Peer Review: Quality Control of Applied Social Research

Peer review needs buttressing and extension as a means of improving the adequacy of policy-related research.

John H. Noble, Jr.

There is a great ferment these days in the science community because of cutbacks in research funds and changes in the ways federal agencies do business. As federal agencies, following Etzioni's advice (*1*), began to shift funds away from basic research and proposals initiated by scientists outside of government to agency-directed studies of a very applied character, peer review came under a cloud. The question was raised whether peer review was an effective method of judging the scientific relevance and quality of research. This apparent insult to the reputation of peer review provoked a sharp reaction from the science community (*2*). And while the debate continues inside and outside of government, little has been done to document the strengths and weaknesses of peer review as a process leading to value judgments about the relative worth of investigations purporting to yield objective knowledge about the "state of the world." Arguments back and forth tend to get confounded with the issue of procurement by grant versus contract. Protagonists of contracts *without* peer review argue that the science community is insensitive to the applied research needs of agencies now impeded from achievement of their missions because of the lack of elementary know-how. "Robin Hooding," they claim, turns many grants for applied research toward the more basic research preferences of scientists. Those favoring grants *and* peer review counter with the argument that contracts receive less rigorous scrutiny than grants under the system of peer review and thus give the public less value for its money.

Dr. Noble is an adviser to the Secretarial Task Force on Disability, Department of Health, Education, and Welfare, 330 Independence Avenue, SW, Washington, D.C. 20201.

It is unfortunate that the debate has become confused with arguments over preferred methods for procuring research—basic or applied. What is important are the criteria for judging research, especially those criteria relating to methodological adequacy. No matter how relevant is the research to government policy, it cannot provide answers unless it is properly designed and executed. There is mounting criticism about flaws in investigations concerned with major policy issues, such as the effects of psychotherapy (*3*), social work intervention (*4*), compensatory education (*5*), performance contracting (*6*), busing (*7*), and the effectiveness of schooling in general (*8*).

This occurs at a time when the amount of untested theory and information competing for the attention of policy-makers boggles the mind. In a period of noticeably diminishing resources, the nation can ill afford trial-and-error approaches to the solution of social problems. It seems not only wasteful but irresponsible to continue attacking problems by mounting long series of "demonstrations," each proclaiming self-prophesied "success." These and other studies that mislead because of error in design, procedure, or inference are a menace to policy-makers and the commonweal.

In this article I discuss several issues which influence deliberations leading to value judgments about the worth of applied studies, especially "policy" or "evaluative" research. In theory, administrators and policy-makers rely upon the products of these investigations to make choices among alternative courses of action. Such investigations are supposed to reduce the policy-maker's uncertainty about specific policy choices—the consequences of past choices or those about to be made in order to affect the future. I shall present empirical evidence concerning the strengths and weaknesses of several variants of peer review as methods for judging, in terms of the factors that influence internal and external validity, the scientific relevance and quality of applied social research.

I maintain that peer review needs strengthening in two ways. First, the use of formal standards and associated reproducible measures of methodological adequacy would guide review panels in their deliberations and make explicit the grounds for judgment. Second, if, in addition to reviewing project proposals and appraising the progress of projects being considered for refunding, peers were to evaluate completed research, it would be easier for them to determine whether research being funded was answering specific questions. In this way, peer review would assist potential users of the research results, including the administrators and policy-makers who authorized funding in the first place, in making their decisions on how the results should be used. Peer review which stops short of final assessment of the research product—letting the "buyer beware"—seems wanting.

Uses of Research and Development Funds

The uses to which research and development (R&D) funds are put creates some ambiguity and confusion when an effort is made to apply explicit evaluative criteria for determining the worth of specific projects. Funds allocated by Congress for R&D purposes are frequently used by agencies not only for generating knowledge, but also for subsidizing the services of their clientele and for seeking the influence and support of important constituencies. To quote Orlans, ". . . social research is not and . . . cannot be strictly apolitical, so its use or nonuse by government officials cannot be understood satisfactorily in apolitical terms" (*9*). R&D funds are used to play several kinds of politics: personal, professional, tactical, and party. Almost conspiratorially, Orlans observes (*9*, p. 32):

> That personal considerations influence the award of funds, the appointment of consultants, and the degree to which advice is not only heard but heeded is known by all but confessed by few during their active careers, since these matters are supposed to be decided solely on professional merit.

By identifying the political issue we can channel discussion of appropriate evaluative criteria to the knowledge-building function of the R&D enterprise. Criteria for evaluating political utilities are obviously different.

What characterizes policy-oriented research is the quest for cause-effect relationships. This is as true for simple case studies as it is for more complicated analyses of time-series, descriptions of correlations among variables in one-time surveys, or direct manipulation of stimuli in experiments. What were, are, or will be the effects of specific policies or interventions? On whom, by whom, where, when, and how? Investigators fully cognizant of the limitations of their descriptive studies usually cannot avoid their own recourse at some point to interpretations that ultimately depend upon causal inferences. The requirements of producing a final report, the formal thinking processes of the investigator, and the expectations of the readership conspire to force causal interpretations.

In view of this discussion the criteria for judging the performance of peer review as an objective evaluative procedure seem clear. Peer review must (i) apply the standards of the scientific method in reaching decisions about the methodological adequacy of research projects in relation to the questions they propose to answer, and (ii) eschew political agendas and biases of every kind.

Threats to Inference

Determining whether knowledge-building R&D projects are capable of answering the questions they propose to answer is the proper sphere for peer review. Methodologies and the degree to which they are implemented ultimately define the probabilities that specific questions have been answered. Alternative study designs can be assessed in terms of the factors they introduce to threaten the internal and external validity of findings. Internal validity is concerned with the question, "Did conditions (policies) of the experiment significantly affect the results?" External validity asks an additional question, "In what other situations would the same policies have the same results?" (*10*).

These threats to validity also qualify the inferences that may be drawn from nonexperimental investigations, such as cross-sectional surveys, and longitudinal and panel studies. Contrary to the beliefs of some social scientists, the same logic that is used in the laboratory can be used to assess the validity of field studies. Campbell has argued convincingly on this point (*11*).

It is possible, in fact, to classify most of the literature dealing with methodology by reference to the specific threat or threats to validity either addressed or neglected by application of given techniques. Sampling theory (*12*), measurement of reliability and test structures (*13*), and significance testing (*14*) all relate to instability as a threat to internal validity [see (*10*)]. Sampling theory is also concerned with threats stemming from selection biases of various types as well as with threats to valid generalization. Studies of several classes of response sets, including "yea-saying," "nay-saying," and the social desirability response set, challenge certain types of instrumentation and the interpretability of resulting measurement (*15*). Not only has optimal research design been defined, but so have the consequences of deviating from normative procedure (*16*); and substantial errors of practice are known (*16*) that permit incorrect inferences, or produce basic threats to the scientific principle of intersubjectivity: the requirement that independent observers be able to perform identical procedures on the same empirical phenomena and thereby arrive at identical results (*17*). Monte Carlo and other simulation methods have been used to demonstrate common errors of inference associated with certain research designs and procedures (*5*).

The science community does an unsystematic job of applying what is known to the evaluation of studies whose authors seek publication—with some amount of slippage occurring as a result (*18*). Peer review is responsible for some portion of the compromised studies being conducted in the first place; the rest undoubtedly gets funded as a result of flawed contracting procedures, or the political process which enables agencies, public and private, controlling access to the information needed by the investigator to throw up barriers to successful execution of study designs.

Recent reports document the magnitude of the problem. Bernstein *et al.* discovered that only 10 percent of 152 comprehensive evaluation projects funded by agencies of the federal government in fiscal year 1970 met minimum scientific standards (*19*). Minnesota Systems Research found, in a probability sample of 179 of 532 projects considered by staff of the responsible funding agency to have at least some research component, that only 6.7 percent were completely able to achieve their stated objectives; another 34.1 percent were judged to hold reasonable promise of doing so in face of time, resource, or study constraints beyond the investigator's control (*20*). To use a homely analogy, more than 90 percent of the eggs are getting broken between the hen house and the store, but the carriers are not to blame for nearly three-eighths of the breakage!

Effectiveness of Peer Review

The Department of Health, Education, and Welfare recently asked Minnesota Systems Research to study the effectiveness of four versions of the single panel peer review (SPPR) method in terms of the ability of senior scientists to discriminate among projects of varying methodological adequacy (*21*). Fifteen judges were asked to rate six projects, first, on 67 dimensions of research methodology and, second, with respect to their overall excellence. The judges reached their conclusions after reviewing project proposals, progress reports and, in some instances, after meeting with the principal investigators. The dimensions rated encompassed such matters as:

1) *Design.* The logic of the research; conceptualization; kinds of contrasts set up; appropriateness of assumptions, definitions, and the use of literature; and overall technical structure of the project.

2) *Sampling.* The way cases were selected; definition of the research case, population, and sample; handling of concerns about representativeness and generalizability of findings; and randomization.

3) *Statistics.* The selection, appropriateness, use, and interpretation of statistics, both descriptive and inferential; explanation of the statistics used; avoidance of artifactual distortions; and use of controls.

4) *Checking.* The handling of validity and reliability issues; examination for biases; pretesting of procedures and instruments; care and workmanship in handling data and analysis; and discussion of assumptions.

5) *Reporting.* Presentation of preliminary findings and progress of the research; readability of interim reports;

discussion of implications; inclusion of all data relevant to points of discussion; and objectivity.

Four versions of the SPPR method were evaluated. They were defined by combining and permuting two factors: (i) whether or not panel members prerated and preranked projects at home before meeting as a panel and (ii) whether or not they met as a panel with the principal investigator in Washington, D.C., to receive a briefing and ask questions about the current status and future plans of each project. All panels but one consisted of four members randomly drawn from a master list of senior scientists, stratified by reference to their previously rated tendency on a single benchmark project to give harsh versus lenient assessment of overall research quality (*22*). Thus, the design of this evaluation makes it possible by means of analysis of variance to discern the effectiveness of the SPPR method in terms of four main effects and their interactions. The statistical model for the analysis of variance can be expressed for the specific score, y, of an individual project, as follows:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl}$$

where μ is the overall mean, α_i is the effect due to the panel member meeting grantee, β_j is the effect due to the panel member having prerated the project at home, γ_k is the effect due to project differences, δ_l is the effect due to judge strictness, and ε_{ijkl} represents the residual error (*23*).

It is important to note that none of the four versions of the SPPR method evaluated here exactly fits the description of ordinary peer review as conducted by agencies funding research. First, agencies do not use the randomization principle as the basis for selecting peer review panels from some predefined group of experts. The resultant agency panels, therefore, are subject to many biases which stem from the political process leading to their creation. Second, ordinary peer review is not tightly constrained to address systematically the many methodological issues that various types of research can raise by a structure, process, or set of instruments comparable to those employed in this evaluation. Instead of exact scores, panel judgments typically take the form of written commentary by individual panelists and a gross ranking of project proposals or progress reports according to merit or priority for funding, or both. Third, ordinary peer review makes use of site visits, when needed, to obtain supplementary information about projects under review instead of having tightly controlled briefing sessions with principal investigators.

Table 1. Mean global and mean composite scores given to six projects by judges asked to rate their methodological adequacy. Prior scoring of a single benchmark project by judges resulted in their assignment to a strictness level: 1, far better than average; 2, somewhat better than average; 3, about average; 4 or 5, somewhat below or far below average [adapted from (*21*), pp. 11 and 16].

Item	Mean global assessment score	Mean composite score
Meeting with grantee	2.68	3.07
Meeting without grantee	3.52	2.51
Prior rating	3.06	2.84
No prior rating	3.14	2.73
Project:		
201	1.81	3.60
202	3.37	2.80
203	3.0	2.87
204	3.56	2.49
205	4.68	1.78
206	2.18	3.19
Judge strictness level		
1 (lenient)	3.08	2.90
2	3.12	2.72
3	2.95	2.70
4 or 5 (strict)	3.25	2.83

With the foregoing differences being kept in mind, ordinary peer review can be described as that version of the SPPR method which permits panel members to study project materials before they meet as a panel and which may or may not permit panel members access to the principal investigator for supplementary information. Site visits to obtain supplementary information are not always considered necessary. It seems reasonable to assume that, in the context of this evaluation of peer review, a meeting with the grantee during a briefing session is a good substitute for a site visit. Any biases developing as a result of meeting principal investigators would probably be more pronounced under the less controlled conditions of a site visit. By evaluating the effectiveness of the SPPR method under controlled conditions, it might be possible to establish the limits of peer review as ordinarily conducted. Ordinary peer review, subject as it is to many more extraneous influences, probably allows greater error and explains less variance among ratings and raters in relation to methodological issues than the variants of the SPPR method investigated here.

In presenting results, two methods for obtaining judgments about research quality are contrasted: (i) the single global assessment versus (ii) a composite judgment based upon separate ratings of 67 methodological issues. To reach a global assessment, judges were asked to rate each of six projects on a five-point scale ranging from 1 "far better than average" to 5 "far below average," after considering the current state of research in the areas and settings addressed by the project. In contrast, the total composite score for each project is the unweighted sum of the separate ratings, reversed so that 1 means, "largely unacceptable, major errors, flaws" and 5 is for "acceptable, far above average," for as many of the 67 methodological issues as the judge considered relevant in forming an opinion. Judges thus weighed not only acceptable methodologies and current constraints for the type of research being evaluated, but also the applicability of each of the 67 methodological considerations as (i) important to include, (ii) optional but relevant as a possible alternative procedure, or (iii) irrelevant. Two nonresponse categories, consistent with this rating system, were also available: N, "research does not have this feature or characteristic" and U, "unable to tell, insufficient information."

Projects which did not have methodological features considered by judges as important to include were rated for the specific features involved according to the system for obtaining total composite scores. Absent but optional features, as well as important issues about which information was insufficient, were not scored, thus always giving the benefit of the doubt. Finally, because of reversed meanings ascribed to high and low values under the two methods for obtaining judgment, high global assessments convey below average quality in contrast to above average for high composite scores.

As shown in Tables 1 and 2, project differences and meeting the grantee by themselves had statistically significant effects ($P < .001$) in single global assessments of research quality. The interaction between the effects of meeting the grantee α_i and project differences γ_k was also statistically significant ($P < .05$). The average ratings given to the projects covered much of the possible range and clearly discriminated among projects. Judges who had met principal

investigators tended to rate projects more favorably than those who had not. There was also a significant interaction effect ($P < .05$) due to judges meeting grantees for specific projects. Prior rating of projects and judge strictness did not have statistically significant effects.

As Tables 1 and 3 indicate, composite scores which take all 67 methodological considerations into account again show that project differences and meeting the grantee had statistically significant effects ($P < .001$), and so at a lesser level ($P < .05$) did the interactions of (i) meeting the grantee α_i and project differences γ_k, (ii) meeting the grantee α_i and judge strictness δ_l, and (iii) prior rating β_j and project differences γ_k. As with global assessments, judges who had met principal investigators tended to rate projects more favorably than those who had not.

All the interaction effects are easily explained. Each project had a different principal investigator. Consequently, the presentation of both oral and written materials differed in style as well as content. It is reasonable to expect the manner and personality of given principal investigators to strike judges in different ways. Apparently the impressions gained from prior rating of some projects are sufficiently salient to carry over and affect final judgments.

In view of these findings, how effective is the SPPR method as a means of judging the quality of applied social research? That approximately 44 percent of possible variance in both global assessments and composite scores can be explained by project differences implies a verdict of "moderate" to even "considerable" effectiveness. Seldom does the main predictor in social research account for more than this amount of total variance. On the other hand, the verdict becomes clouded when consideration is given to the fact that another 13 percent of the variance in global assessments and 20 percent in composite scores can be explained by statistically significant biasing factors.

But why qualify the power of the SPPR method when it would be more diplomatic and in line with conventional wisdom to gloss over possible weaknesses while emphasizing its virtues? Besides, the results of this experiment must be considered tentative until replicated. Clearly, the panelists were able to discriminate among the six projects and, regardless of underlying differences of opinion, rank them according to their methodological adequacy (*24*). What more could a funding agency desire than knowledge that the experts it convened to evaluate a batch of projects had reached agreement on their overall worth? In simplistic terms, all that remains for the agency to do is to allocate funds to the projects according to their rank until the available funds are exhausted. Later, it can use the same procedure to reappraise projects subject to refunding decisions.

Logic and evidence suggest that the type of peer review which only yields graded batches of projects is not sufficient. First, how can one panel's judgments be compared to another's unless both adhere to the same standards and deliberative process? The agency whose funding policy calls for support of top-ranking projects which qualify in terms of some panel's rating and the agency's test of "relevance" is by no means assured that its investment will bring anything near equal quality (*25*). Furthermore, who, when faced with different sets of standards, would want to be judged by a harsher set? Why not shop around for the easiest "touch" when systematic differences in standards among agencies are discernible and there is a choice?

Second, the observed behavior of panelists who assume the role of "substantive expert" to apologize on the grounds of field constraints for methodological inadequacies tends to soften the judgments of other panel members (*21*, p. 34). The tendency is reinforced to the extent that panelists receive instructions or agree among themselves to permit "current constraints" to influence their judgments. Past mistakes can thus become the rationalized standard for future practice instead of methodologies to protect against threats to internal and external validity. It might be possible to correct for this consensual biasing effect by requiring panelists to explain and defend their ratings of methodological adequacy in an adversary process. Ratings thus anchored could facilitate comparison of the meanings ascribed to ratings by

Table 2. The global assessment score [adapted from (*21*), p. 11].

Sources of variation	Degrees of freedom	Sums of squares	Mean squares	F	Variance explained (*23*) (%)
Meeting grantee (α)	1	16.67	16.67	22.52*	8.64
Prior rating (β)	1	0.17	0.17	0.23	0.0
Project (γ)	5	84.96	16.99	22.96*	44.06
Judge strictness (δ)	3	1.04	0.35	0.47	0.0
Interaction effects					
$\alpha\beta$	1	0.02	0.02	0.05	0.0
$\alpha\gamma$	5	11.29	2.26	3.05†	4.12
$\alpha\delta$	3	3.08	1.03	1.39	0.47
$\beta\gamma$	5	5.56	1.11	1.50	1.01
$\beta\delta$	3	1.57	0.52	0.70	0.0
$\gamma\delta$	15	16.98	1.13	1.53	3.17
Total interaction	32	38.50			
Residual error	53	39.62	0.74		38.53
Total	95	180.96			100.00

* $P < .001$. † $P < .05$.

Table 3. The composite score [adapted from (*21*), p. 16].

Sources of variation	Degrees of freedom	Sums of squares	Mean squares	F	Variance explained (*23*) (%)
Meeting grantee (α)	1	7.61	7.61	32.23*	11.06
Prior rating (β)	1	0.29	0.29	1.21	0.07
Project (γ)	5	30.81	6.16	26.09*	44.44
Judge strictness (δ)	3	0.65	0.22	0.92	0.0
Interaction effects					
$\alpha\beta$	1	0.09	0.09	0.37	0.0
$\alpha\gamma$	5	3.28	0.66	2.78†	3.15
$\alpha\delta$	3	2.06	0.69	2.90†	2.02
$\beta\gamma$	5	3.86	0.77	3.26†	4.01
$\beta\delta$	3	1.54	0.51	2.17	1.24
$\gamma\delta$	15	3.48	0.23	0.98	0.0
Total interaction	32	14.30			
Residual error	53	12.52	0.24		34.01
Total	95	66.17			100.00

* $P < .001$. † $P < .05$.

different judges either on the same or different panels. In time, development along these lines might lead to standardized rating scales suitable for adoption by the peer review panels of all agencies funding applied social research.

Last, the findings, cited above, of Bernstein *et al.* (*19*) and Minnesota Systems Research (*20*) reinforce the conclusion that peer review and other "batch" assessment practices need buttressing as a means of quality control. It is disturbing to learn that only 10 percent of projects meet minimum scientific standards or completely achieve their objectives. But realization that 51 percent of projects—if representative of what happened to the approximately $45 to $50 million spent for evaluation research in fiscal year 1970 by federal agencies—fell below 4.24 on a seven-point scale of methodological adequacy (where "6" stood for the minimum standard) provokes outrage and demands strong corrective action (*19*, pp. 4 and 72).

A New Quality Control System

How can we improve the present system for quality control if peer review by itself cannot do the job? It seems clear that any new system will have to gain the acceptance of the science community if it is to succeed. It must be objective, fair, and governed by due process. Otherwise, the system will be viewed as a tool of special interests wishing to suppress the free spirit of scientific inquiry and, if sponsored by government, boycotted as an instrument of censorship and state control. Responsibility for designing and developing a new system, therefore, should be lodged in an organization which has the sanction and support of both the science community and government. Needless to say, this organization and any organizations assuming responsibility for administering the system will have to guard against possible infiltration by persons in the science community, industry, or government who may on occasion have an interest in "rigging the jury" (*26*).

Implementation of the new system will require one or more organizations with broad scanning opportunities and knowledge of persons of scientific reputation so that the appropriate expertise can be drawn to the task of reviewing a wide spectrum of research. To assure that experts are consistent in identifying both the strengths and weaknesses of projects under review, it may be desirable to assemble panels consisting of both substantive and methodological experts and have them operate as adversaries. The principle governing resolution of differences of opinion should be robust: the mere possibility of some alternative explanation of fundings will not suffice—only plausible rival hypotheses will be considered invalidating.

The system will have to provide some means for uncontaminated communication between panels of experts evaluating specific projects and the principal investigators of those projects. Interim and final reports seldom contain sufficient information concerning methodologies and procedures to allow "cookbook" replication. Resolution of ambiguities may require access to original data for additional analysis. When conflicting results are produced by different studies, the panel may have need of the original data from the several studies in order to pool them according to criteria that allow direct comparison (*27*). Permanent staff of the organization managing the system could assume responsibility for anonymous communication between the panel of experts and principal investigators. It may also be desirable to conceal as much as possible from panel members the identities of principal investigators. This will require some editing by permanent staff of the text and references of reports before they are transmitted to the panel for review (*28*).

What form of feedback should the panel provide to the several parties interested in the results of evaluations? A single formal report may be adequate. In addition to findings concerning the validity of studies, it will also be helpful, when there are sharp differences of opinion, to receive any minority reports and sufficient information to place them in context. Panels should be able to make recommendations on preferred methodologies or specific corrective measures needed to overcome deficiencies or limitations uncovered in the course of the evaluation. If panelists cannot reconcile conflicting evidence concerning an issue of substantial importance, they might even undertake design of a tiebreaking study.

Fairness dictates that principal investigators be given an opportunity to make a formal reply to an adverse evaluation. This reply should probably become part of the report through which the panel communicates its findings to all interested parties. Once the panel has filed its report, the burden of decision on how to proceed will then rest upon potential users. At a minimum, one would hope that the agency which sponsored the research would use the panel report to qualify press releases and dissemination plans. Administrators and policy analysts will ultimately have to decide for themselves how much to rely on the findings of a specific study as a guide to policy. For them, the panel report might serve the good purpose of increasing their awareness of the real chances for error—thus enabling a decision to be either hedged or made with confidence (*29*). The new system might even play a "broker" role in facilitating communication between researchers and policy-makers.

How far agencies sponsoring research should go in using panel reports to monitor the performances of grantees and contractors and of their own R&D administrators and technical consultants is an open question. In any event, it is safe to assume that creation of a quality control system for applied social research such as the one proposed would have the immediate impact of tightening standards and realigning expectations among all affected parties.

Concluding Remarks

Aside from the need to tighten standards and realign performance expectations, there are other good reasons for establishing a new quality control system for applied social research. Most important would be the effect that a new system might have by unmasking and curtailing the use of scarce R&D funds for service subsidy and the seeking of influence. In addition to opportunity costs, there are considerable other nonmonetary costs associated with the practice of masking service subsidy and the seeking of influence under the guise of knowledge-building. First, the rhetoric promising "open competition among competing ideas" begets cynicism and disillusionment in those who compete and lose to those who had an "inside track." Second, projects serving mixed purposes create ambiguity and, because of the servicing requirements they impose, distract and dissipate the energies of technically qualified R&D administrators—thus preventing them from doing

the job for which they were hired in the first place.

A new quality control system will undoubtedly influence the direction, scope, and content of research dissemination and utilization activities of government agencies. Evaluation of finished research can guide not only utilization of what has been produced but also the funding of new research. Program administrators and policy analysts, unschooled in the language of research technology, should be able to obtain somewhat of a less cluttered view of the productivity of the investments they authorize for R&D and program evaluation. If need be, the system can be used to improve the existing contractor rating system by making available objective evidence about bidders' past performances. Finally, a successful new system will make it more difficult for agencies to commission studies favorable to partisan points of view.

The stakes seem large enough to justify the expense of creating a new quality control system for applied social research, even at a cost each year of 1 percent or more of the total R&D and evaluation budgets available to government agencies. Conceivably, the National Academy of Sciences or the National Science Foundation, through its program of Research Applied to National Needs (RANN), might underwrite development of a new system as a basic contribution to the advancement of applied science.

References and Notes

1. A. Etzioni, Washington *Post*, 11 June 1972, p. B-3.
2. N. Wade, *Science* **179**, 158 (1973).
3. H. J. Eysenck, *J. Consult. Psychol.* **16**, 319 (1952); *Int. J. Psychother.* **1**, 97 (1965); E. E. Levitt, *J. Counsel. Psychol.* **21**, 189 (1957).
4. S. P. Segal, *J. Health Soc. Behav.* **13**, 3 (1972).
5. D. T. Campbell and A. Erlebacher, *Disadvantaged Child* **3**, 185 (1970).
6. Comptroller General of the United States, "Evaluation of the Office of Economic Opportunity's performance contracting experiment," Report to the Congress B-130515 (General Accounting Office, Washington, D.C., 8 May 1973).
7. D. J. Armor, *Public Interest* **28**, 90 (1972); T. F. Pettigrew, *ibid.* **30**, 88 (1973); J. Q. Wilson, *ibid.*, p. 132.
8. H. A. Averch, S. J. Carroll, T. S. Donaldson, H. J. Kiesling, J. Pincus, *How Effective is Schooling?* (Rand Corporation, Santa Monica, Calif., 1972).
9. H. Orlans, *Ann. Am. Acad. Polit. Soc. Sci.* **394**, 28 (1971).
10. Nine threats to internal validity exist which act as rival explanations for changes or differences that are thought to be the effects of intervention or treatment. They are: (i) intervening historical events; (ii) maturation effects of processes within respondents or observed social units; (iii) instability of measures or sampling units or both; (iv) learning effects due to repeated measurements; (v) changes in measuring instruments or observers; (vi) regression artifacts due to extreme cases being selected for treatment; (vii) selection biases due to unequal recruitment of comparison groups; (viii) differential loss of respondents from comparison groups; and (ix) interactive biases due to different rates of maturation or autonomous change among respondents. The six threats to external validity, limiting generalization of results to other settings, other versions of intervention or treatment, or to other measures of the effect, are: (i) interactive effects of testing and treatment; (ii) interactive effects of selection and treatment; (iii) reactive effects of atypical or artificial treatment conditions; (iv) confounding due to multiple treatments; (v) responsiveness to irrelevant components of measures; and (vi) irrelevant replicability of complex treatments. D. T. Campbell, *Am. Psychol.* **24**, 409 (1969).
11. D. T. Campbell, *Psychol. Bull.* **54**, 297 (1957).
12. L. Kish, *Survey Sampling* (Wiley, New York, 1967).
13. L. J. Cronbach, *Psychometrika* **16**, 297 (1951).
14. T. J. Duggan and C. W. Dean, *Am. Sociol.* **3**, 45 (19(D. Gold, *Am. Sociol. Rev.* **22**, 332 (1957); D. E. Morrison and R. E. Henkel, *Am. Sociol.* **4**, 131 (1969).
15. A. L. Edwards, *The Social Desirability Variable in Personality Assessment and Research* (Dryden, New York, 1957).
16. See, for example, O. L. Deniston and I. M. Rosenstock, *Health Serv. Rep.* **8**, 153 (1973); references (*10*) to (*15*) above either give explicit guidance on preferred methods or imply normative procedure.
17. K. S. Crittenden and R. J. Hill, *Am. Sociol. Rev.* **36**, 1073 (1971).
18. D. Crane, *Am. Sociol.* **2**, 195 (1967).
19. I. H. Bernstein, P. P. Rieker, H. E. Freeman, "A review of evaluative research: The state of the art, methodological practices, and dissemination of research findings," unrevised draft of paper presented at the 68th Annual Meeting of the American Sociological Association, Washington, D.C., 28 August 1973.
20. Minnesota Systems Research, Inc., "Methodological adequacy of federal R & D projects," unpublished final report, Minneapolis, Minn., December 1973, executive summary, p. 1 and appendix table 10.
21. ———, "Assessment of methods to evaluate the scientific rigor of social research: A study of six rehabilitation research projects," unpublished interim report, Minneapolis, Minn., May 1973.
22. One panel member prerated and preranked the six projects at home but could not attend his panel meeting.
23. Estimated components of variance, calculated by reference to the following type of formulas appropriate to the fixed-effects ANOVA (analysis of variance), underlie the proportions of variance explained.

$$\hat{\sigma}^2_\alpha = \frac{(J-1)\,(MS_\alpha - MS_\epsilon)}{n\,JKLM}$$

where MS_α is the mean square associated with effect α, MS_ϵ is the mean square error, n is the number of observations per cell, J represents the levels of effect α, K represents the levels of effect β, L represents the levels of effect γ, M represents the levels of effect δ. The computed estimate of a variance component is negative when the mean square error is larger than the mean square associated with an effect. In such a case, it is reasonable to assume that the component of variance actually equals zero. See G. M. Vaughan and M. C. Corballis, *Psychol. Bull.* **72**, 204 (1969). The proportion of variance explained by each effect is the expression of a given component of variance as a proportion of the sum of all components.
24. The 15 judges on the four panels made a total of 225 paired comparisons of the six projects with respect to their overall methodological adequacy (15 judges times 15 paired comparisons of the six projects). Kendall's U, an index of interjudge agreement for paired comparisons which varies from zero (complete disagreement) to one (perfect agreement), was 0.45. Although statistically significant ($P < .001$), clearly there was as much disagreement as agreement with respect to the possible realm of agreement. See (*21*, pp. 27–28 and E-23).
25. In actual practice the test of "relevance" may often compromise selection of projects which meet the most rigorous standards. The cynic might go so far as to hypothesize an inverse relationship between methodological adequacy and "relevance," as defined by a sometimes arcane bureaucratic process that is extremely sensitive to political considerations of every kind and the impetus of frenzied "use or lose" spending at the end of each fiscal year.
26. See, for example, R. Gillette, *Science* **174**, 800 (1971); V. Cohn, Washington *Post*, 25 April 1973, p. A-2.
27. R. J. Light and R. V. Smith, *Harv. Educ. Rev.* **41**, 429 (1971).
28. Perfect two-way anonymity between panel members and principal investigators becomes increasingly difficult to achieve as principal investigators gain stature in their respective fields or release interim reports of findings to funding agencies and the press. The drive for notoriety and fame is perhaps nowhere stronger than in the science community [see, for example, J. D. Watson, *The Double Helix* (Atheneum, New York, 1968)]. Nevertheless, it is still important to establish the principle of anonymity and a ritual to support it. There should at least be a code of ethics to govern relationships between panel members and principal investigators whose studies are being evaluated.
29. The technical problem here is that of finding an appropriate decision-rule—one which, under the circumstances, limits chances of reaching hazardous erroneous conclusions. For example, setting the alpha value of statistical significance at too stringent a level increases the possibility that true differences will be declared false (type II error), whereas too lax a level increases chances of asserting true differences where none exist (type I error). When the social consequences of erroneously asserting true differences are not abhorrent, it seems reasonable to increase chances of Type I error by establishing a larger than conventional alpha value of statistical significance. See W. L. Hays, *Statistics for Psychologists* (Holt, Rinehart and Winston, New York, 1963), pp. 265–269.

30. I wish to express my special thanks to Ms. Sylvia McCollor, Minnesota Systems Research, Inc., for her help in the preparation of this article.