighting scheme as a point of departure for policy debate, he or she should at least have complete assurance that the weights will not change if the first stakeholder is asked to give the scheme again next week, but one categorically cannot have this assurance. There is no way to guarantee even that level of constancy or certainty.

Scriven (1994) has made a valuable contribution with his idea of probative inference. He makes a convincing case that one should in principle be able to infer the value of something from a purely factual description by having a careful conceptual definition of what that something is and means—a principle followed also in the contingent valuation of nonmarketed goods in economics (Kopp, Pommerehne, and Schwarz 1997). For example, Scriven (1995) persuasively argues that a "pot" is a leak-proof vessel that is able to withstand a great deal of heat, with a handle that stays cool, and so forth. Still, if one wishes to rank pots in value, one must weight those several dimensions, and in principle there is no way to do that either correctly or reliably, as just reviewed. But furthermore, many public programs are not so easy to define in this way, even superficially, for example, the university program of applying to the federal government for research grants. What is the conceptual definition of that program? What is a research grants program? What is its meaning to the university? That is difficult to say. One actually determines the meaning in both easy and hard cases by going through the finding procedure reviewed earlier—formally or informally—a procedure that happens to be easier to carry out for pots and vacuum cleaners than for the program of obtaining federal research grants. But in both instances, even with pots, the procedure necessarily brings the impossibility of objective weighting to the foreground.

These various problems with the weighting component are compounded by the fact that a great deal of evaluator energy can be and often is consumed in the demanding effort to find reasonable weights. We spend a great deal of time and other resources trying to put values on things. Given the limits on the resources available for evaluations, what tends to suffer are the other, more important components: finding, limiting, and impact assessment—especially the component mix that involves keeping track of outcomes on which there are no good impact data and subjecting them to a reasoned speculative analysis.

After reviewing Cronbach’s views on the theory of valuing, Shadish, Cook, and Leviton (1991, 354-358) take issue with his reluctance to admit the value of summary statements of any kind on overall merit or worth. They suggest that, conceding all of the pitfalls of summary statements, some stakeholders want them and there is no great harm and possibly some good in providing them—as long as multiple versions are presented rather than only one, the various versions are accompanied by descriptions of their respective weighting schemes, and they are presented in tandem with rather than instead of the individual impact assessments. Although it is tempting to be flexible on such matters, there is nevertheless substantial reason to be basically sympathetic to Cronbach’s hard-line position. Net-worth summaries generally are not needed or requested when relevant values are few, sharp, and clear. They mainly are wanted when there are close calls, when there are too many values to keep track of easily, when they are not clearly and confidently held, or when they are closely competitive. Summary assessments are wanted by stakeholders when they cannot arrive at them effortlessly by themselves and then, to provide them, the evaluation specialist must begin to research the literature and ask questions of the stakeholders about scales and weights in order to prepare the ground for the calculations. It is probably rare for stakeholders to be aware that there are ideas such as common scaling and weighting at the foundation of valuing in the case of multiple outcomes. This is a chunk of disciplinary arcana about which most people have not thought. When the analysis finally is delivered, however, there are three undesirable stakeholder reactions that may easily crop up. The analysis might be distrusted as tenuous because the critical scaling and weighting, they now realize, have been a matter of the judgment of the moment, or the judgment is too readily accepted as not tenuous, or it is accepted with an avoidance of attention to an underlying stirring of misgivings in order to escape the expenditure of time and thought necessary for the stakeholder to reach his or her own conclusions based on the profile of raw impacts. Professionals should avoid either encouraging or accepting any of these three reactions. Perhaps there are occasions when there would be no danger that any of them would occur, but it would seem healthy to be reluctant to interpret situations as being so innocent.

How is the stakeholder expected to reach a summary conclusion *without* going through the process of common scaling and weighting? The answer is in the same way that we generally make life decisions. We are so constituted as to be able to sift through the probable consequences, seize on those that are particularly important to us, at least at the moment, and arrive at a decision. Sometimes this happens swiftly and effortlessly. Sometimes we have to mull things over and perhaps wait a few days until the decision finally makes itself. Research suggests that one cannot expect to improve on this process by trying to make it rational (Mohr 1996). Scriven (1994, 374-378) suggests the possible solution of a qualitative rather than a quantitative weight and sum approach, but that must also fall afoul of the common scaling and several of the weighting problems noted above and, in any case, still leaves the decision maker with a profile of performances—parsimonious but crude—by which to decide on overall merit. In sum, we are past masters at reaching satisfying conclusions in the face of a set of raw, incommensurable outcomes. One also may arrive at a conclusion by a supposedly rational analysis, including explicit common scaling and weighting, but if that conclusion is not supported by a gut-level feeling of satisfaction, it is not likely to be acceptable.

AN APPLICATION: FEDERAL RESEARCH GRANTS

We move now to the task of illuminating these issues by means of a policy example, that of research grant maximization in the modern university. To begin with some necessary background, the issue that eventually led to this article was raised in the following manner: Informal communications made apparent to the author a growing unease with research grants on the part of officials in charge of finance at many major research universities. The problem is that university accounting, if closely queried, suggests that research grants may cost universities money out of pocket. That is, indirect cost recovery, even on federal grants, (for which indirect cost allowances are relatively liberal—some granting organizations give none at all) is arguably inadequate to cover the indirect costs, so that the institution has to dip into other funds to support the research. The amounts involved, it was suggested, could be in the neighborhood of 25%; that is, around 25% of the amount of research grants, on average, must be added by the university from other sources to the funds received for direct and indirect costs. In that case, the policy issue arises of whether to be selective or at least restrained in some way in applications for research grants or to continue, as most universities do, to try to maximize them (Geiger and Feller 1995). The question is important and difficult, but it becomes particularly acute when the charge is made, as it frequently has been, that the high tuition payments of undergraduates in research universities go in substantial measure to support grant-aided research that, adding insult to injury, then distracts the institution monumentally from its primary mission of undergraduate education and lowers the overall quality of instruction (e.g., Grossman and Leroux 1996).

One way to investigate the problem is to look closely at the accounts in one or more institutions. That sort of study is not within my area of expertise, but in addition, the detailed allocation of costs to categories is so difficult in universities, with decisions on such matters frequently being left to individual discretion rather than to solid rules, that it is at least doubtful whether a satisfactory answer could ever be obtained by this method, even for one institution, let alone with applicability to universities in general. Another approach, and the one taken for this evaluation, is statistical; that is, one may look at research grants and other financial data across a range of institutions to try to learn whether on average, certain impacts of grant revenues on other aspects of finance seem to prevail.

But another issue also arises. Assuming for the moment that research did reveal the unfortunate financial impacts feared by many officials, that would not be the whole story. Research grants probably have many impacts, and some of them might be positive. In that case, a policy of all-out pursuit of grants might be optimal, even if there were some costs. Clearly, the case then would be one in which the theory of valuing would play a prominent role. Given the possible positive and negative effects of current policy, is that policy worthwhile all in all, or is it basically harmful?

Empirical Analyses	
Tuition:	0
Instruction:	0
Faculty quality:	+
Salary:	+
Equipment:	+
Professional standing:	+
Additions to endowment:	0
Private gifts, grants, and contracts:	0
Quality of incoming undergraduate students:	0
Quality of incoming graduate students:	+
Speculative Analyses	
Quality of graduate teaching:	+
Quality of undergraduate teaching:	+
Distortion of academic power structures:	0
Depth and innovation in research:	–

Figure 1: Outcomes and Impacts

Finding. To pursue the finding component, I personally interviewed and otherwise met with a broad range of about 25 faculty, administrators, and students at one university whose lives might be touched in a major or minor way by the research grant activity. Not included were parents of students, as I felt that their concerns were fairly obvious and were covered by students and other stakeholders. Because of funding limitations, the interviews and other finding activities were not duplicated at additional institutions. Last, there were no interviews of officials at granting agencies. The purpose here is not to evaluate the federal research grant programs from the perspective of the government or the country as a whole but rather to evaluate one part of that overall picture at this time: the impacts on the educational institutions that carry out much of the research. It is not a question here, in other words, of the extent to which the government should give grants—we can assume for the moment that the federal government feels it worthwhile to do so—but rather the extent to which universities should pursue them.

Figure 1 presents a list of all of the potential impacts turned up by this finding process. Those outcomes that are investigated with hard data are: tuition; instructional expenditures; additions to the endowment fund; private gifts, grants, and contracts; salaries; equipment; research quality of the faculty; quality of incoming graduate students; quality of incoming undergraduate students; and reputation or professional standing of the institution. Speculative analyses are included for four outcomes found to be salient but on which more rigorous, data-based analyses were not feasible.

Limiting. Perhaps the most influential factor in making decisions on which of the outcomes to investigate rigorously was the availability of existing data and restrictions on resources for getting more. At the same time, many (not all) of the outcomes that are considered extremely important, even critical, by one or more stakeholders are included, because data do happen to be available that can tell at least a partial story on those impacts. Had this not been the case, this particular evaluation would not have been undertaken. That these outcomes are important suggests an implicit weighting scheme. However, the interviewees were not asked systematically for weights. The views of the stakeholders on selected outcomes simply emerged naturally, albeit in rough terms, out of the discussions that were held.

There were four principal sources for the data. The first and most important in terms of supporting a broad section of the project was the CASPAR database supported by the National Science Foundation in awards to the Quantum Research Corporation.[1] Second, the recent National Academy of Science–National Research Council (Golberger, Maher, and Flattau 1995) ratings of doctoral programs were used for measures of research quality of the faculty and professional standing of the disciplines and institutions. Also included in the data set were the National Research Council (NRC) ratings from a decade earlier (Jones, Lindzey, and Coggeshall 1982), which are available as part of the CASPAR database. Third, the Association of American Universities/Association of Graduate Schools Project for Research on Doctoral Education was used for data on average Graduate Record Examination (GRE) scores of entering graduate students in selected programs.[2] Last, *American Universities and Colleges* 1983, 1992 and the *Princeton Review* 1992 were used for average Scholastic Aptitude Test (SAT) and American College Test (ACT) scores of incoming freshmen in selected years.

The research design used in all cases is the ex post facto or correlational design. There was no way to assign research grants either at random or in any other way to some institutions and withhold them from others to implement a stronger, quasi-experimental or experimental design. The great threat to validity in the ex post facto design is selection, especially that due to spuriousness. That is, it is possible that any apparent relation emerging between research grants and the various outcomes is not directly causal but results rather from some omitted third factor (or set of factors) that causes both the research grants and the outcomes to vary. This is a serious limitation, making the results throughout essentially tentative. The picture is not entirely black, however. The threat of spuriousness can be circumvented in many individual analyses, as will emerge from comments on this subject as the various results are reported. Given that there is not the sort of random design here that would permit either causal or population inference, significance tests (actually, *t* values) are used only as a rough measure of the size or strength of the relations investigated.

IMPACT ASSESSMENT: EMPIRICAL ANALYSES

Tuition. The question is whether the acquisition of research grants has on average put upward pressure on tuition. If so, then some of the most serious fears and accusations regarding sponsored research possibly are borne out, and one can speculate that many universities have chosen, consciously or unconsciously, to solve the problem of out-of-pocket costs due to grants in part by raising tuition. This issue will be investigated in two ways. The first is to answer the question, "Does a change in the level of research grants in 1 year bring about a change in the level of tuition 1 or 2 years later?" If an increase in research grants generally is followed by an increase in tuition—and similarly for decreases—this would be substantial evidence both for the existence of the problem and the direction of at least part of the solution employed by the institutions. Spuriousness is not to be feared here. It is easy to see that grants and tuition may have risen together over time, either for independent or for common reasons, but it is not so easy to see what third factor might cause research grants to go down at times and tuition subsequently to remain approximately constant (it almost never decreases) rather than rise in a substantial subset of the institutions. For a relation to emerge, this latter behavior also must be part of the pattern.

Table 1 shows results from the regression of tuition changes in thousands of dollars from 1 year to the next on similar changes in federal research grants in prior years, controlling for the initial size of the institution as measured by total revenues. For example, we have the effect of grant changes from 1992 to 1993 on tuition changes from 1993 to 1994, controlling for total revenues in 1992. The institutions included are those that received the most grant money and that together accounted for more than 95% of all Federal R & D (Research and Development) funds in 1992 on the CASPAR data base. The size of the group was 203 institutions. If the worst fears outlined earlier were realized in the data, we would expect to see large positive coefficients throughout the table, meaning that an increase in research grants fairly consistently resulted in an increase in tuition 1 year or 2 years later. This is not, however, the pattern that is observed. Somewhat more than half of the coeffi-

TABLE 1: Regressions: Tuition Changes From Federal Grant Changes, Controlling for Total Revenues

	Real Dollars			*Constant Dollars*		
Annual Change	*Adjusted* R^2	b	t	*Adjusted* R^2	b	t
One-year lag[a]						
1994-1993	0.37	0.10	1.32	0.30	0.08	1.18
1993-1992	0.43	0.06	1.16	0.36	0.05	1.10
1992-1991	0.36	–0.04	–0.61	0.25	0.00	0.06
1991-1990	0.19	–0.18	–1.72	0.09	–0.15	–1.59
1990-1989	0.27	–0.12	–1.67	0.16	–0.07	–1.25
1989-1988	0.23	0.02	0.31	0.11	0.03	0.55
1988-1987	0.17	–0.08	–2.16	0.10	–0.08	–2.57
Two-year lag[b]						
1994-1993	0.36	0.03	0.46	0.29	–0.00	–0.00
1993-1992	0.45	–0.11	–2.04	0.38	–0.08	–1.98
1992-1991	0.37	–0.08	–1.10	0.25	–0.03	–0.50
1991-1990	0.18	–0.09	–0.80	0.08	–0.03	–0.37
1990-1989	0.25	0.03	0.34	0.15	0.04	0.76
1989-1988	0.22	–0.01	–0.46	0.11	–0.02	–1.10
1988-1987	0.15	0.09	0.63	0.07	0.07	0.48

a. Tuition 1994-1993 from grants 1993-1992, controlling for total revenues in 1992.
b. Tuition 1994-1993 from grants 1992-1991, controlling for total revenues in 1991.

cients are negative, with the overall indications being that the connection between the variables may well be more fortuitous than causal. Many factors, of course, affect tuition levels, but it does not appear as though research grants played a discernible role during these 7 years.

A special word is in order regarding the *t* values. They are not used at any time in this article for statistical inference but rather only as a rough indicator of strength of relationship. The guideline employed was to consider a *t* value of 2.0 or more to indicate a fairly strong relationship, with "strong" therefore expressed in terms of the probability that a regression coefficient this large or larger would have emerged through a random assignment of the research grant figures to institutions (rather than using the institutions' actual numbers), given the size of the group and the variances of the variables. If the *t* value is around 1.5 or 1.0 or less, it means that the relation observed easily could have resulted even from such a random process. Thus, a *t* greater than 2.0—significant beyond the 5% level under the null hypothesis—is not interpreted to mean that the relation is "real" (it certainly is that), or causal (it could be spurious), or descriptive of a larger population (these institutions are considered to be a population in themselves), but rather "large," as scaled in those probability terms. The great majority of the relations in the table then are quite small in those terms. Of course, we do not expect that research grants would cause huge changes in tuition levels, but with a group size this large, even numerically small effects, if pretty consistent, would result in small error variances and therefore sizable *t* values. Such small effects would be large in probability terms.

It seems most appropriate to use undeflated data here—that is, amounts expressed in actual dollars—because the relation depends on the short-term reactions of individuals to expenses, which probably would depend on perceptions of actual rather than deflated dollars. It is not like comparing tuition levels in 1994 with those in 1975, when deflated or constant dollars clearly would be desirable. Table 1 also shows the results using constant 1987 dollars, however, and the pattern conveys the same message.

The effect of research grants on tuition could differ between private (N = 70) and public (N = 133) institutions. One would think, for example, that public institutions, under scrutiny of their legislatures, would feel less free to raise tuitions in response to this particular kind of financial pressure. Tuitions have indeed gone up during this period much more in private than in public institutions (Clotfelter 1996, 3-4). The patterns in these data, however, are quite similar across the two, with most of the coefficients being quite small and divided almost in half between negative and positive. Perhaps public institutions indeed would have had a difficult time trying to justify tuition increases because of pressure from research grants, but the similar record in the private institutions indicates that there was not a compelling need in either sector to raise tuitions for that particular purpose.

The second manner in which this possible relation will be viewed is as follows. Perhaps moderate fluctuations in research grants from year to year would have little discernible effect on tuitions. If the out-of-pocket costs are there, however, and tuition is a frequent response to them, then at least when grant levels rise quite sharply, tuition levels should follow suit. To pursue this line, a time series strategy is pursued using the interrupted time series design. The CASPAR data set has data going back to fiscal year 1975 and when this analysis was carried out, continued through 1994. To implement a good interrupted time series design with this modest number of data points, it is well to have the interruption or intervention point come in about the middle of the period, allowing a reasonable number of points to set the slope of the series both before and after the intervention. The intervention of a sharp rise in federal research dollars was implemented by considering only the 20 institutions with the greatest increase in federal research grant revenues in 1983 over 1982 (the smallest gain in the group being $5.5 million). If the causal effect was present, we should see a jump in tuitions (an intercept change) for these institutions in the following year, 1984, or perhaps an increased upward slope in tui-

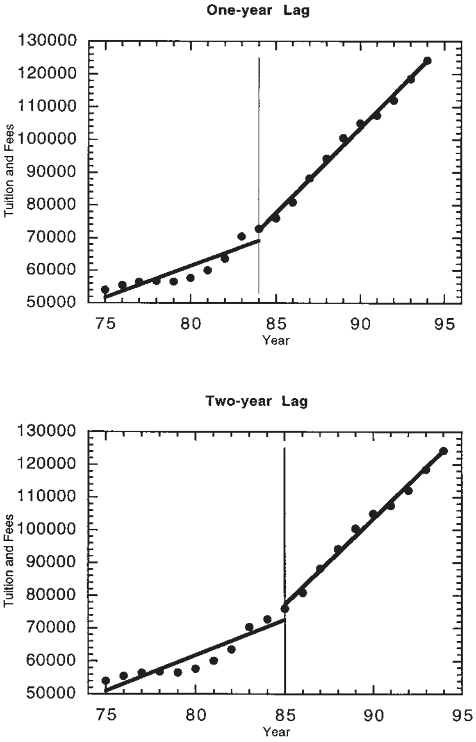
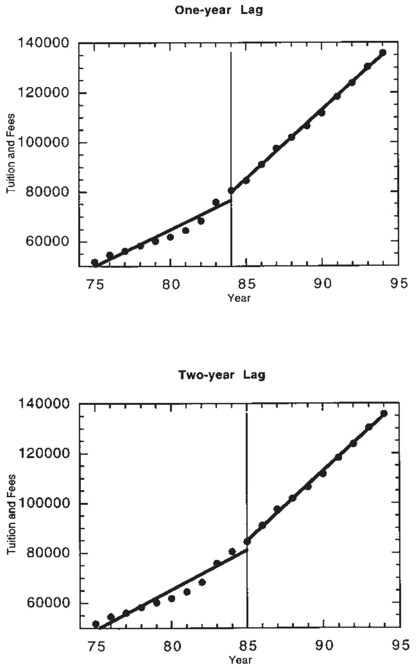


Figure 2: Average Tuition for 20 Gainers From 1982 to 1983

tion revenues for awhile after 1984, or both. The time series for the average annual tuition revenues in thousands of dollars for these 20 institutions is shown in Figure 2, using constant 1987 dollars, and with 1984 scored 0 (and the other years accordingly) to capture the "intercept effect" of the intervention directly from the regression results. With an interaction term included in the model to allow for the possible change in slope, the intercept shows a jump in moving from 1983 to 1984 of about $5 million, with a *t* value of 2.68 (see Figure 2). In addition, the slope of the tuition series is seen to increase dramatically (*t* = 10.73) after this surge in grants. A 2-year lag for the effects to be felt on tuition might be more appropriate, therefore, these results also are shown in Figure 2 and are quite similar.

Can research grants be responsible for such a change in slope? Intuitively, it seems unlikely because a surge in grant funds would not have such protracted effects. The slope would be expected to decrease to the "before" level after a few years instead of continuing to rise steeply until the end of the


One-year Lag

Two-year Lag

Figure 3: Average Tuition for 20 Losers From 1982 to 1983

series. In fact, it is well-known that there were dramatic rises in tuition revenue growth in all of higher education starting in the mid-1980s (e.g., Clotfelter 1996). To test whether these changes both in intercept and slope might have been due to the local increases in research grants or, on the other hand, to the more general trend, the same analysis was carried out on institutions that lost grant revenues from 1982 to 1983. The data for the 20 biggest losers (among whom the smallest loss that year was about $3.7 million) are shown in Figure 3, again allowing 1-year and 2-year lags. The pattern is strikingly similar to that of the gainers. Even when federal grants decreased substantially from 1982 to 1983, tuition jumped a year later ($t = 3.81$) and continued to rise at a notable rate to the end of the series (t for the change in slope = 11.89). One should keep in mind in reading these plots that the data are in constant dollars, so that the year-to-year increases in tuition for both the gainers and the losers are in addition to the effects of inflation.

Similar analyses were carried out using 1981 to 1982 and 1983 to 1984, instead of 1982 to 1983, as the intervention years. The comparative results for the gainers and losers of those years strictly corroborated this initial analysis. In sum, the slope and intercept changes among the gainers probably should be seen as evidence of the absence of particular pressure on tuition because the comparative time series analysis shows that the extreme gainers do not stand out from the extreme losers in this regard.

One might ask why, if not due to the need to support increased research grant funding, tuitions rose so dramatically in the late 1980s relative to general inflation. This question and the parallel question of why general academic expenditures rose comparably at the same time have no satisfactory answers at this point. The best effort to analyze the sources of the marked recent increase in the cost of higher education was carried out by Clotfelter (1996). He indicates that the average annual growth rate in expenditures was a dramatic 3% greater in the 1980s and early 1990s than it had been during the 20 previous years. Concentrating his detailed analysis on three major research universities and one liberal arts college, he used what would seem to be a thorough and powerful model to pinpoint the factors most responsible for the average annual growth rate in expenditures between academic years 1983 and 1984 and 1991 and 1992. Yet, the seemingly exhaustive set of sources incorporated into the model are able to account for only about half of this growth rate. Because one of these sources, and indeed by far the most powerful one, is the growth rate in financial aid and because the inclusion of financial aid and the tuition and fees it supports has the effect of increasing both the total expenditures and total revenues simultaneously and fairly symmetrically, the unexplained residual is actually closer to 62.5% than to 50%. Thus, the rise in these costs is inadequately understood. The above analysis suggests, however, that attributing the tuition increases to the rise in research grants would be an error. On average, that impact appears to have been nil.

Instruction expenditures. Although out-of-pocket costs due to grants might be absorbed in any number of ways, one additional potential problem was of substantial interest to stakeholders, namely, expenditures for instruction. If research grants caused money to be withdrawn from instructional outlays, it would reflect a particularly worrisome trend. These same two analyses therefore were carried out using instructional expenditures in place of tuition revenues. In the analysis comparable to Table 1, we would expect to see mainly negative coefficients. In fact, using all 203 institutions, the coefficients were positive three times out of the seven using a 1-year lag and five times out of the seven using a 2-year lag. Whether positive or negative, almost all of the coefficients were very close to 0. Breaking down into public and private institutions, the hypothesized negative coefficients emerged four times out of the 7 years in the public institutions, for both 1- and 2-year lags, and two times out of the seven in the private institutions, again for both 1- and 2-year lags. However, all of these coefficients and their corresponding *t* values were again so close to 0 that the proper conclusion is one of no effect; that is, grant changes had no general effect on instruction expenditures, and the public and private institutions did not differ from one another in this regard. For our purposes, it seems evident from this first analysis that grant success did not constrain rises in instructional spending during the period considered.

Taking the comparative time series approach on instructional expenditures, the hypothesis would lead us to expect that expenditures would decline after a surge in grant revenues. Given the long and steep rise in tuition observed and the similar data in Clotfelter 1996 and elsewhere on the rise in overall expenditures during this period, a decline in instruction expenditures is highly unlikely. This surmise is borne out by the results. One might expect, however, that increases in intercept or slope would be greater for the losers, as grants presumably were not exerting downward pressure, than for the gainers, as grant increases should have counteracted the general upward trend at least to some degree if the hypothesis is correct. This expectation is not supported. The results are mixed, but on the whole, increases tended to be larger for the gainers. A fair conclusion from these regression results is once more that, on average, research grant revenues had little or no effect on expenditures for instruction during the period.

Another possibility is that in response to financial pressures induced by grants, institutions sometimes raised tuition and sometimes reduced instructional expenditures, but did not consistently resort to either one. To explore this line, the quantity tuition change minus change in instructional expenditures was formed as the dependent variable. That way, either a frequent rise in tuition or drop in instructional expenditures, or both, in response to a rise in grant revenues would tend to produce a positive regression coefficient. A table similar to Table 1 shows, however, that most of the coefficients are very small and negative.

Thus, the analysis throws considerable doubt on the validity of the hypothesis that federal grants had perverse effects on tuition revenues and instruction expenditures. On average, institutions either felt little or no financial pressure at all because of their quest for research grants, or they did but were not strongly constrained to respond to that pressure by raising tuition or reducing expenditures for instruction. If there was a need to compensate, the average institution found other ways than these. This does not mean of course that research grants have not been costing universities money out of pocket. It is possible that the grants did not pay for themselves but rather that the shortfall was made up in different bits, major and minor, here and there over many expenditure categories and varying by individual choice both across institutions and over the years.

It should be remembered too that if research universities spend considerable funds for research that are not covered by external grants and contracts, much of this must be seen as part of the central mission of the institution, so that funds would be expended for this purpose even if there were no such thing as external grants. One may quarrel with the policy of spending money on research of any sort instead of devoting it all to instruction, but given that a research university will be a research university, the issue of whether to spend the money on externally supported or nonexternally supported research is surely a minor and esoteric one in broad public policy perspective.

Allowing for the possibility, however, that grants do not pay for themselves—and that they also lead universities to spend more than they otherwise would for research—it then becomes desirable to know what else these grant dollars seem to be doing to the universities so that stakeholders can judge whether it is worthwhile to try to maximize them.

Research quality of the faculty. Certainly, one of the important reasons why universities try to maximize their external research grants is to attract high-quality faculty. That goal in turn is seen as instrumental for the vital aim of maximizing the prestige of the institution (Feller 1995, 6-12, 34). In fact, policy toward sponsored research and the concern that it may carry a burden of internal financial support are important in large measure because, if grants also increase faculty research quality and university prestige, it would be difficult to curtail them voluntarily even if the cost burden were substantial. A university can fall behind in the prestige race, and this could make life difficult in uncountable ways (Feller 1995). A salient issue then is the impact of marginal changes in the level of grants on the research quality of the faculty.

Faculty research quality is difficult to measure, but data are available that allow a reasonable attempt. Reputational measures for this purpose are rejected here for two reasons. One is that their validity as indicators of quality initially must be considered doubtful. A small amount of light in fact will be shed on that question later in the article. The other is that they are important in their own right—as measures, that is, not of quality but of reputation or professional standing among those who supply the ratings. Another option is to base an indicator of quality on publications and citations. That is not perfect by any means because there are fads in research that do not stand the test of time, because people cite themselves and their friends more often than they should, because poor work gets cited often only to disprove or discredit it, because books are neglected in favor of journal articles, and for other reasons. Nevertheless, there is a core of validity to publication and citation measures that recommends their use in the absence of better bases, at least as elements of a more complex indicator.

The NRC produced a variety of ratings of faculty and program quality both in the early 1980s (Jones, Lindzey, and Coggeshall 1982) and the early 1990s (Goldberger, Maher, and Flattau 1995) that either contain or permit good indicators. One of their great advantages for a study such as this is that they covered essentially all doctoral granting institutions. In the former report, an indicator was included directly that was essentially the average quality of the journals in which faculty of the program published articles in 1978 and 1979. The quality of the journals in turn was determined by the frequency of citation of their articles in other journals in the field. The report called this variable "estimated overall influence of the articles," and it will be referred to in this article as "quality 1978-79." In the more recent report, this measure was not repeated, but data were presented both on publications of the faculty of the program and the number of times those were cited in the relevant journals. In addition, two Gini coefficients were published, one on publications and one on citations. The Gini coefficient, which norms between 0 and 1, is a measure of inequality. In this case, the larger is the coefficient for a particular academic program or department in a particular school, the more the publications or citations were concentrated in just a few members of the faculty. Other analyses (e.g., Katz and Eagles 1996; Jackman and Siverson 1996; Lowry and Silver 1996) have used these coefficients as stand-alone variables. They might be used instead, however, as coefficients—that is, as modifiers of the publication and citation data. In extensive exploratory work with a variety of measures, one seemed to stand out as being highly reliable and consistent and to make sense in terms of its results. It will be labeled *quality 1988-92*. It is the average number of citations per publication for the program in 1988 through 1992, multiplied by 1 minus the Gini coefficient for publications and multiplied again by 1 minus the Gini coefficient for citations. In this way, the citations per publication for a program are weighted by the extent to which both the citations and the publications were spread out among the entire faculty rather than concentrated in the work of a few. Once these calculations were made for each program in the social (excluding history), biological, and physical sciences and in engineering, a weighted average of the quality of the various programs in each institution was produced using the number of faculty in the program as the weight variable. (The programs or disciplines could not be analyzed separately for our purposes because adequate grant data were not available by program.)

On the grant side, the variable "grant change" was constructed by taking the average of federal research grant revenues in the years 1984 through 1987, just before the dates covered in the second NRC research quality measure, and

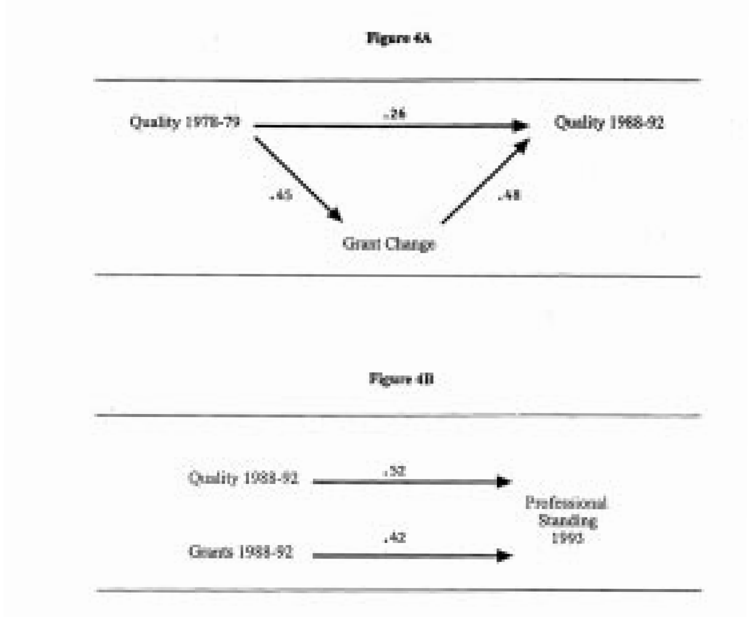


Figure 4a: Faculty Research Quality
Figure 4b: Professional Standing

subtracting from them the average for 1974 through 1977—just before the first NRC measure.

The issue is the extent to which this change in level of grants over the decade affected quality 1988-92, controlling for quality 1978-79. In real dollars (the appropriate basis when a psychological relationship is at issue rather than a simple economic progression or comparison), the coefficient obtained is $\beta = .48$. In terms of the standardized coefficients that were used, this means roughly that one standard deviation increase in grant support over the period led on average to a half standard deviation increase in faculty research quality ($t = 7.45$, $R^2 = .41$; see Figure 4a).[3] It appears then that research grant totals did matter substantially for faculty research quality. Spuriousness could be a problem here in that an institution might have decided to put money into attracting excellent research faculty, who then both obtained grants and published excellent work. However, the influential work still presumably would have depended on the grants, so that the relation is in that sense causal rather than spurious. In fact, the data do not permit one to pinpoint the manner in which grants operated—whether by leading to the acquisition of good new faculty who then received grants and published or by making the current faculty more productive and influential—but there seems to have been a decided effect one way or the other. Furthermore, the Gini coefficients directly restrict the impact of faculty stars on the quality measure unless they constitute a large proportion of the department, so that an institution would not contribute much to this spuriousness unless it bought not just one or two but a great many such stars during this period. Thus, there may be some spuriousness here from this source, but it is probably minor.

Several faculty stakeholders suggested that the effect of grant performance on salaries and on equipment is quite important. Obtaining grants raises salaries, thus enabling a university to attract and retain high-quality faculty, and brings in research equipment, especially the latest equipment, which has the same effect. One should look at salary and equipment, therefore, as subobjectives (Mohr 1995, chapter 3) or as mediating variables in the program theory (see Mark 1990 and sources cited there)—intervening outcomes that might need to be attained as a result of grant performance so that faculty research quality in turn might be affected.

There are some limitations in the availability of data on these variables in CASPAR; nevertheless, fairly reasonable measures could be constructed. The effect of a change in average grant revenues from the period of 1974 through 1976 to the period of 1984 through 1986 on average instructional salaries for 1987, controlling for instructional expenditures in 1977 (separate data on salaries do not begin until 1987) was β (standardized) = .09, $t = 2.5$, a modest relation. For the second link in the chain, the effect of 1987 salaries on the subsequent faculty research quality measure (controlling for grant changes, 1977 instructional expenditures, and quality 1978-79) was β (standardized) = .31, $t = 2.35$. Thus, the reported effect of grant performance on quality noted above does appear to be explained in some small degree by the effect of grants on instructional salaries as an intermediate link.

For the comparable equipment hypothesis, we look at the effect of growth in federal grant revenues between the period of 1979 to 1982 and the period of 1984 to 1987 on average federal equipment expenditures for 1985 through 1987, controlling for the comparable equipment expenditures for 1981 to 1982 (the data series on equipment expenditures begins in 1981). The relevant standardized coefficient is impressive, as one would expect: $\beta = .35$, $t = 8.18$. For the possible effect of equipment on faculty quality, however, the controlled coefficient is negative ($\beta = –.05$, $t = –.45$). Several different methods of getting at this relation also resulted in a weak negative coefficient. Taken at face value, this analysis suggests that better results in terms of influence of publications frequently came from the less well-equipped labs or, in other words, that equipment was no guarantee of success in terms of influential work.

Professional standing. As noted, because we have an independent measure of faculty research quality, the widely known reputational measure produced by the NRC in both periods, often treated as a quality indicator (especially in the popular press), can be used for what it really is—a measure of the professional standing of the program within its field. Again, a weighted average was calculated across the same disciplines to obtain a measure at the level of the whole institution. Of considerable interest here are the relative effects of obtaining grants and faculty research quality on national reputation. Feller (1995) refers to these concepts as financial competition and intellectual competition, respectively. Figure 4b shows the results of the regression of professional standing 1993 as the dependent variable on quality 1988-92 and average federal grants, 1988 through 1992. The controlled effect of grant performance in standardized terms was $\beta = .42$, $t = 8.67$. The controlled value for quality 1988-92 was $\beta = .52$, $t = 10.77$, R^2 for the three-variable regression $= .72$. Thus, these two sources of reputation had quite independent effects, both of which were strong. The reputational measure emerges with some apparent validity as an indicator of quality, although a far from perfect one. Of particular interest to us in this article, we see that the ability to get grants had considerable importance in determining national reputation.

Additions to endowment. The hypothesis is that the more successful the institution is in winning grants, the more successful it will be in attracting endowment funds, because good performance of any appropriate sort (not only athletics) is a selling point. A table similar to Table 1 but for additions to endowment was constructed to explore this possibility. There were more negative coefficients in the table than positive, although all coefficients tended to be small in probability terms (t values). Apparently, federal grant performance did not help much in the quest for endowment funds in this 7-year period. Further analysis showed, however, that the tendency came mostly from private institutions. In that group, the coefficients were negative 5 years out of the 7, both for the 1-year and the 2-year lags, whereas for the public institutions, they were positive 5 years out of the 7. The suggestion is that perhaps too many other things affected endowment acquisitions in private institutions, where their importance is so critical, whereas grants as selling points might have stood out just a bit more in the public institutions.

Private gifts, grants, and contracts. This category is similar, except that it composes additions to current-fund revenues rather than to the endowment. The efficacy of federal grants seemed marginally better here, tending to be positive four, five, and even six times out of seven, with the public and private institutions faring about the same. Possible spuriousness is a particular problem here, however, in that the same developments that led to increases or decreases in federal grant-getting capability might have led to similar changes in the ability to attract private funds.

The analyses of tuition, instructional expenditures, additions to the endowment, and private gifts, grants, and contracts indicate that research grants have not had either an appreciable or a reliable causal impact for the average institution on other components of university finance. These other areas are complex, depending on a host of considerations and forces. They apparently have not been determined in a major way by grant-getting performance.

Quality of incoming undergraduate students. This factor could be important for several reasons, including not only the prestige of the institution and the quality of instruction but also the level of donations in future years. The question therefore becomes pertinent: Do research grant levels affect the quality of the undergraduate classes that a school can attract? The entering test scores for undergraduates are readily available from publications such as the *Princeton Review* (1992) and *American Universities and Colleges* (1983, 1992). Moreover, they are available for several years back, so that it is possible to explore relations over a decade. In particular, the data permit a look at the impact of grant growth from the period 1976 through 1979 to the period 1988 through 1991 on the 1992 SAT/ACT scores, controlling for the same scores in 1980. Some of the institutions require one of these entrance tests and some the other, but a number of them—43 in 1992 and 36 in 1980—accepted either and had data on the averages for both. These allowed a regression to be run for the "effect" of ACT on SAT scores. The R^2 yielded was .91 for 1992 and .86 for 1980. To improve the usable sample size, the regression equations were used to transform all ACT averages into their estimated SAT equivalents. This permitted analysis covering about 80 institutions, depending on the particular variables included in the regressions. About seven fewer schools were available when only SAT scores were considered, but the results are almost identical.

These results indicate that federal grant performance made essentially no difference in the quality of incoming undergraduate classes (as measured by their standardized test scores). The relevant regression coefficients in the extensive exploratory analyses were very close to 0. For example, the coefficient on grant change in the regression mentioned above—1992 test scores on grant change, controlling for 1980 test scores—was .001, $t = 1.66$, $N = 83$. The t value in this case looks almost respectable, but it is affected by one outlier institution that gained more than $50,000,000 in federal grant revenues

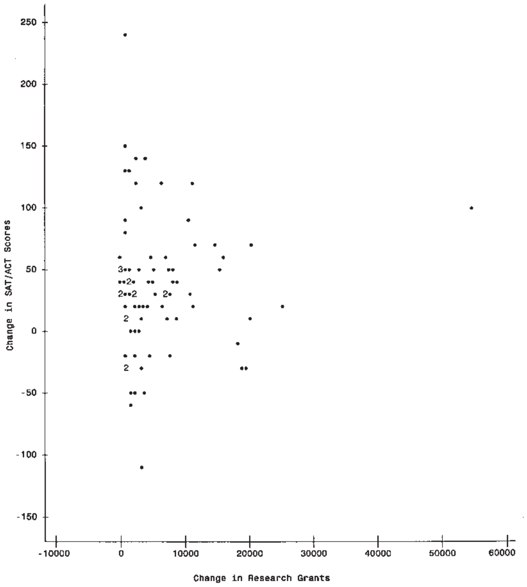


Figure 5: Change in Average SAT/ACT Scores From Change in Research Grants
NOTE: SAT = Scholastic Aptitude Test; ACT = American College Test.

over this period. Without that one school, the slope coefficient is .0006 and the *t* value is 0.65. A picture of the relation may be obtained by asking a slightly different but comparable question. Figure 5 shows the effect of change in grant performance on the change in total SAT scores from 1980 to 1992 (rather than 1992 controlling for 1980; see Mohr 1995, 116-119). It is abundantly clear from this scatterplot that although there were a great many instances of change in both grant levels and test scores over the period, such changes did not proceed in tandem.

A possible distortion could enter in that the test scores for some schools were already so high in 1980 that they had little or no room to grow, and these institutions might be exactly the ones with a strong growth in grants, undermining a positive correlation. The institutions with average SAT/ACT scores of more than 1,100 (which included our outlier), were eliminated, leaving 47 schools for the analysis. The results for the regression reported above were only marginally better: β = .002, *t* = 0.89. Moreover, the plot of change in scores on change in grants is still essentially triangular. One therefore may conclude that grant performance had little or no effect on the quality of undergraduates that these institutions were able to attract during the period covered by the analysis.

Quality of incoming graduate students. Although the results on undergraduates are not too surprising, one would strongly suspect that success in winning federal grants would have quite a direct and pronounced effect on the caliber of the graduate students an institution is able to attract. The measurement of this quality is of course not an easy or straightforward task, but GRE scores would seem to be a reasonable indicator for this purpose. Unfortunately and rather surprisingly, average GRE scores for incoming graduate classes in the various programs across the country are not collected by any central source. It is well-known that almost every program calculates them, but there has not been sufficient interest to build a national database from these records. A small amount of data is available from the source in note 2. Eliminating humanities programs and programs for which matching grant information is not collected, four disciplines were available for analysis: biochemistry, economics, mathematics, and mechanical engineering. The years of GRE data from this source covered 1989 to 1993, inclusive, so these 2 outside years were used to try to establish whether change in grant performance between the averages of 1985 to 1988 on one end and 1990 to 1992 on the other would match up with the GRE scores, the hypothesis being that the greater the increase in grant performance, even over so short a period, the greater is the increase in average GRE scores. Unfortunately, the participating schools with GRE scores for these disciplines did not match perfectly for the years 1989 and 1993, and of those that did match, some did not have the grant information for those disciplines for 1985 to 1992, so that the *N*s for our analyses end up being generally less than 15 programs in a discipline for any single regression.

The four disciplines available for analysis were not randomly selected and therefore cannot be considered representative of all, but they are fortuitous in one respect. Two of them, biochemistry and mechanical engineering, depend heavily on grants for faculty research and the other two, economics and mathematics, do not. One therefore may look for a positive effect of the ability to obtain grants on GRE scores in the first two and little or no effect in the second two. The regression results gauging the effect of change in the ability to obtain grants on 1993 GRE scores for the disciplines, controlling for the 1989 scores, look quite dismal for the success of the hypothesis. The slopes and *t* values are very small not only for economics and mathematics but also for biochemistry and mechanical engineering. It is more important in this case, however, to look at scatterplots rather than regression output because there are outliers and, with so few data points, the outliers can have a pronounced effect on the results. Inspection of the plots in Figures 6a and 6b shows a fairly clear tendency when bracketing the outliers, for average GRE

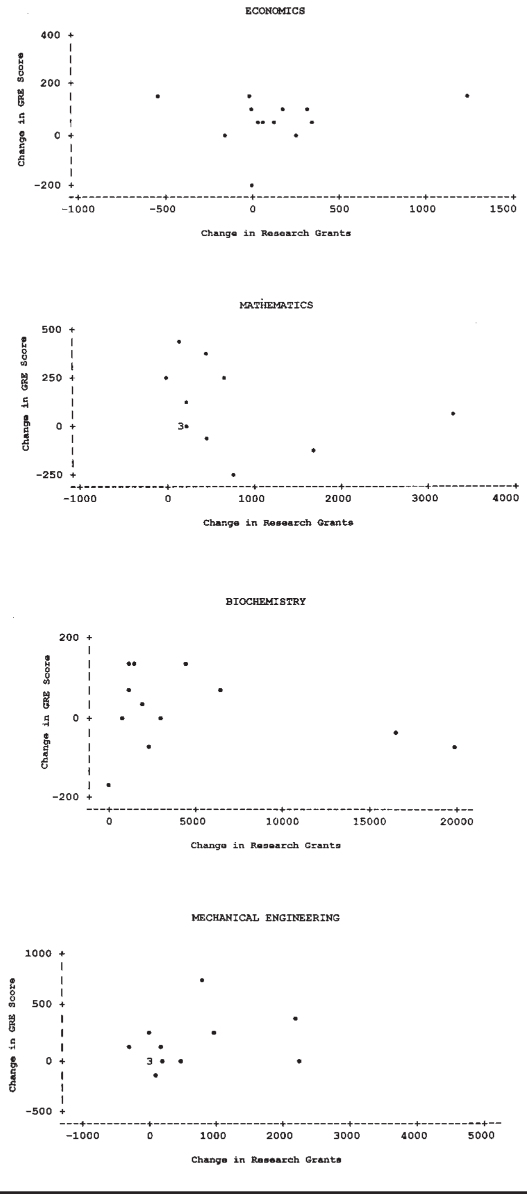


Figure 6: Change in Average GRE Scores From Change in Research Grants for Economics, Mathematics, Biochemistry, and Mechanical Engineering

NOTE: GRE = Graduate Record Examination.

scores to move reliably in concord with grant changes in biochemistry and mechanical engineering but not at all in economics and mathematics. Such results by no means can be considered strong evidence. Nevertheless, the evidence we do have does provide a modicum of support for the hypothesis.

SPECULATIVE ANALYSES

Graduate teaching. Is it the case that the more research dollars a program is able to attract, the better the job it does of graduate teaching? It would be difficult to devise a measure of this outcome. Perhaps the average standing of the universities or research units by which graduates were first hired would be a reasonable measure, but such data are not available and would be expensive to collect. It stands to reason that because the graduates are being trained for a research career, more grants would mean more opportunity to learn and practice in the latest research areas and therefore would produce better quality graduates (controlling, let us say, for entering GRE scores). On the other hand, the quality of their exposure to research and theory probably would have as much, and perhaps even more, impact as the quantity, and quality and quantity are not perfectly correlated as the data on grants and faculty research quality indicate. Therefore, one would expect to find a moderate relation between grant performance and value added in graduate teaching.

Undergraduate teaching. This is a highly sensitive area. Many consider that a research orientation detracts from the quality of undergraduate teaching because research faculty do not care as much about undergraduate students as do faculty in teaching colleges. Others consider that the opportunity for students to be exposed to research findings and to research minds and even to participate in research substantially enhance the quality of undergraduate teaching in research universities, especially in the junior and senior years.

The problem with this debate is that there is no agreement on how to measure the quality of undergraduate education. This would seem to be a highly lamentable lack, given that there is so much interest in the topic. However, a direction for the solution to this problem of conceptualizing quality is suggested by the theory of valuing defended in this article. Several dimensions have been prominently mentioned by stakeholders as being critical for the assessment of value added by an undergraduate education. These include standardized test score performance, the extent to which students have learned the subjects they have been taught, the ability to think critically, the ability to speak and write articulately, and the motivation toward lifelong learning. It follows then that measurements of such outcome dimensions might be developed where they do not already exist and applied to a range of colleges and universities. Some observers then might be interested in the evaluation of particular institutions and would want to look at a profile of outcomes on these dimensions for each institution of interest. For purposes such as the analysis in this article, on the other hand, one would want to group institutions into those with strong versus weak research orientations, for example, and compare profiles in the two categories. Stakeholders then would use their own informal weighting schemes to arrive at conclusions regarding the impact of a research orientation on the quality of the education and bring those conclusions to the policy debate, whether it be within a university or in a state or national forum.

There are some data that might help us at least in a minor way in our present task. Astin (1993) measured research orientation of the faculty—which is of course not the same thing as grant performance—and several dimensions that might be affected by this variable. Astin's conclusion is, "Institutions . . . that heavily emphasize research tend to produce a generally negative pattern of student outcomes" (Astin and Chang 1995, 45) or, "These results show clearly that, with the exception of performance on standardized tests, there is a significant institutional price to be paid, in terms of student development, for a very strong faculty emphasis on research" (Astin 1993, 338).

What is interesting is that these conclusions depend heavily, as this article and scholars cited in it have maintained, on how one values the various outcomes. In fact, research orientation in Astin's 1993 data apparently had strong positive effects on GRE verbal and quantitative scores and on the Law School Admission Test (LSAT) scores. This same orientation had negative effects on trust in the administration; satisfaction with faculty, with the overall quality of instruction, with the overall college experience, and with individual support services; self-rated leadership ability, popularity, and social self-confidence; completion of the bachelor's degree; growth in interpersonal skills; graduating with honors; college grade point average; and student orientation of the faculty (a measure of caring about the personal well-being of students, not about their academic growth) (Astin 1993, 338). There is clearly a pattern here. Some critics presumably would value the test scores over all the rest and therefore conclude that a research orientation of the faculty is a good thing, although one might support some tinkering with local policies to try to improve retention (completion of the bachelor's degree) and perhaps a few other items.

Another consideration is important in this context. The policy question for most research universities would be whether to rein in their grant-seeking efforts to some extent, not to eliminate them almost entirely and become a teaching college. In that perspective, it is unlikely that fewer grants would lead to better undergraduate teaching. In particular, if fewer grants meant a diminution in the research quality of the faculty while still preserving an overall research orientation, then one might lose a little in the advantages of grants while not gaining anything in the advantages of a teaching orientation.

This analysis suggests that marginal changes in grant performance probably would relate positively to changes in the quality of undergraduate instruction as measured by GRE and LSAT scores. How obtaining a grant would relate to the other major criteria mentioned above (critical thinking, etc.) is a matter of great interest and, at this point, is anybody's guess.

Distortion of academic power structures. The concern would be that the more the institution were oriented toward getting grants, the more would power in the disciplines—department chairmanships, editorships, tenure committee chairmanships, and so forth—be determined by grant-getting prowess rather than by quality of research and of theory production. It may be that a grants orientation does have this effect on power. In evaluation perspective, however, one must be concerned with the counterfactual. What would be the case without a grants orientation? Would power be more likely to flow to the best scholars? It is difficult to find experience with which to answer such a question but given that the best scholars so often seem to avoid positions of power whereas others seek these positions out, it is doubtful that decreasing the grants orientation would move us toward the power distribution favored by those who have this concern.

Depth and innovation in research. The concern here would be that the more the institution were oriented toward getting grants, the more superficial, technical, or like "normal science" the research activity would be, in place of more thoughtful, theoretically powerful activity. Perhaps if an institution were oriented more toward depth and innovation than obtaining grants, for example, there would be both fewer grant applications and average work of better quality and greater influence. More will be said about this in a future publication from an analysis of residuals in certain of the faculty quality regressions, but for now, it can be reported that there is at least some validity in the concern. On the other hand, making that sort of change in the prevailing research orientation within a given institution may be much easier said than done.

CONCLUSION

Figure 1 shows the impacts discovered on the full range of outcomes included in the analysis. Rather than present the regression coefficients, it seems more suitable in this case, where more than one coefficient was used for some of the criteria and the data in many cases were not as complete as one would like, to place just a plus, a minus, or a zero, depending on the direction of the impact. A plus for faculty research quality, for example, means that the higher the grant growth, the higher is the growth in quality. It is important to note that these pluses and minuses do not indicate good and bad—that depends on how one would value the individual impacts.

The following is a brief summary of the results for ease of reference. Such a summary is a critical element in implementing the impact-profile approach.

Tuition. Tuition and research grants rose steeply in tandem beginning in the early to mid-1980s, but the evidence seems to deny unequivocally a causal impact of the grants. On average across the institutions (see Table 1), there is no evidence for pressure on tuition from sponsored research. There were tuition gains after especially high grant increases, but the same occurred after especially large grant losses.

Instruction. As with tuition, there was apparently no impact of grant-getting performance on expenditures for instruction.

Research quality of the faculty. There seems to have been a definite effect of grant-getting performance on faculty quality, measured in terms of publications, citations, and the extent to which these were spread out among the whole faculty of a program rather than concentrated in a few stars.

Salaries. Salaries for instruction were affected positively to a modest extent by grant success. There is also evidence that the salary changes translated into changes in faculty quality.

Equipment. Federal grant success correlated well with federal equipment expenditures. The latter, however, had no relation to faculty research quality.

Prestige or professional standing. Both grant-getting performance and the research quality of the faculty were strong and independent predictors of professional standing, as measured by the careful and well-known reputational ratings published by the NRC.

Additions to the endowment fund. Grant-getting performance had little or no effect, although the signs of the relations tended to be more positive among public than among private institutions.

Private gifts, grants, and contracts. Grant-getting performance probably had little or no effect.

Quality of incoming undergraduate students. Grant-getting performance appears to have had zero impact on average entering SAT/ACT scores.

Quality of incoming graduate students. The time period covered by the analysis was relatively short and the sample sizes were exceedingly small, but growth in federal grant revenues appear to have had a positive impact on average entering GRE scores in disciplines in which such an outcome would be expected.

Quality of graduate teaching. There are no statistics. A moderate positive effect is most likely.

Quality of undergraduate teaching. There are no statistics. A small positive effect of marginal changes in grant-getting performance is most likely, depending to a certain extent on how *quality* is defined. A positive effect of faculty research orientation on standardized test scores has been documented in other research.

Distortion of academic power structures. There are no statistics. It is probably true that power structures are affected by grantsmanship so that scholarly quality does not become the primary determinant of power. It is unlikely, however, that scholarly quality would go on to prevail if grantsmanship became less of a factor.

Depth and innovation in research. There are no statistics, although some data will be presented in a subsequent report. An orientation toward maximizing research grants may have a small negative impact on this factor, suggesting that universities might want to reconsider the role that grants play on campus relative to other determinants of quality.

The conclusions are first that there is no profit to be gained in trying to put all of these impacts on a common scale of dollars or a scale normalized by plausibility or tolerability anchors. If forced, one certainly could come up with answers to the questions about dollar values or anchoring points, but they would probably seem like guesswork and be subject to revision once the value implications of the final summing up were in view. For some of our important measures, such as the research quality of the faculty for example, the effort would be particularly unappealing because, with Gini coefficients and so forth, people simply do not think in terms of such scales and cannot readily interpret the numbers for conversion into dollars or decisions as to anchor points. The standardized regression coefficients and *t* values, however, are meaningful.

Weighting seems even more futile. There are too many outcomes, too many interactions and curvilinearities, and too many points of view. For example, the importance attached to instructional expenditures might be very small if the impact were positive but much larger if the impact were negative. The importance of salaries as an outcome, and even whether to weight this outcome positively or negatively, might depend entirely on the impact of salaries on faculty research quality and so forth (e.g., if grants raised salaries but these had no impact on faculty research quality, the pressure on salaries could be negatively valued by some stakeholders). What is the importance of depth and innovation as an outcome? It is unlikely that more than a handful of stakeholders could put a monetary or importance value on this criterion with even moderate confidence in the judgment—at least not until all the data were in and the results were being considered in their entirety.

Furthermore, the effort needed to change the reported quantitative impacts from regression coefficients into dollar values would be as great again for the author as already has been expended on this article. And to what advantage, because the proper transformations would be in doubt and the quantitative impacts themselves are approximate? Alternatively, one might have presented several illustrative sets of weights, primarily to demonstrate how this sort of thing is done (it is not as simple as it might seem; see Edwards and Newman 1982; Scriven 1994). First, such a presentation has no value for summing up without common scaling, which is also costly in time and effort. Second, the more one were to introduce illustrative curvilinearity and interaction, the more the example would tend to become bogged down in complexity; but the less these two complications were featured, the more oversimplified and misleading the example would be. Moreover, such illustrations would not settle the issue of value but would only serve as an invitation to individual stakeholders to go through the elaborate and difficult quantitative exercise for themselves, using their own weights and arriving at their own conclusion on value. They are free to do so—even without finding an illustration of the method in every evaluation report—but few actually would do so because we rarely (not never, perhaps, but rarely) go through such an exercise for important, multiple-outcome decisions in life. The reason, as argued above, is that such common scaling, weighting, and summing on our own would neither feel nor be constructive.

On the other side, human beings can come to conclusions on merit without the trappings of systematic rationality. We are in fact expert in making decisions when the elements involved are incommensurable. We do it all the time. Reaching such conclusions regarding what house to buy, what college to go to, what employee to hire, and so forth sometimes takes a little time, thought, debate, and evolution, but not only can it be done without the numbers, a good case also can be made that it really cannot be done in any other way. The numbers that truly matter are those quantitative impacts that give the stakeholder or decision maker some better idea than pure impression of what the degree of causal connection between major variables has been in the past, as, for example, the data on professional standing and on entering graduate and undergraduate test scores reported here.

Instead of common scaling and weighting, the method that should be used by stakeholders is to read over a summary of results, such as that presented above, several times, referring back to the longer descriptions in the evaluation report when interest or uncertainty make that desirable. After a few such readings, the list is conquered. It is not unmanageable for the human mind. What is important and unimportant to the individual stakeholder soon sorts itself out. After a period of reviewing and re-reviewing, a sense of the relative worth of the alternative program options (e.g., trying to maximize grants vs. restraint or selectivity of some sort) will take hold. If it does not—if one has trouble making up one's mind—the policy debate will probably help to develop or crystallize one's view. It will be most helpful, however, to keep the summary list of impacts at the top of one's stack of notes during the discussions and to refer to it frequently as the group's as well as one's own views move toward closure.

Paradoxically, impacts help to determine goals (March 1978). Thus, the time to apply weights as a stakeholder is after the data are in. That way, many of the contingencies upon which the choice of weights might depend—the curvilinearity and interaction problems—have become reality, and one does not necessarily have to think about value structures for other possible realities. The purpose of explicit weighting at the end rather than the beginning, however, would not be so much to help determine one's view of the overall value of the program or policy, but rather to reveal something about oneself (i.e., to get some insight into what one thinks about the relative importance of the various program outcomes, such as cost versus faculty research quality, under one real set of circumstances).

To conclude, it has been the contention of this article that the policy analyst, in providing these impacts in their entirety, performs a critical service in the necessary process of valuing, but that going further than providing the impacts is both unjustified and unavailing.

NOTES

[1] Quantum Research Corporation is located at 7315 Wisconsin Avenue, Suite 631W, Bethesda, MD 20814; phone: 301-657-3070; fax: 301-657-3862. I gratefully acknowledge the contribution of this database to the research. It is an invaluable resource for research on higher education.

[2] The Association of American Universities/Association of Graduate Schools, Project for Research on Doctoral Education, Educational Testing Service is located at Rosedale Road, MS 19-T, Princeton, NJ 08541. I very much appreciate the project's willingness to make the data available to me in appropriate summary form, including the special preparation of diskettes and printed tables.

3. Standardized coefficients are used here in all analyses involving faculty research quality because, given the complicated measurement scale that involves ratios and Gini coefficients, the results using unstandardized coefficients are essentially uninterpretable, both in absolute terms and relative to other coefficients. The primary problem with standardized coefficients, that they depend on local variances and therefore render the results specific to the data set at hand, is mitigated somewhat by the fact that these data are considered a self-contained population in any case. The unstandardized coefficient of quality on grant change in millions of dollars is .05. The mean of the quality scores is 3.6 and the standard deviation is 2.4. Thus, a rise of one standard deviation in grant change, $25 million, led to about half a standard deviation increase in quality or 1.2.

REFERENCES

American Council on Education. 1983. *American Universities and Colleges*. Vol. 12. Washington, DC: Author.

American Council on Education. 1992. *American Universities and Colleges*. Vol. 14. Washington, DC: Author

Astin, A. W. 1993. *What matters in college: Four critical years revisited*. San Francisco: Jossey-Bass.

Astin, A. W., and M. J. Chang. 1995, September/October. Colleges that emphasize research and teaching: Can you have your cake and eat it too? *Change* 44-49.

Blackorby, C. 1990. Economic policy in a second-best environment. *Canadian Journal of Economics* 23(4): 748-71.

Clotfelter, C. T. 1996. *Buying the best: Cost escalation in elite higher education*. Princeton, NJ: Princeton University Press.

Cronbach, L. J. 1982. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Edwards, W., and J. R. Newman. 1982. *Multiattribute evaluation*. Beverly Hills, CA: Sage.

Feller, I. 1995. The determinants of research competitiveness among universities: A critical review of the issue and the literature. Paper prepared for the AAAS Conference on Assessing Research Competitiveness, April 21-23.

Geiger, R., and I. Feller. 1995. The dispersion of academic research in the 1980s. *Journal of Higher Education* 66:336-60.

Goldberger, M. L., B. A. Maher, and P. E. Flattau (eds.). 1995. *Research-doctorate programs in the United States: Continuity and change*. Washington: National Academy Press.

Grossman, R., and C. Leroux. 1996. Research grants actually add to tuition costs, study claims. *Chicago Tribune*, January 28: 1, 15.

Jackman, R. W., and R. M. Siverson. 1996. Rating the rating: An analysis of the National Research Council's appraisal of political science Ph.D. programs. *PS: Political Science and Politics* 29(2): 155-60.

Jones, L. V., G. Lindzey, and P. E. Coggeshall. 1982. *An assessment of research-doctorate programs in the united states*. Washington, DC: National Academy Press.

Katz, R. S., and M. Eagles. 1996. Ranking political science programs: A view from the lower half. *PS: Political Science and Politics* 29(2) :149-54.

Kopp, R. J., W. W. Pommerehne, and N. Schwarz, ed. 1997. *Determining the value of non-marketed goods*. Boston: Kluwer.

Lowry, R. C., and B. D. Silver. 1996. A rising tide lifts all boats: Political Science Department reputation and the reputation of the university. *PS: Political Science and Politics* 29(2): 161-67.

March, J. G. 1978. Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics*, 9(2): 587-608.

Mark, M. M. 1990. From program theory to tests of program theory. In *Program theory in program evaluation: New directions for program evaluation*, ed. L. Bickman, 47:37-51. San Francisco: Jossey-Bass.

Mohr, L. B. 1995. *Impact analysis for program evaluation*. 2nd ed. Newbury Park, CA: Sage.

———. 1996. *The causes of human behavior: Implications for theory and method in the social sciences*. Ann Arbor: University of Michigan Press.

Princeton Review. 1992. *Student access guide to best colleges*. New York: Villard Books.

Rosenthal, D. H., and R. H. Nelson. 1992. Why existence value should not be used in cost-benefit analysis. *Journal of Policy Analysis and Management* 11: 116-22.

Scriven, M. 1994. The final synthesis. *Evaluation Practice* 15: 367-82.

———. 1995. The logic of evaluation and evaluation practice. In *Reasoning in evaluation: Inferential links and leaps: New directions for evaluation,* ed. D. M. Fournier, 68:49-70. San Francisco: Jossey-Bass.

Shadish, W. R., Jr., T. D. Cook, and L. C. Leviton. 1991. *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Tribe, L. H. 1972. Policy science: Analysis or ideology? *Philosophy and Public Affairs* 2:66-110.

Lawrence B. Mohr is a professor of political science and public policy at the University of Michigan. His major areas of research interest are organization theory, program evaluation, and the philosophy of social research. Within program evaluation, he emphasizes evaluation theory and methods.