

Peer Review for the 21st Century:
Applications to Education Research

Prepared for a National Research Council Workshop
Washington, D.C.
February 25, 2003

Submitted August 1, 2003

Edward J. Hackett
Department of Sociology
Arizona State University
Box 872101
Tempe, AZ 85287-2101
ehackett@asu.edu

Daryl E. Chubin
National Action Council for
Minorities in Engineering, Inc.
350 Fifth Avenue-Suite 2212
New York, NY 10118-2299
dchubin@nacme.org

Preface

To declare our biases at the outset, we offer a brief sketch of aspects of our respective backgrounds and experiences that influence our ideas about peer review. We both have been involved with the grants peer review systems of various agencies for many years. Hackett has been PI of several NSF grants for research (and one for graduate training), took a turn as a rotating program officer at NSF, served on more panels than he can remember (including a couple of NIH Initial Review Groups [or “study sections”]), and wrote his first (ad hoc) mail review for NSF sometime in the late 1970s. At last count he has handled about 1,000 proposals in one or another of these capacities. Chubin has been a PI on several grants, an NSF division director in the Education and Human Resources Directorate (approving recommended awards), and a frequent panelist and reviewer. Before that, as staff of the Office of Technology Assessment, he directed analyses of federal agency research funding mechanisms, including grants peer review, to inform committees of the U.S. Congress. Later, as senior policy officer for the National Science Board, NSF’s Presidential-appointed governing body, he participated in revisions of the agency’s merit review criteria in 1999.

We have also collaborated in studies of the peer review system, using interviews, questionnaires, observation, and examination of the “gray literature” that abounds in Washington (and is hard to find elsewhere). We’ve even contributed to that fugitive press. In the open literature, our work on peer review has appeared in the book *Peerless Science: Peer Review and U.S. Science Policy* (SUNY Press, 1990) and some articles. We’re among the “usual suspects” rounded up to speak about peer review to audiences in the U.S. and elsewhere.

None of this should be understood as making us experts on the topic. It simply explains our distinctive perspective on peer review: we see it as participants and observers, insiders and outsiders, as donors and beneficiaries, as gatekeepers and gatecrashers.

This paper offers an institutional analysis of peer review, a style of sociological analysis developed to an art form by Robert K. Merton, the founder of the sociology of science, who died in early 2003. An institutional analysis examines a social phenomenon, asking not only what is done “on the surface” but also what deeper purposes are accomplished, what concealed values are served. In Merton’s language, one seeks the “manifest and latent functions” of the institution. Our mode of institutional analysis may lack Merton’s magisterial prose and historical sweep, but in exchange it is enriched specifically by a vigorous half-century of research on the social dimensions of science and more generally influenced by social theories and concepts that have arisen over the years.

Our aims here are to use an examination of peer review to inform the development of a policy-minded review system for education research. Throughout we emphasize the importance of legitimacy and the role of peer evaluation in building and sustaining a knowledge-producing community.

Orientation

Some years ago, when the recognition that federal support for science could have significant returns to the U.S. economy, a member of Congress asked his colleagues, “When did we decide that peers would slice the melon?” It’s actually hard to say exactly when or why that decision was made, but it is certain that the Congress’s skeptical stance toward peer review – in contrast to population-based formula funding and direct appropriations (see below) – has endured.

Peer review of scientific *manuscripts* dates back to the Philosophical Transactions of the Royal Society in the 17th century. The origins of *grants* peer review are more recent and much murkier. The National Advisory Cancer Council, established in 1937, was authorized to review applications for funding and “certify approval” to the Surgeon General. The Office of Naval Research developed an informal variant of peer review, which may have been brought to NSF by Alan T. Waterman. Peer review is not mentioned in NSF’s founding legislation, but according to J. Merton England, historian of NSF’s early years, “no doubt everyone understood that there would be some sort of peer review” (see Chubin and Hackett, 1990: 19-20 for elaboration as well as original quotes and references).

If peers don’t “slice the melon” for science, then who might? There are alternatives to “expert” review. Congress might, and does, exercise its prerogatives: the practice is called direct appropriation (“earmarking,” “porkbarrelling,” or worse...). In fiscal year 2002 Congress earmarked \$1.8B for projects at colleges and universities, continuing a steep upward trend that began in 1996 (and reversed a two-year interruption in a decade-long rise; see *The Chronicle of Higher Education*, 27 Sept 02, A20). Not all

of this money is for science and not all is for research, but earmarks for academe are a useful indicator of the phenomenon. While \$1.8B is a rather small amount compared with the roughly \$100B federal investment in R&D, it seems somewhat larger in comparison to the federal budgets for basic research (about \$25B) and basic research performed on campus (about \$12B) (all data from AAAS analysis of R&D budget; <http://www.aaas.org/spp/rd/guihist.htm>).

Critics of earmarking complain that its allocations circumvent technical expertise, that is, they do not draw upon the collective wisdom of the community. The symbolism is worrisome: earmarked funds have a corrosive effect on the meritocratic values of science, especially when compared to the difficulty of writing a successful grant proposal. How discouraging it would be to learn that others have received funding without competition! Earmarks may have similar effects on reviewers (cynicism) and even on recipients (stigma) of earmarked awards. In general, the culture of academic science is quite sensitive to procedural compromise.

Supporters of earmarking counter that the practice belongs in the U.S. system for funding R&D because it reflects the democratic values of distributional fairness (“geographical equity”) and representative decisionmaking. Stated more forcefully, earmarking *complements* peer decisions, compensating for oversights and even repairing the "market failure" of meritocratic decisions that may have an elitist edge to them. Debates about earmarking and other political instruments for allocating research funds are powerful reminders that scientific expertise and institutional capability are quite concentrated in relatively few places within the U.S. (U.S. Congress, 1991: 89-93). For

that reason alone, perhaps a small amount of earmarking has a place in a pluralistic funding system.

If peers don't always slice the melon (advice needing consent) and Congress doesn't typically slice the melon (consent without advice), one might rely on a single, strong manager who makes decisions according to his or her best judgment (as is done in the Defense Advanced Research Projects Agency [DARPA]; for details see www.darpa.mil/body/information/proposal.html). In effect, this is peer review with one peer, so this steward had better be on a par (intellectually and in "stature" within the field) with those applying for support. As pointed out by Susan Chipman, a highly accomplished program manager for the Office of Naval Research, the job is much like running a distributed research center, and at turns the manager plays the roles of advocate, broker, collaborator, evaluator and, we would add, terminator. The person should also understand the field and its needs (which should be clear and widely shared) to insure that decisions and allocations are wise, legitimate, and effective.

In the DARPA model, the program being managed has well-defined objectives and the manager is held accountable for performance outcomes (because *process* accountability is low). DARPA packages many of these enabling conditions under the distinction it draws between projects and programs. *Projects*, which focus on a common objective or idea, have a beginning and an end, and a specific, hoped-for result that may have very high risk. *Programs*, by contrast, emphasize particular academic disciplines or general technologies and tend to be very open-ended. While the selection of program proposals often places heavy emphasis on previous publication histories and peer review, DARPA tries to distinguish itself as an agency that is based almost entirely on good ideas

with clear, exceptionally beneficial consequences. It sponsors projects, not programs. (www.darpa.mil/body/information/proposal.html, p. 2).

This arrangement has worked spectacularly well in some fields and for some purposes, though it would probably not work as well for all fields of science and would probably face some difficulty scaling up to the size of NIH (about \$27B in 2003). It also requires program managers and their bosses to be willing to accept failure and to exercise a bit of ruthlessness in cutting losses when a good, high-risk idea fails to pan out in practice (e.g., in year 3 of a 5-year award).

Nonetheless, the strong manager approach may have a place in agencies that otherwise depend on peer review for their allocation decisions. Elements of the strong manager's agility and risk-taking are a valuable adjunct to peer review at NSF, for example, where Small Grants for Exploratory Research (SGER, commonly called "sugar" grants) may be awarded in amounts up to \$100,000 on the program officer's recommendation, provided that the total of such grants does not exceed 5% of a program's budget. The SGER mechanism allows program managers to support risky research, exploratory projects, and research opportunities that pop up in a brief and closing window of time (e.g., studying response to a natural disaster).

A third research funding possibility would use a formula to allocate resources. Allocations may be made to states or universities or institutes, then suballocated to groups or individuals according to a variety of additional criteria. Some have proposed formulas for evaluating the past performance of individual scientists, then awarding funds accordingly. Sheer numbers of researchers at a university, or residents of a state, or some measure of current need or potential payoff may also factor into the equation. Fair and

effective formulas would be hard to devise, and the relative merits of alternatives may be debated endlessly. Among the serious questions one might ask are: How would newcomers fare in such a system? How would old-timers be put out to pasture? Would the system encourage researchers to take risks and to persist in a line of inquiry that proves recalcitrant? Who would develop and administer the formula, preserving it from scientists' efforts to "game" the system (by doing precisely those things that are amply rewarded, even if they are not most beneficial or intellectually daring)? The answer, of course, is that no one would have such authority and responsibilities.

Finally, we come to the allocation mechanism of peer review, or "merit review with peer evaluation," as NSF and other agencies prefer to call their review systems. (Using the term "merit review" relieves the agency of the obligation to identify, locate, and deploy a proposer's "peers," and sustains a subtle distinction between *peer review*, which focuses on scientific quality, and *merit review*, which takes account of a broader array of criteria.)

Peer review may be considered a "boundary process," in the sense that it spans the boundaries of several social worlds (compare Star and Griesemer, 1987, on boundary objects; Galison, 1997, who discusses trading zones in high-energy physics; and Guston, 2000, who extends the idea to boundary organizations). Thinking about peer review as a boundary process emphasizes its position at the intersections of science and policy, of academe and government. Peer review sometimes also straddles disciplines, when it is applied to interdisciplinary fields or research initiatives, and may also cross the boundaries of knowledge production and professional practice, of research and policy. For example, it would be natural for a federal agency to feel that it "owns" the peer

review practice and is free to decide when, how, and to what effect it is performed. But it is equally natural for academic scientists to see the process as “owned” by their discipline—it is, after all, conducted by them in their language and aims to advance knowledge in their discipline. It only happens that the meeting and the money are in Washington.

Calling peer review a boundary process is more than a trendy trope: it directs attention to the mix of communities, purposes, evidential standards, argumentative procedures, ethical precepts, theoretical frameworks, epistemic cultures, principles of fairness and the like that mingle and collide in the review process. Those engaged in boundary processes may experience difficulty in achieving mutual understanding, and a variety of linguistic and operational accommodations may be necessary (hence pidgin and Creole languages arise and hybrid practices are invented; Galison, 1997: Chapter 1).

To understand peer review in comparison with alternative mechanisms of resource allocation, and particularly to design or re-engineer peer review to strengthen a community of scientists, we must first take a step back and think more generally about the purposes and values of peer review. In the next section we explore the purposes of peer review, following with a discussion of the competing values it is asked to serve. These purposes and values establish a framework for thinking about peer review and for choosing approaches and evaluating their effects. Each science funding organization will have its own purposes and traditions, so there will be differences within this family of approaches – even *within* the same federal agency or department.

Keep this notion foremost in mind, too, when reading further: Peer review allocates scarce resources – money, time, space – and the career capital they generate.

The result of peer review decisions will concentrate or disperse the available resource over the pool of eligibles. It will anoint “winners” or “hedge bets.” At one extreme, it will be highly selective by restricting the competitors to those with certain characteristics (what is known as “set-asides” by age, gender, discipline, prior accomplishment, or location at an institution with a track record or facility to conduct the research). At the other extreme, peer review will appear to be a lottery, yet with the criteria of choice related to the chosen projects. Along this continuum, the process must be sufficiently fair and translucent so that competitors find the opportunity real and the outcome justified (even if they are judged to fall short when the awards are announced).

What Is Peer Review?

Peer review is surely a technique for "grading the grain," to borrow the term of Kees Le Pair, an administrator at the Dutch Technology Foundation. Grants peer review influences the distribution of research funds, while manuscript peer review guards the entry into the scientific literature. But peer review does much more, so efforts to implement or reshape peer review systems should take account of these additional purposes.

Peer review is a source of **expert advice** to the proposer, in hopes of improving the product, and to the decisionmaker, in hopes of yielding wiser allocations. Indeed, having resources at stake motivates and sharpens the critique in much the same way that gambling games are a lot more interesting when there's money on the table. In the aggregate, the collection of review-based advice guides and shapes a research area through both mechanisms: members of the field get the idea that work on a particular

topic is no longer interesting or valuable, and the program gets much the same message (and may eventually choose to codify it in a revised program announcement).

Peer review is a **flywheel** that lends stability to research in an area. In this respect the review process embodies Kuhn's (1977) "essential tension" between originality and tradition in science: promising new ideas are tested against the cumulative store of shared knowledge and established theory. Peer review challenges whether new ideas are truly new and worth pursuing, in principle distinguishing between sound innovation and reckless speculation. This is not to understate the importance of originality in science: The division director for social and behavioral sciences at NSF asked his program officers to name the reason most commonly mentioned in their "Form 7s" or program recommendation for declining a proposal. Answers varied widely from weak methods to missing literature to inconsistent theory. In fact, the most commonly cited reason was lack of originality. That is, when program officers summed up the arguments and criticisms presented in written mail reviews and panel discussions to form a recommendation for declining a proposal, "there is nothing new here" turned out to be the most common reason they gave. Reviewers drew upon their collective knowledge of the field to critically evaluate a proposal's claim to stake out a new line of inquiry, and by their rejections of such claims they reassert the established knowledge of the field.

A second, related way that reviewers defend tradition against claims of originality is when they reject novel ideas as impractical, unlikely, unworkable, or as implausibly inconsistent with the established body of knowledge. Huge imaginative leaps may be excluded by this process, but the tight weave in the fabric of established findings in a field is reinforced as a result. Huge leaps would create "pulled threads" in the fabric,

causing damage and distortion. They would also entice other scientists to fill the void with further research, in somewhat the same way that paradigmatic puzzle-solving or normal science fills in an intellectual space. Reviewers may be cautious about endorsing studies that take leaps not so much because they do not appreciate creativity but perhaps because they are troubled by the potential expenditure of research energy needed to fill the space. In this view, sharp disagreements among reviewers about the merits of an idea may indicate a promising but risky new research path (e.g., “premature” ideas).

Consensus, in contrast, might indicate a sufficiency of problems left to solve in the research area, or the grip of a school of thought, or some plain risk-aversion (see below).

Peer review lends inertia of a third sort, too, giving the successful proposer some momentum and an endorsement that might keep a project on course through the vagaries and disappointments of research. This is easiest to see by imagining alternatives. If funding is provided without a semi-public statement of a research plan and its endorsement through peer review, inconveniences and obstacles met along the way might derail the project, driving its investigator to shift aims and methods. Fortified by the critical scrutiny and imprimatur of peer review and the public nature of the promise implied in accepting a grant, a project may be more likely to remain on course through various setbacks.

Peer review is a **communication channel** that circulates research ideas in their formative stages to key “influentials” in a field. In some disciplines (e.g., ecology) this signals others to stay away from an area, organism, or research problem, thus avoiding duplication of effort. In others, it calls attention to a problem that is promising, perhaps urgent, attracting other researchers and setting off a race for priority (think about work on

cancer genes). Communication through peer review also helps prepare the ground for new ideas by first circulating them in the speculative format of a proposal, which will be followed (as results are achieved) by colloquia, presentations at meetings, and manuscripts submitted for publication. By the time a body of research is finally published, aspects of its findings and methods may be generally familiar to many in the field, speeding acceptance and utilization of the new work. Of course, ideas circulated in this fashion also attract criticism and advice, contributing to the research process and the quality of the final product.

Peer review is an **entry point for adding value beyond quality** to research decisions. This occurs, for example, when NSF program officers try to balance their portfolios by taking account of geographic distribution, age, gender or ethnicity of investigator, research participation of four-year colleges or historically-black colleges and universities, the “hotness” of a topic or method or other such considerations. At NIH, “Advisory Councils” (the second of a two-level review process) are empowered to recommend some proposals for funding out of priority-score order, perhaps because they address urgent national needs. While possible in principle, this is seldom done in practice.

With the Government Performance and Results Act of 1993 requiring research agencies to show that their investments yield societal benefits, we must ask if science and technology experts are the reviewers best qualified to render such judgments. The question leads to a search for ways to enrich or augment peer review. One relatively recent innovation allows more direct and systematic citizen participation in scientific and technical allocation decisions. The Dutch Technology Foundation for several years has

augmented traditional peer review with “lay review” by citizens. Advisory councils of the individual NIH institutes apply similar criteria, albeit sparingly, as part of a two-stage review process. The "Director's Council of Public Representatives," formed by Harold Varmus in 1999, tries to do this by involving citizens in “the broad development of NIH programmatic and research priorities” (for details see www.copr.nih.gov). Activist and support groups for various diseases call for further efforts in such directions, applying pressure for more democratic participation in science and technology. In short, who participates redefines who is a “peer” (Sarewitz, 1996) and the purpose of the “review.” Public dollars define the public’s interest in the process.

Peer review is also an **assertion of professional authority**, with both practical and symbolic attributes. Peer review creates a buffer or a boundary, allowing scientists to make decisions in a privileged space, apart from the general public and politics, and to do so according to principles that reinforce their professional culture. It is a realm in which expert decision-making is expected and respected (congressional qualms about melon-slicing notwithstanding). This may seem inconsistent with the preceding principle—public participation – but should be understood as reflecting peer review’s role as a boundary process: it crosses borders, taking on distinctive characteristics in each region. A good review system delimits the amount and character of public participation, shaping the form of input allowed and preserving professional autonomy while permitting lay participation. Federal agencies seek a delicate balance: deference to expert evaluation untainted by politics, yet sensitive to societal needs and “extrascientific” values. These values often encompass questions of research application, risk and benefits to whom, and long- versus short-term consequences (what NSF terms “broader impacts”).

In all, peer review is more than a tool for allocating resources or a measuring stick for “grading the grain” of science. Recognizing these diverse and somewhat incongruent purposes underscores the thoroughgoing significance of a review system for a community of researchers and practitioners. Precisely because peer review spans the boundaries of several social worlds—science and policy, research and practice, academe and bureaucracy, public and private—its purposes and meaning may be understood differently across communities and at different times in the history of a single community.

The Competing Values of Peer Review

Not only do we ask peer review to do many different things, in doing so we ask it also to serve a set values that are not entirely consistent with one another. At a particular juncture, a field may be better served by emphasizing certain values at the expense of others. It is unlikely, however, that any workable peer review system would long be at the extreme of any particular value dimension. We recognize that these value dimensions are not independent of one another and that their poles are not complete opposites. But the dimensions are sufficiently disparate to discuss individually and their poles are sufficiently inconsistent to impede efforts to satisfy both values in each pair simultaneously.

Openness-Secrecy

Peer review is open to the community of qualified scientists as proposers or reviewers. The process of peer review, at least at the level of procedures, criteria, rating

scales and such, is knowable, transparent (well, translucent), and held to account for its workings and outcomes. And the review criteria are open and applied equally to all.

(Thanks to Bob Cook-Deegan, 1996, for developing this dimension).

But peer review is also secret. Confidentiality is pivotal, and anonymity is preserved throughout much of the process. The meetings are closed and their minutes are minimal. Proposals, reviews and panel discussions are privileged. To outsiders, who participates and how they are chosen can seem mysterious, and reviewers' identities are generally not disclosed. To meet the requirement of publishing panel membership lists, for example, the NSF Sociology Program lists on its website the name of every person who served on the panel within the past several years. Concealing identities by revealing them in this way protects the identities of those serving, reducing the chances that they are lobbied or hassled or worse. As a counterinstance of openness in reviewer selection, the Fund for the Improvement of Post-Secondary Education (FIPSE) website invites potential reviewers to nominate themselves, but it is unclear how this works in practice.

Effectiveness-Efficiency

Peer review is asked to be effective – to recommend projects that would benefit the field and confer some greater social benefit, to offer advice to proposers, to circulate ideas within a community, and the like. Peer review is also asked to be efficient, to do all this at very low cost, with cost measured in terms of dollars spent on reviews (infrastructure, travel, reviewer compensation) and hours expended by proposal writers and reviewers. How realistic are these expectations?

For reasons based in idealizations of the public service ethic of academic life (an issue too large to go into here), reviewers are paid little for their services, so it is easy to be efficient in terms of monetary costs. Among peer review systems education seems to pay its reviewers quite little for their time, which may contribute to the difficulties of getting expert reviewers and thorough, competent reviews (August and Muraskin, 1998).

Setting aside costs measured in dollars, the review process is also asked to impose light burdens of time and involvement on reviewers and proposers. But a thorough review of fifteen single-spaced pages might take half a day, and is unlikely to be accomplished in less than two hours. A sound proposal takes many days of work to prepare, and a low success rate—around 10% at OERI in recent years—offers a low expected return for the investment of effort. Similarly, the value of openness demands relatively few limits on who may submit a proposal, but it would be more efficient all around to spare some would-be proposers the effort. (Hence the invention of a two-stage process with the first a “preliminary proposal” that can be screened into or out of the more competitive second stage.)

In all, efficiency in the proposal and review process may be achieved only at the expense of other desiderata. To make the best choices and to provide the best advice to proposers would require more extensive, intensive, expensive review. Each agency must decide the value of the advice its staff can afford to solicit and deliver.

Sensitivity-Selectivity

In the selection of research projects, the peer review system is asked to be highly sensitive and highly selective at the same time. A sensitive review system would detect

the merit in every worthwhile proposal, while a selective system would filter out all projects of dubious quality or significance. In effect, a sensitive system captures the signal, however faint, while a selective one removes the noise, however innocuous. Given the risky nature of scientific research, the difficulties in communicating original ideas clearly and persuasively, and the possibility that the phenomenon of interest in some fields may itself be in question (the Higgs process, the top quark, prions), the two poles are in tension. A system sensitive to every scintilla of scientific merit would probably support some projects that don't work out. But a system so selective that only projects beyond skepticism are chosen for funding would surely filter out some good ideas with the bad. (Besides, some researchers simply write better than others. Still others construct better proposals than conduct the research once funded. Just what is the review rewarding?)

Different fields at different times may be better served by adjusting the review process toward one pole or the other. Even within a field certain problems or objectives might be better served by a review system adjusted toward sensitivity or selectivity. Selectivity might work best on problems situated within well-developed theoretical frameworks with clearly established methods, whereas sensitivity may be preferred when the problem or objective is ill-defined or when there is little agreement about the most promising approach.

Innovative-Traditional

This is Kuhn's (1977) essential tension (or Merton's norm of organized skepticism) in operational form: an innovative review system would reward novelty,

risk-taking, originality, and bold excursions in a field, while a traditional system would sustain the research trajectory established by the body of accepted knowledge by imposing skeptical restraint on new ideas. Peer review is expected to identify, encourage, and support frontier work, but to deliver us from fads and passing fancies. So it is simultaneously criticized for confusing brilliance and hubris at one extreme, incrementalism and narrow-mindedness at the other.

Meritocratic-Fair

Peer review is expected to be meritocratic, judging proposals and scientists equally in accordance with the stated criteria (and nothing else). NIH says this well in its instructions to reviewers, asking them to evaluate all the science and nothing but the science in the proposal. Recognize, however, that this asks reviewers to do an extraordinary amount of boundary work (Gieryn, 1999) on the fly: they are expected to extract science from speculation, rhetoric, common sense, practical benefit, and whatever else the proposer orchestrated into the document. Reviewers are to find and evaluate all the science in the proposal, crack or smelt or separate (or your favorite separation metaphor) it from the irrelevance in its context, and render judgment.

At the same time, however, peer review functions within a political context that expects it to embody and apply standards of fairness to ideas apart from consideration of a scientist's reputation, personal characteristics, geographic or academic position; the economic potential of the proposed work; or its relevance to pressing national needs. The best science for a discipline may not be the best science for the nation. Decisions that adhere strictly to meritocratic criteria may violate societal standards of fairness.

For example, principles of distributional fairness or capacity maintenance might indicate that decisions at the margin should give preference to investigators who currently have inadequate funds. Similar arguments may be made for primarily undergraduate institutions, traditionally minority-serving colleges and universities, and for states that tend to receive little research support. One could support such decisions in terms of growing research capacity, or educational or economic investments, or politically savvy allocations. Whatever their name or purpose, such decisions deviate from strict application of meritocratic principles. Nevertheless, most people are fine with that much of the time.

Reliability-Validity

As an assessment tool peer review is asked to be both reliable and valid. A reliable measuring instrument has little random error; a valid measuring instrument measures what it is supposed to measure (and nothing but what it is supposed to measure). To be reliable, ratings should show high levels of agreement between raters and consistency from one group of raters to another. (To the extent that this does not occur, some agencies have employed statistical standardization procedures to compensate for differences.) As August and Muraskin (1998) have shown for the OERI review process, standardization entails some very strong and somewhat unrealistic assumptions. When inter-rater agreement is low, implying that the peer rating system has low reliability, some have questioned the soundness of the entire review process.

To be valid, a measure must take account of the scientific merit of a proposal in all its complexity without becoming distracted or enchanted by other properties of the

proposal. But scientific merit is an abstract and multi-faceted quality, so valid review must attend to each of the many elements that make up a good proposal. The combined assessments of several diverse experts may be needed to achieve a rounded evaluation of a proposal. With a multifaceted proposal evaluated from several divergent perspectives, it is not surprising that inter-rater agreement may be low. Different experts might properly reach different judgments about the quality of the proposal when their particular area of concern is given central importance and evaluated through their particular set of epistemic lenses.

Peer review might be considered a problem in inductive inference: sounder inferences are those built upon broader and more varied foundations. Program managers seeking informed reviews of the various facets of a proposal would seek diverse experts and avoid duplicative perspectives. In effect, they are implicitly following Donald Campbell's "fish-scale model" of collective omniscience, which develops comprehensive understanding from a pattern of partially overlapping specializations. Put differently, given the limited number of reviews that can be elicited for any one proposal and the range of reviewer backgrounds necessary to cover the intellectual content of the proposal, it is not surprising that agreement is low. With a small number of reviews possible, the pursuit of validity in substantive coverage limits the level of agreement likely to be obtained.

Rigor-Responsiveness

Peer review is expected to demand analytic and methodological rigor in the studies it supports, yet at the same time to be responsive to emerging needs and

possibilities. (We are grateful to Meryl Bertenthal for suggesting this value tension.) A review system that favors rigor would support only those studies that use the strongest research designs and analytic approaches, rather than engaging in exploratory and speculative work. In contrast, a system that favors responsiveness in the work supported would relax methodological standards to address problems that are current and urgent, even if they are ungainly or ill-structured. Compelling problems demand to be studied with the best methods available, even if the results produced are only approximate: with time, theory, and experience the methods will catch up. This is a particular concern for education research as it is possible to overemphasize experimental rigor – in the quest for legitimacy and recognition among the sciences – at the expense of educationally vital research issues. We close with some considerations of how education research might negotiate the tradeoffs inherent in peer review.

Challenges Facing Peer Review for Education Research

Peer review does many things and serves many values in the process, but it cannot simultaneously deliver on all things equally well. Which purposes and which values are most important for which sorts of education research now – and in the near term? And which agencies should assume which responsibilities? The answer depends in part on whether one focuses narrowly on the mechanics of peer review or more expansively on the process of community building. Consider the following issues that might arise in the management of a research program:

- How should departments, agencies, and designated units within them establish and maintain or shift their agenda of research priorities?

- Given a set of research priorities, how should those priorities be funded – through individual projects, centers, or other arrangements (e.g., cooperative agreements, contracts, and work orders)?
- How much should reviewers be asked to consider, if anything, beyond intellectual merit of the proposals? And who should establish those criteria and develop tools for applying them in the review process?

In sum, are such matters “management prerogatives” to be decided within agencies, or community concerns to be brought before the reviewer community? Such decisions structure programs in fundamental and enduring ways, so it seems appropriate for reviewers, as representatives of the broader community, to participate in the process. NSF advisory panels often spend an hour or so at each of their semi-annual meetings discussing research opportunities and procedural issues, sometimes with representatives of NSF management—Division Directors or Assistant Directors—present and participating. As part of the agency’s “extended family,” the community takes a hand in shaping the conditions under which its research is supported.

Having a peer review policy and process in place is like having a score without an orchestra, a shop manual without a mechanic. The blueprint for peer review should be accompanied by a plan for developing a professional identity for program officers and their reviewer community. It is certainly possible to prescribe a lot of review procedures—rules for minority reports, for example, when one panelist’s score lies far outside the range. But it is less possible to anticipate everything that must be done to make the process run smoothly. Some learning, reflection, and calibration must occur within the agency and the reviewer community.

If the success rate is low—and 10-20% is low—it is quite likely that the review system will be asked to render judgments that are beyond the resolving power of evaluative process. In any competition the “pay line” (as NIH calls it) between the last proposal funded and the first proposal declined will inevitably appear to be arbitrary. But if that first decline is an excellent proposal by accomplished researchers, and if many such proposals are declined over time, then the legitimacy of the review system itself will be called into question. Despite the resources harbored by the agency, proposers and reviewers alike may be less willing to participate in the processes that lead to their award.

Similarly, involving the best researchers in the review process probably leads to better and more legitimate reviews—that is, reviews that will be accepted by the community. But such people are also the most likely proposal writers, and it is generally unwise to allow someone to review proposals for a small competition in which he or she is also a contestant. Strategies for handling such conflicts of interest would have to be developed and must be accepted by the community or the legitimacy of the process will erode. With such erosion, prepare for charges of cronyism.

Developing a review process that has widespread legitimacy and building a community of researchers are among the highest priorities for education research. Peer review is a strategic instrument for accomplishing these goals. We emphasize that community is more than a warm and fuzzy feeling: it entails responsibilities, relationships, shared purposes, consultative decision making, and trust that together combine to transform research findings into a body of knowledge, conjectures into theories, researchers into a knowledge community. Strong managers would help with these aims. Weak managers—those who “facilitate” the process without reading

proposals, entering discussions, and engaging reviewers—are of little value in this important endeavor.

In education research perhaps more than other fields, it will be important to construct a bridge between research and practice or policy. Legislation such as the U.S. Department of Education’s No Child Left Behind Act (NCLB) is a federally mandated call for accountability. Outcomes, by school, must be reported on a schedule. Many strings are attached. Repercussions are a certainty. There is no way that such a requirement, independent of state and local performance standards, can be fulfilled without redirecting education research and evaluation.

Concomitantly, it becomes essential to involve advocates and the engaged lay public in the review process. K-12 school districts are community- and neighborhood-bound. Parents and other residents will bring valuable perspectives to the table. Research-based findings should be one input – a resource for all – to the deliberations of school boards, PTAs, and teacher conferences. So while NCLB complies with the federal Government Performance and Results Act requirement that agencies show societal benefits for their investments of public funds, it becomes a driver that must also conform to vagaries of “local control” and statewide standards. Such intense engagement with practice, policy, and the public surely reduces distance as it expands scrutiny. But it also might weaken claims that education research – being a specialized, expert, professional enterprise – “knows better.” It is a difficult tradeoff, indeed an ongoing tension between discretion and audit, to be confronted in Washington, D.C., and across the U.S.

Finally, should the U.S. Department of Education develop a pluralistic model that combines peer review to allocate funds and develop a research program (in the manner of

NIH and NSF) with a strong-manager system (DARPA-style) targeted projects that address identified needs? The mission of the Institute for Educational Sciences (formerly OERI) is to support applied research, so perhaps the research program and community building can be left to NIH and NSF, which fund basic research on education. With a relatively modest budget and a low funding rate, it is hard to endorse further dividing functions and operations within IES. Yet there may be advantages to developing both a program of research intended to endure and cumulate, and a changing set of guided projects directed toward problems in practice, evaluation, and policy. Both the Administration and the Congress would likely weigh in – with impunity – on such decisions.

Whatever path is chosen, building community and all that that entails is an essential step toward a stronger national program of basic and applied research on education. To strengthen education research, scarce dollars will need to be awarded by agency stewards working hand-in-hand with a community of scholars and practitioners in defensible and measurably productive ways. Peer review is but a mechanism for making progress. More than ever in the early years of the 21st century, it must assure all that quality in teaching and learning is being served in the national interest.

References

August, Diane and Lana D. Muraskin. "Strengthening the Standards: Recommendations for OERI Peer Review." Bethesda, MD: August and Associates, 1998.

Chubin, Daryl E. and Edward J. Hackett. Peerless Science: Peer Review and U.S. Science Policy. Albany, NY: State University of New York Press, 1990.

Cicchetti, Dominic V. "The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation." Behavioral and Brain Sciences 14: 119-186, 1991.

Cook-Deegan, Robert Mullan. "Merit Review for Federally Funded Science and Technology: A White Paper for the Council of the National Academy of Sciences." Washington, D.C.: 1996.

Galison, Peter. Image and Logic: A Material Culture of Microphysics. Chicago: University of Chicago Press, 1997.

Gieryn, Thomas F. Cultural Boundaries of Science: Credibility on the Line. Chicago: University of Chicago Press, 1999.

Guston, David. Between Politics and Science: Assuring the Integrity and Productivity of Research. Cambridge, England; Cambridge University Press: 2000.

Kuhn, Thomas S. The Essential Tension. Chicago: University of Chicago Press, 1977.

Langfeldt, Liv. "The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome." Social Studies of Science 31 (6): 820-41, 2001.

National Academy of Public Administration. "A Study of the National Science Foundation's Criteria for Project Selection." Washington, D.C.: February 2001.

National Institutes of Health. "Recommendations for Change at the NIH's Center for Scientific Review: Phase I Report, Panel on Scientific Boundaries for Review." Bethesda, MD: 2000.

National Science Foundation. "Report to the National Science Board on the National Science Foundation Merit Review System: Fiscal Year 1999" (NSB-00-78). Arlington, VA: NSF, 2000.

Sarewitz, Daniel, Frontiers of Illusion: Science, Technology, and the Politics of Progress (Philadelphia, PA: Temple U. Press, 1996).

Starr, Susan Leigh and James R. Griesemer. "Institutional Ecology, 'translations,' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39" Social Studies of Science 19: 387-420, 1989.

U.S. General Accounting Office. "Peer Review: Reforms Needed to Insure Fairness in Federal Agency Grant Selection." (GAO-PEMD-94-1). Washington, D.C.: June 1994.

U.S. Congress, Office of Technology Assessment, Federally Funded Research: Decisions for a Decade, OTA-SET-490 (Washington, DC: U.S. Government Printing Office, May 1991).

U.S. General Accounting Office. "Peer Review Practices at Federal Science Agencies Vary." (GAO-RCED-99-99). Washington, D.C.: March 1999.