CALIBER
ASSOCIATES

EVALUATION OF THE EXPERT PANEL REVIEW SYSTEM FOR IDENTIFYING PROMISING AND EXEMPLARY PROGRAMS

Prepared by:

William A. Morrill
Maureen Murphy
Rebecca Adamson
Marianne Hooker

Caliber Associates
10530 Rosehaven Street
Suite 400
Fairfax, VA 22030
(703) 385-3200

March 1, 2001

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (CONT.)

Page

TABLE OF CONTENTS (CONT.)

Page

I. CONTEXTUAL BACKGROUND AND METHODOLOGY

I. CONTEXTUAL BACKGROUND AND METHODOLOGY

The review system that is the subject of this evaluation came into existence through Section 941(c), (d) and (e) of Title IX, Public Law 103-227, enacted by the Congress and signed on March 31, 1994 (copy appended). Congress, in this legislation, established a specific structure for the identification, designation and dissemination of promising and exemplary educational programs, as well as other related processes such as panel review of grant applications and contracts. The structure created for identifying promising and exemplary programs was not an entirely new idea, but a replacement for the existing national diffusion network (NDN). The replacement of the NDN with a new system was initiated in Congress, and accepted by the Administration. An analysis of that decision is not the subject of this assessment, but it appears to have occurred as the result of various complaints about the existing system not countered by champions for the NDN.

The new system came into existence not only with some history concerning the general objectives, but also with a set of contextual factors important to both its history and its future. Two widely shared and long held goals impel the process. First, educators, parents and students, researchers, policy makers and the public at large want the best, most effective educational programs to be available to schools for the education of children. That goal grows stronger as the increased emphasis on accountability reveals shortfalls in achieving positive educational experiences for *all* children. Second, all concerned want to be assured that educational programs recommended for application come with a known level of effectiveness. Given the unsatisfactory experience for too many students, three concerns take on added urgency: the educational challenges posed by an increasingly diverse student population, difficulties in replicating promising practice, and the need to identify what works and for whom.

In addition to the widely shared goals, the context for the implementation of a new review system includes the perceived and actual state of knowledge concerning effective educational programs, efforts to improve the state of this knowledge base, and the strategy to link the review system to other elements of knowledge development and utilization activities. Major components of this context are not conducive to the easy implementation of a new system.

A continuing stream of *assessments of the knowledge base*,[1] including reviews by the National Academy of Sciences (NAS), the National Academy of Education (NAE), and the National Educational Research Policy and Priorities Board (NERPP), have called attention to weakness in the documentation and lack of rigor in the evidence to support the knowledge base about effective educational programs. Few programs have been subjected to rigorous

[1] The meaning of the term knowledge base in this report is the cumulative state of knowledge (clear and certain apprehension of facts, truths or principles) about educational matters whether in written form or not. A related term, databases, is used to describe knowledge available in accessible written form.

evaluations, particularly in the past, and national clearinghouses publicize programs without persuasive evidence of quality. Difficulties in replicating attractive innovations and past experience with fads underlie the skepticism about the quality of the knowledge base. The problem is further complicated by the difficulty of assessing both the quality of educational materials and their use in live classroom settings. All of these circumstances create a substantial set of problems for a new review system intended to define and apply criteria and assess evidence for the selection of promising and exemplary programs.

The weakness in the knowledge base continues to be accompanied by a drastic *national under-investment in educational knowledge building* (research, development and communication). Studies done for the NERPP Board, the Panel on Educational Technology of the President's Committee of Advisors on Science and Technology (PCAST), NAS and NAE continue to call attention to this resource issue. Work done for the NERPP Board estimated national spending each year on educational research, development and communication represents *less than a quarter of one percent of total national educational spending* (estimates were required due to the lack of any collected comprehensive data). Such a level is miniscule compared with any similar public and private functions or enterprises, where levels for new knowledge building range from 5 to 10 percent of annual spending. *Thus, strengthening the body of empirical evidence of educational program effectiveness is unlikely to progress rapidly without substantially higher levels of annual resources. Specifically, rigorous research methodologies and large scale trials must be a more frequent component of educational knowledge building, and they are expensive. Likewise, there are unlikely to be enough well substantiated exemplary programs to produce a strong group of candidates until such resources are applied.*

The context described above concerning the state of the knowledge base and its improvement would lead logically to a set of modest and prudent objectives and expectations as to what might be achieved, and when, in a new review system to designate promising and exemplary programs. There was a sense of healthy caution among the research community and knowledgeable observers about the implementation of a revised system. *A different set of perceptions about the knowledge base as related to the new system existed, however, at a more popular level. These perceptions operated from the conviction that there was a substantial pool of exemplary and promising educational programs operating in the "real world," and that those could be found and designated through a straightforward, panel review system. The legislative language is based as much in the second perception as it is in the first.*

A final contextual issue of importance concerns the *linkage of the review system* to designate promising and exemplary programs *with other components of the knowledge building* (research and development) and communication (dissemination) activities from both a national and Departmental perspective. The pre-existing program was deeply imbedded in the dissemination system. The work of involved Departmental staff and the research community before and shortly after enactment contemplated not only continued linkage with the dissemination system, but also links to the research and development efforts of particularly promising programs. This integrated system concept—particularly with respect to dissemination—had been a continuing issue for the Department. Resource constraints and a multiplicity of demands also caused continuing problems in developing and enforcing a coherent knowledge building strategy with a firm set of priorities linked to national needs. A linked strategy was, nonetheless, part of the agenda for the new system.

The new review system came into existence with a strong set of common goals, but with a contextual situation full of issues and challenges to the translation of those broad goals into an effective operating system.

1. EVALUATION SPECIFICATIONS AND METHODOLOGY

This evaluation has been commissioned by the NERPP Board as the second in a series of assessments of standards recommended by the Board and promulgated by the Department to implement provisions of Title IX of Public Law 103-227. The first evaluation assessed the implementation of standards governing the peer review process for grant applications and contracts issued in September 1995 (known as the Phase I standards). This evaluation has been commissioned to review implementation of the standards recommended by the Board and adopted by the Department in November 1997. These standards govern the expert panel process for the review and selection of educational programs to be designated by the Secretary as promising and exemplary (known as the Phase II standards). A copy is appended.

The evaluation specifications called for an examination of OERI practices in implementing the new review system, including the activities of the four panels, a review of dissemination plans and an examination of analogous public processes. They also included the development of conclusions and recommendations with respect to future direction.

A three-part methodology has been used to accomplish the purposes of this evaluation. The first involved review of the written record: statutory provisions, initial standards, application guidelines, proceedings of workshops, descriptions of other processes, and material concerning the contextual background. For the most part, there is a substantial written record with occasional gaps, though the amount of comparative analysis about analogous public processes is sparse because there are few similar undertakings.

The second part of the methodology has been to conduct written and oral surveys of participants in the process, who have helped to document and provide comments on the process as they experienced it. All participating groups were included in a survey:

- The *panelists* from each of the four topical panels—Math and Science, Educational Technology, Gender Equity and Safe, Disciplined and Drug Free Schools. The chairs and selected other panelists were interviewed orally, and mail surveys were sent to the remainder.

- The *reviewers*[2] who screened and scored the applications for panel consideration. These reviewers carried several different labels over the course of time and among panels, but they are referred to as reviewers for consistency. One special crosscutting group known as the Impact Review Panel deserves separate mention because of its special character. This group was created for the Math and Science panel to provide a review of the quality of the evidence of effectiveness on the math applications. Its work was subsequently extended to reviews for other panels. The responses of the IRP reviewers have been grouped together with those of other reviewers in the summary data in this report.

- A random sample of *applicants* identified and contacted with the assistance of the Department's support contractor—RMC Research Corporation.

- The involved Departmental and supporting contractor staff. All of the Departmental staff and some of the supporting staff were interviewed orally.

The results of those surveys are contained in summary form throughout this report, organized according to the topic under discussion. In addition, a response rate summary and more detailed reports of the panelist, and reviewer and applicant surveys are appended to this report.

The third part of the methodology involved the use of a technical working group to provide expert opinion at critical junctures in the project. Each member brought to the discussions invaluable operational and analytic experience with the issues being confronted by the evaluation. Their thoughtful comments on the draft report have been taken into account. None has expressed opposition to report findings. Their names are provided in an appendix.

The following sections of this report are organized to reflect the major dimensions of the expert review system—goals and objectives, scope and criteria, processes, and staffing—followed by a section on conclusions, options and recommendations.

[2] Not all reviewers were identifiable, due to the loss of records identifying reviewers of the Gender Equity panel who were not members of the panel itself (members of this panel also served as initial reviewers). Thus, surveys were not attempted for this group.

II. Goals and Objectives

II. Goals and Objectives

As noted in the previous section, the goals of the review system are widely shared—to identify effective educational programs and to be able to assure teachers and school systems that those identified can be used with a known level of confidence as to their effectiveness in specific applications. The previous discussion of the contextual background sets up an immediate tension between the increasing urgency of the first goal and the less-than-desired strength of the knowledge base to help meet the second. Given the interplay between them, a central issue for the implementation of the goals becomes the degree to which programs can be identified as promising or exemplary in a way that enables States and school districts to make wise decisions about their use.

Knowledge is usually gained incrementally and cumulatively, and the knowledge sought for educational institutions is complex, as is the amassing of persuasive evidence that the policy and programs under consideration are effective. The desired improvement in the knowledge base can only be built over time. In the meantime, what are schools to do to make good choices? The quality of evidence and the ability to replicate effective performance are at the heart of the matter. Neither is an easy issue. Replication is imperfect in all human activities and, indeed, in most science. Difficult judgements are required.

The statute was specific about the process the Department was to follow for implementation, but much less specific about how a few very broad standards were to be used in making choices. Though present by implication, many of the detailed objectives—including reconciliation of competing objectives—were left to Departmental implementation through regulation or judgment in expert panel deliberations.

The definition of educational programs, for example, was not made clear in the statute, and the Department did not subsequently define it in a precise and uniform way, leaving to individual panels the determination of what "programs" they would review. The panels later made different choices as to what constituted a "program" in their assigned area. The definitions of promising and exemplary were equally broad.

The statute stated that exemplary programs would be determined based on empirical data that included, but were not solely limited to, test data and programs that could be implemented at the State, local and classroom level (Section 941(d)(2-4)). Promising programs were to be based on the judgment of experts and practitioners "that the program shows promise for improving student achievement …" (Section 941(d)(2); however, many of the details of what products were expected, what standards of evidence would be used in particular panels, and how the review system would fit into the Department's development and dissemination of knowledge of improved and high quality programs were left to Departmental implementation. We now turn to those more specific matters to evaluate the implementation of the expert panel system with respect to goals and objectives.

As a general matter, the Department has treated the implementation to date as a pilot or development project that will evolve through time rather than as a fixed system, making every effort to remain consistent with statutory prescription. Much has been achieved from this perspective, particularly given the contextual starting place. While the evaluation may concentrate more on lessons pointing to improvement or change, it does not intend any diminution of the contributions that have been made. Further, the challenges provided by the contextual background and many tough implementations issues faced by the Department should not be underestimated.

1. WHAT PRODUCTS WERE EXPECTED AND ACHIEVED?

The products explicitly anticipated in both legislation and regulation were the designations of programs as exemplary and promising. At least implicitly in legislation, and also explicitly reflected in the thinking of Departmental staff and in some of the views in the research community, additional research and development were also anticipated for designated programs to further enhance them, particularly with respect to programs designated as promising. This view was consistent with an integrated concept of the expert panel review system in which promising programs would be further developed and evaluated with Departmental involvement in order to move them to exemplary status or abandon further development.

As noted in the prior section, the statute and implementing regulations provided no standard definition as to what was expected with regard to the character of the educational program to be reviewed. The individual panels created their own expectations, each different and related to their topic. These ranged from curricula to systems and other program activities.

At this writing, the expected designations have been delivered almost completely. The follow-on additional research and development has yet to be planned and initiated for reasons discussed later in this section. In several panels, an unanticipated product has been developed. This product summarizes the panel's reflections and refinements on the criteria and scoring rubrics used and the material submitted as guidance to the field on future submissions, the directions of future research and development, and the desirable direction in innovation. A number of panelists regard this material as a far more important contribution than the particular designated programs. *Not all of this material has yet been published, and the Department should consider appropriate ways to make sure this potential contribution to the future of those fields does not get lost among issues related to release of the designations.*

A further aspect of the anticipated products is the manner of their dissemination. That topic is covered in the process section (Section IV).

2. WHAT STANDARDS OF EVIDENCE WERE SOUGHT AND DELIVERED?

The statute establishing the expert panel system paid modest attention to the specification of a process for developing the standards of evidence to be used in selecting exemplary or promising programs. Beyond the general specification of empirical data and the admonition to use, but not rely solely on, test data in the selection of exemplary programs, the statutory provisions provided scant guidance about standards. Indeed, the statutory structure and language mirrored the perception discussed in Section I that various topic fields already possessed the appropriate criteria and knowledge base, and that all that was needed was a process to "mine" them and designate the winners.

This perception, of course, ignores three or more decades of struggle to identify agreed upon standards to warrant effectiveness, resulting in an existing database largely devoid of more than anecdotal or case study evidence concerning the unambiguous impact of educational programs. With the Gender Equity panel and the Math and Science panel work beginning, the Department began to develop an implementation regulation including criteria for selection of exemplary and promising programs.

The initial efforts as reflected in the June 3, 1996, notice of proposed rulemaking started down the path of a single set of uniform criteria to be applied across all panels. Two initial panels were already dealing with quite different educational areas (gender equity and math), and new panels were likely to become more diverse. While this approach was a thoroughly understandable impulse, the commentary from the first two panels and others made clear that a single set of criteria was unlikely to be a workable idea in the short term, if ever. The final regulation issued by the Secretary on November 17,1997 backed off the original plans for a single criteria set off "until the work of all four pilot panels is concluded…." What remained in the final regulation were four terse statements of criteria with all details to be added by the individual panels for each topic. These criteria were: "(a) evidence of success, (b) quality of the program, (c) educational significance, (d) replicability."

As we discuss in more detail in the next section, the four panels then moved to add their own statement of criteria within each of the categories, some adopting several and others only a single criterion within each. These criteria were included in the solicitation of applications (copies appended). All were somewhat different in content or expression. We later discuss those differences, the differing importance attached to the four categories, and what is likely to become of the notion of a single set of criteria. It is appropriate to note here that the respondents to the survey were generally satisfied with the four categories, though those who made specific comments about the matter noted their overlapping nature. *It is our judgment that the critique is correct, but it does not represent a major problem relative to others needing more urgent attention.*

The major point to be made, however, is the continuing absence of and unresolved issues about the *standards of evidence* to be used to identify exemplary programs with respect to evaluation methodology and measures. This problem impacts not just the expert panel review system but the entire range of education research, development and evaluation (the knowledge base). While comparison group and time series methodologies, including the random assignment "gold standard," are well recognized routes to assessing effectiveness, such studies exist in such small numbers that sole reliance on these methodologies will produce an inadequate pool of designation candidates, absent a large increase in research and evaluation activities and resources. Alternative methodologies to help close the gap are much discussed and little pursued. This situation leaves the Department in a quandary as to what to do in meeting the program goal.

The Department recognizes this problem, as reflected in its continuing efforts to pursue the issue of standards of evidence more generally. Specifically, the work of the Rand Corporation on research standards and the assignment to the National Academy of Sciences both seek progress on the issues. External work underway in the Campbell Collaboration and the recently initiated Education Qualify Institute are pursuing the issues from somewhat different perspectives. All recognize the evolutionary nature of the enterprise and the strongly held viewpoints involved.

Given the experience to date with the expert panel system, it seems likely that the *measures* in criteria sets will continue to need to be different among panels, particularly if the Department continues to identify a diverse set of programs that include curricula, cross-cutting programs and technology among the topics for designation. *It is less clear, however, that the methodological basis for producing evidence should vary substantially, and present variation appears more an issue of resources and time rather than of merit.*

3. WHOSE PRODUCT IS IT?

The statute unambiguously identifies the designations to be those of the Secretary of Education. While the Secretary may not, under the statute, designate a program not recommended by an appropriate panel, the Secretary is free to accept or reject panel recommendations on whatever grounds he or she sees fit, with or without explanation.

This process contrasts sharply with that of the National Research Council (NRC) of the National Academy of Science, and may for this and other reasons be a cause for reconsidering the structure. In the National Research Council process, appointed panels are vested with the authority to make the appropriate technical judgments, subject to an NRC review process and the authority of its president to assess the quality of the evidence supporting the panel findings, seeking clarification or amendment where the evidence is inadequate to support the findings, or as a last resort, denying publication of the panel findings for explicit reasons. Thus, the findings of the panels reflect the best evidence and judgment of the panel members rather than the corporate position of the National Academy of Science, and are understood to be the best assessments of the knowledge about a topic by a group of experts at a given point in time.

Such an approach has considerable attraction for the expert panel system for exemplary and promising programs on both factual and conceptual grounds. As a factual matter, the panels and their reviewers have supplied work, expertise and judgment about the programs. The Departmental staff has neither the time nor the resources to second-guess the panels on behalf of the Secretary, and it may also lack the expertise to do so. The law already specifies that no program may be included without panel recommendation, and the legislative process already provides for the Secretary and his or her staff to offer programs to be reviewed by the panels (an authority not exercised to date in the system). In short, identifying the technical judgments with the panel is consistent with the current and likely future fact case.

That approach is also appealing for contextual reasons. Both provisions of law and long-standing political convictions have sought to keep Federal officials out of curriculum matters, yet the Secretary under this program is designating curricula and other programs as exemplary. A Congressional hearing has already made an issue of this question. Further, as a political appointee, the Secretary is subject to interventions from other political actors of import to the Department on behalf of one or another program. Such pressure may be more easily deflected in a process that vests technical judgments with the panels. In reviewing possible legislative changes to improve the program, the NRC model warrants serious consideration.

4. WHERE DOES THE REVIEW SYSTEM FIT IN THE DEPARTMENT’S RESPONSIBILITY TO DEVELOP AND COMMUNICATE IMPROVED, HIGH QUALITY PROGRAMS?

The statute clearly linked the new expert panel system to dissemination activities, embedding the provisions within its section dealing with a national dissemination system and referring explicitly to a variety of instrumentalities for the communication of designated programs to concerned audiences. While the statute did not take the linkage further, thinking within the Department and in some of the early meetings foresaw a closed loop, integrated system that linked new research and evaluations with technical assistance to the designation process and communication of results.

While the developmental concept was clearly dominant at the beginning, the implementation through the first four panels has been highly concentrated on the designation process alone, with essentially no linkages to research, evaluation and technical assistance, and only modest efforts on the communication of results. Much of this outcome has to do with limited resources in terms of staff capacities and budget funds for linkage activities. Some of this outcome, however, is attributable to choices and developments in the implementation that lead the review system away from its developmental perspective.

Competition is an almost inevitable part of a designation system such as this one because choices are made among programs as to which meet certain tests required for designation and which do not. The competitive aspects of the designation create tensions with the developmental perspective, particularly if those aspects are emphasized. The implementation decisions that emphasized the competitive dimension of the review system include the use of an application process rather than a canvas of the field to identify programs for consideration. There was also the promise to hold the identity of applicants not achieving designation confidential to encourage rather than discourage submissions. Further, in the view of some panelists and reviewers, the strong criteria with respect to the provision of evidence of effectiveness in multiple settings and student groups led in the case of curricular materials programs toward submissions from commercial publishers who had the scope and resources to produce the desired evidence. For this group of applicants, gaining or not getting a designation produced a "high stakes" game with important potential consequences.

While none of these choices were poorly motivated (indeed, the emphasis on evidence was highly desirable), they all pushed the review system in the direction of a "contest" rather than a developmental activity. In a contest environment, the interaction between contestant and reviewer tends to be highly regulated and limited, and the review process is rigid and uniform. Neither of these features is useful in a developmental environment. There is also a question whether Federal support is desirable or needed in the dissemination of commercial curricular materials for which the publishers have substantial incentives to include their designations as a part of their marketing activities.

Beyond the competitive aspects of the review process, the selection of topics for a panel represents an important opportunity for linkage with the priorities selected for the Department's research, development and communication activities. The selection of topics also provides an opportunity for input from the field about urgent problems and information needs. The topics selected in the initial round of implementation appear to have been chosen on more pragmatic grounds from the panoply of Departmental program interests. These programs, all important in their own right, were picked because they were ready to go, had resources available, or were not otherwise being addressed. Since the implementation was initiated, the Department has moved forward in developing its knowledge building priorities in a more strategic fashion. *In the future, those priorities, together with an assessment of the state of the knowledge base to support a panel, appear to be central criteria for selection or continuation of topics for panels.*

While other, more detailed, points about linkage are raised in other sections, *this assessment concludes the developmental perspective originally envisioned by the Department with strong linkage between the expert panel review system and other related components of the knowledge building and communication national program is appropriate. Further, if that perspective is to be maintained, the competitive dimensions of the system should be de-emphasized, recognizing that some competition will always be present in the system.*

III. Scope and Criteria

III. SCOPE AND CRITERIA

This section turns to questions about the scope and criteria used in the initial implementation of expert panel review system. Some issues concerning both scope and criteria were introduced in the prior section. In questions of scope, the prior section dealt with the selection of topics in the framework of linkage to other components of the knowledge building and communication programs and priorities, and the strength of the knowledge base to support a panel. Here, we deal more directly with the scope of individual panel and its impact on their reviews. With respect to criteria, we dealt generally in the prior section with issues concerning the strength of the evidentiary standards; and here we deal specifically with the criteria used by the panels.

In addition, this section deals with how the panels distinguished between promising and exemplary programs, and the implications of the criteria used to distinguish them on the future use of these two categories.

1. WHAT ISSUES DID THE SCOPE AND TOPIC OF THE PANELS PRESENT?

For the most part, the scope of the panels presented no unusual issues. Gender Equity, Safe Disciplined and Drug Free Schools (SDDFS) and Education Technology reflected a coherent set of activities and programs around which a group of experts and practitioners could be structured to review and designate applications without too much of a stretch. Math and science was quite another matter.

The maturity and substantial specialization in the fields of math and science presented a real challenge in assembling a sufficient mass of expertise to review a set of unknown proposals that could potentially cover a highly varied set of specialties in both fields. Respondents who were interviewed about this panel overwhelming believe that its scope was too large, and would split math and science in the future. While the splitting could be handled in various ways, it is important for all panels to be organized with a scope that permits a panel of reasonable size and consisting of experts with current experience with theory and practice in educational programs to be organized and supported by competent reviewers.

The math and science panel was composed of intelligent members who worked seriously at their assignments, but the lack of depth in mathematics on the panel was not helpful when the math designations were attacked after initial release by the Secretary. The accuracy or inaccuracy of the attack is not within the scope of this assessment, but the scope of the panel, particularly in areas of high potential controversy, needs to be carefully considered before it is established.

Turning from scope to content, earlier sections commented on the wide topical variation in the selection of the first four panels, from core curricular topics in math and science to less traditional and crosscutting topics such as gender equity, SDDFS and technology. We also discussed the pragmatic rather than strategic basis for their selection.

Reading was an early candidate for selection that would have provided a base of two curricular and two crosscutting panel types. This potential topic was dropped because of the initiation of an NRC review of early reading. The curricular topics present a somewhat different, and perhaps slightly easier, challenge than do the crosscutting topics, but each type has its place in a group of panels in a designation system.

The earlier discussion identified two criteria that would appear important to panel topic selection together with other criteria developed for assessing the topic panels:

- The degree to which the topic is related to knowledge building and communication of national priorities
- The degree to which criteria exist or can be developed quickly to distinguish quality on an objective basis
- The degree to which the knowledge base and related data exist to support the gathering of credible evidence against all criteria for panel deliberations
- The degree to which a manageable and skilled panel can be organized, staffed and supported to complete its work.

The four pilot panels were deficient on one or more of the criteria. Math and science and educational technology were most firmly rooted in national priorities. All panels were faced with developing their own criteria, though the math and science panel could draw on prior work by AAAS and NSF. For different reasons, the knowledge and databases were less than desirable. Math and science education are, for example, more mature topics about which there is substantial literature; however, the science base undergirding the learning of math and the sharp disagreements about appropriate pedagogy among some math scholars provide a less than ideal knowledge base from which to make judgments. The crosscutting topics are all far less mature, and thus still building a knowledge base. All panels faced a substantial management challenge as start up pilot activities with limited resources.

Were the Department to start over at this juncture with the implementation of a new review system, it might well select a different mix of topics. However, given the context and the pragmatic constraints, there are limited grounds for criticism in retrospect, and the pilot panel implementation, as we shall describe, provides highly useful learning for the future. The selection of topics in the future should be more systematic and strategic with respect to both the topics selected and their frequency, as the report later will discuss. The criteria identified above should prove useful as part of such future decision making.

2. WHAT CRITERIA WERE ADOPTED?

Each panel was tasked at its outset with the development of specific criteria for the four different categories specified in the implementing regulations. Relatively little specific guidance was provided, though there was some clear learning in progress from the initial panels to the later ones. An example of the increasingly sophisticated approach was the process of the Educational Technology panel that pretested its criteria in the field before developing a final set. All of the panels took this task seriously, and struggled with it.

Many panelists and reviewers said they would make changes in specific criteria in a subsequent round of program selections for designation. The Educational Technology panel as a group formally revisited the criteria and created a set of scoring rubrics for each as part of the completion of its work. *The work on criteria can be seen as one of the major contributions of the panels to the initial round of implementation, more important in some of their minds (and ours) than the selection of particular programs.*

Table 1 summarizes the criteria adopted by panels for review of the initial applications. More detail can be found in the appended application solicitations. As will be quickly seen, there is substantial variation among the panels and between the categories with respect to the number of criteria, specificity and objectivity. With the exception of Educational Technology, the largest number of criteria were produced for the quality of the program, and the Educational Technology panel strengthened its quality criteria in the closely allied educational significance category. The least number of criteria showed up in the evidence of effectiveness and for the most part lacked specificity, a comment also made by a number of applicants who responded to the sample survey. This brevity is interesting given the responses to surveys by panelists and reviewers ascribing great importance to this category, discussed further below.

A substantial amount of attention was paid to the selection criteria by panelists, reviewers and supporting staff. The effort was made not only to get reactions to the usefulness of particular criteria, but also more particularly to assess the relative importance of the criteria and what the respondents observed about the quality of the submissions with respect to it. A more detailed summary of the panelist, reviewer and applicant survey responses is to be found in the appendix.

The panelists and reviewers assigned the greatest significance to evidence of effectiveness and success, followed by the quality of the program, a ranking echoed in supporting staff oral interviews. The respondents would give more emphasis to both of these categories of criteria, in the same order of importance. Perhaps not surprisingly, the quality of the evidence submitted to support the effectiveness of the program was regarded by panelists as average to poor, in equal numbers, while the reviewers ranked it as average to poor, in a 2-to-1 ratio.

EXHIBIT 1:
MATRIX OF CRITERIA

Quality of Program

MATH AND SCIENCE:

1. The program's learning goals are challenging, clear, and appropriate for the intended student population.
2. The program's content is aligned with its learning goals, and is accurate and appropriate for the intended student population.
3. The program's instructional design is appropriate, engaging, and motivating for the intended student population.
4. The program's system of assessment is appropriate and designed to provide accurate information about student learning and to guide teachers' instructional decisions.

EDUCATIONAL TECHNOLOGY:

1. The program addresses an important educational issue or issues and articulates its goals and design clearly.

GENDER EQUITY:

1. Is based on sound theory and practice.
2. Has up-to-date and accurate content that reflects current law.
3. Is free of stereotypes and bias.
4. Is organized and well written.
5. Is engaging, appealing, and easy-to-use.

SAFE SCHOOLS:

1. The program's goals with respect to changing behavior and/or risk and protective factors are clear and appropriate for the intended population and setting.
2. The rationale underlying the program is clearly stated, and the program's content and processes are aligned with its goals.
3. The program's content takes into consideration the characteristics of the intended population and setting and the needs implied by these characteristics.The program implementation process effectively engages the intended population.

Usefulness to Others

MATH AND SCIENCE:

1. The program can be successfully implemented, adopted, or adapted in multiple educational settings.

EDUCATIONAL TECHNOLOGY:

1. The program is adaptable for use in multiple contexts.

GENDER EQUITY:

1. Is described in a tangible way that others can use.
2. Is affordable in terms of time, money, and human resources.
3. Is accessible and available to others.
4. Takes into account characteristics of special populations, for example, students with disabilities, students who have English as a second language, students of color.

SAFE SCHOOLS:

1. The program provides necessary information and guidance for replication in other appropriate settings.

Educational Significance

MATH AND SCIENCE:

1. The program's learning goals reflect the vision promoted in national standards in math and science education.
2. The program addresses important individual and societal needs.

EDUCATIONAL TECHNOLOGY:

1. The program develops complex learning and thinking skills for its target audience.
2. The program contributes to educational excellence for all.
3. The program promotes coherent organizational change.

GENDER EQUITY:

1. Draws strategies from different fields, such as public health, criminal justice, and social justice.
2. Considers current consensus on how to address issues.
3. Demonstrate improvements over alternative approaches to the challenge.

SAFE SCHOOLS:

1. The application describes how the program is integrated into schools' educational missions.

Evidence of Effectiveness and Success

MATH AND SCIENCE:

1. The program makes a measurable difference in student learning.

EDUCATIONAL TECHNOLOGY:

1. The program has rigorous, measurable evidence for its achievements for at least one of the three educational significance criterion (learning, equity, and organizational change).

GENDER EQUITY:

1. Process measures
2. Outcome measures
3. Characteristics of successful policies, practices, programs or products

SAFE SCHOOLS:

1. The program reports relevant evidence of efficacy/effectiveness based on a methodologically sound evaluation.

What came through consistently in respondent judgments is the need to strengthen the emphasis on evidence of effectiveness and other indicators of quality, a judgment with which we concur. This judgment carries with it significant implications not only for the expert panel system, but for the Department's research and evaluation program as well. *Strengthening effectiveness criteria will require more acceptable effectiveness methodologies; more work to integrate quantitative and qualitative studies and expand the measurement tools; and many more resources to collect the data and conduct the analyses.*

Judgments about the appropriate direction of the criteria also carry with them implications for handling promising, as opposed to exemplary, programs. These implications are discussed in answer to the following question.

3. HOW WERE EXEMPLARY VERSUS PROMISING PROGRAMS DISTINGUISHED?

The statute provided general distinction between exemplary and promising programs that focused on the strength of the evidence supporting their effectiveness or success. The Department's draft regulations suggested a more precise specification by indicating that the strength of the evidence would mean generalizability. Exemplary programs would be expected to demonstrate effectiveness in each of the four categories of criteria in multiple contexts or locations, while promising programs would also meet all four criteria, but not in multiple contexts or locations.

Commentors disagreed, suggesting promising programs might be found that did not meet all criteria, that the evidentiary requirements for promising programs were too stringent, or that the distinctions were too narrow and arbitrary. The Secretary was persuaded by the critiques, so the final regulation backed the distinction off to the more general statutory formulation and the judgment of individual panels.

The panels took a relatively standard approach to the issue, with some specific differences. All panels provided careful descriptions of the review process they intended to follow, including varying detail about the scoring plans. Two panels did not provide further guidance to the applicants on how they intended to distinguish between exemplary and promising, leaving that to the judgment of the panel as the result of the reviews. Math and Science provided some guidance on how they would approach the distinction, and the SDDFS panel supplied a description of the threshold scores needed for promising and exemplary designations.

The panelist and reviewer interviews reveal the centrality of the evidence of effectiveness and quality of the program in the criteria they used, in general, as the basis for determining the differences between exemplary and promising. Respondents overwhelmingly indicated they were comfortable with the distinction. There seems to be little doubt the system worked reasonably well in the pilot implementation.

Two questions remain for the long run. First, if the direction of the criteria for the future of the system is to give more emphasis to the evidence of effectiveness and quality of program categories, as most suggest it should, won't that direction have implications for the promising programs under a single review and criteria structure?

It is possible that, as the quality and evidentiary criteria applicable to all programs under consideration become stronger, promising programs will either begin to look like junior versions of exemplary programs—all well along in the development cycle—or, more likely, promising new and innovative ideas will begin to get squeezed out of the system. It is not clear such a trend would be desirable. *Consideration should be given to providing separate criteria or a separate review cycle for promising programs, in which the procedural rules are more nearly designed for a developmental activity (e.g., interaction with program staff about research or evaluation plans), as opposed to a set of criteria appropriate for judging fully developed exemplary programs.*

Second, the limited amount of rigorous research and evaluation that exists to support tougher effectiveness criteria and high evidentiary standards suggests that either there will be a paucity of good candidates for exemplary designations, or there will be pressure to ease the standards. This concern will be particularly acute until there is an increase in the level of resources to support more research and evaluation. Comprehensive exemplary and promising review cycles raise expectations that might be made more realistic with alternative approaches. *One way to ease an almost certain transition period would be to run promising only designation cycles on appropriate topics until the knowledge base and candidate programs are strong enough to support high evidentiary standards.*

Either of these alternative formulations appears to be consistent with the statute, and might be undertaken on a developmental basis. Though they are consistent with law, it appears wise to discuss such approaches with Congress if the Department decides to proceed with such a plan.

IV. PROCESSES

IV. PROCESSES

Having considered goals and objectives and scope and criteria, we now turn to processes the pilot panels used, from the determination of how programs would be solicited through the process by which programs were considered and selected. In this section, we also include the issue of communication (i.e., dissemination), though very little actual experience exists with this topic because so little time has passed since the panels completed their work and the designations were announced (some are just now completing the cycle). Finally, this section considers whether and how the reviews should be continued.

1. HOW WERE CANDIDATE PROGRAMS SOLICITED?

The authorizing statute provided relatively extended guidance as to the sources for seeking candidate programs for consideration in the designation review process. It called for a Departmental process to:

- Work closely with OERI internal institutions – the Institutes, ERIC and ORAD
- Review successful programs supported elsewhere in the Department
- Seek candidates from other Federal agencies sponsoring education programs
- Reach out to external institutions that might be operating or developing candidate programs.

The implication of the statutory language suggested quite substantial work by OERI to assemble a list of likely candidates. That approach presented a range of practical problems, and the Department moved in another direction—to solicit candidates on an application basis. The application approach was outlined in the regulations, and subsequently in the solicitation material (See Appendix).

The practical problems of a comprehensive approach were threefold: the potential volume of material to examine was enormous; the identification of programs and their status in existing data sets were likely to be unreliable; and the amount of evaluation materials was known to be limited. As a practical matter, an application approach was the only feasible choice.

For the long run, however, an application process has shortcomings relative to a more comprehensive "screening." While the application process permits judging the quality and effectiveness of a particular program against stated criteria, it does not facilitate identifying best practice in the field, since the universe is defined by the applications received rather than the full range of programs. An application process emphasizes the competitive features and the contest environment of the designation system. An application process with substantial information and supporting data requirements may pose such a large burden on small organizations that some promising programs are never submitted. *While an application process may be a practical necessity for several rounds of review, the Department should consider moving to a more comprehensive approach to assembling programs for review, particularly in the case of exemplary programs.* The objective for exemplary programs should be to move toward a capacity to identify best practice in continuing panels dealing with important topics for student learning and achievement.

2. WHAT WERE THE CHARACTERISTICS OF THE PROCESSES EMPLOYED?

While the statute prescribed that less than one-third of the members of a panel could be in Federal employment, it provided no other guidance on panel processes. Further, when the Department decided to delegate the determination of criteria to the panels, it also delegated authority to establish all of the procedures through which the criteria would be applied, subject only to a budget constraint.

Because of common practice among panel review systems, certain basic structural components of the process were the same across panels. All panels had reviewers screen and score the applications for the panel (generally at least two reviewers per application). All panels paid particular attention to the evidence of effectiveness, and a special panel, known as the Impact Review Panel (IRP), was set up by the Math and Science panel to review the evidence of effectiveness for math applications. Gender Equity and Educational Technology reviewed the effectiveness evidence, but also sought IRP review. Though the SDDFS panel wanted their own reviewers of the evidence criteria, they also were forced to take IRP review of effectiveness criteria.

Beyond that commonality, the remaining procedures and processes were highly varied in detail among panels and even within panels through time, reflecting the developmental perspective in the review system. The sequence of reviews with respect to different criteria (e.g., does the effectiveness review or the quality review come first or last?), the number of reviews and reviewers, scoring rubrics, documentation, the number of panelists who read the applications and reviews, and other process all varied between panels and over time. One early panel had interactive exchanges between the panel and the applicants, while later panels had none. One panel wanted to do some site visits to confirm program claims, and were denied the opportunity by lack of budget resources. Support staff were put in the awkward position of denying some desired processes on budget grounds, notwithstanding panel authority to do what they wanted. Yet in other cases, support staff were unable to argue with procedures that were less commonly used, since the panels understood that they had complete discretion in such matters.

From a developmental perspective, the delegation of authority over procedures in a new system made sense, and panels and their staff learned many valuable lessons from the philosophical to the mundane (e.g., never reschedule a meeting of busy panelists once it is set). Yet neither the support staff nor higher levels of the Department would favor a repeat of the complete delegation of the first round. Nor, in fact, does there appear to be any substantial justification for as much delegation and variation that occurred. *While the variation in process directly related to panel-specific features of criteria might be warranted, rules with respect to the number of reviewers, interactions with applicants, documentation of actions, and review rules at the panel level should be made uniform and standard within and between panels. This requirement is particularly true as long as the panels remain a direct Federal operation.*

The variation in processes among and within the panels proved particularly troublesome, and will continue to be so if the panel's topic or its actions stir up controversy and challenges to appropriateness or fairness in its selections (the competitive or high stakes contest dimension of the system). Variation in process treatment will strengthen challenges to the fairness of the procedure, however warranted such variation may be from a developmental perspective. *All variations will need to be carefully justified and documented.*

In operational terms, panelists and reviewers alike considered the processes adopted by the first four panels appropriately flexible for their purposes. As discussed above, there is every reason to believe the process suffered from too much rather than too little flexibility. The responses to questions about the duration of the process and the burden on the participants were a somewhat different matter. Most respondents indicated the process took far longer than they imagined at the outset.

Part of the extended duration of the process can be attributed to the start-up cycle, which will not have to be repeated in subsequent cycles on the same topics. New topics will require some additional time to develop criteria and associated review processes. Part of the elongation of the process, however, appears to be occurring in the Departmental review and processing after the panel's work is completed. Some of the additional review is associated with controversy surrounding panel results. *Some, but not all, of that processing may be reduced with subsequent cycles, but Departmental planning should aim to compress this review to the minimum.* Compression may be aided by the suggestion made in the goals and objectives section regarding the role of the Secretary.

Applicants also were troubled by the duration of the process, particularly because they had little or no information about its likely duration or the status of their applications. Only the Educational Technology panel provided a set of timing expectations. As long as the application process exists, the Department will need to provide better information, if it wants to maintain incentives for new or repeat applications.

Finally, with respect to burden, the panelists and reviewers had somewhat different views. Although panelists contributed far more hours than reviewers, only a third of the panelists regard the burden of their service as high, and almost 60 percent regarded the burden as acceptable. The reviewers, by contrast, were more evenly divided between reporting a high burden and an acceptable burden. The survey did not reveal why the reviewers regard their assignment as more burdensome, but *the finding suggests that more care should be taken in the future to explain to reviewers in advance about the nature of the commitment they are making.*

3. WHAT COMPLICATIONS DO FEDERAL GENERAL PROCEDURAL RULES IMPOSE?

Two general rules of particular import to the expert panel review system are the Freedom of Information Act (FOIA) and the Federal Advisory Committee Act (FACA). FOIA provides rules concerning the public availability of information in the possession of the Federal government that, in general, seek to maximize the information available to the public on request and limit the grounds on which the government can decline to divulge information (e.g., national security, personal data, privileged information, and proprietary data). FACA has similar purposes with respect to Federal advisory committees, by maximizing the amount of their activities open to the public and minimizing the activities conducted in private.

Neither of these two general rules was explicitly applied to the work of the four pilot panels, but questions pertaining to both have arisen during the initial implementation. In the case of FOIA, the rules have been cited in connection with the controversy over the math designations, which led to requests for release of the identities of all applicants for designation. Unsuccessful applicants had been promised confidentiality by the program in order to encourage applications. Executive branch legal review initially led to the conclusion that applicant identity could be protected as promised only if the applicant could write a persuasive case alleging damage in the event of disclosure. Applicants were so advised, and a limited number wrote such a persuasive case for confidentiality, thus avoiding disclosure. *Whether confidentiality promises can be defended in the future may be somewhat uncertain.*

The question is whether the lack of confidentiality will discourage applications. A random survey of applicants asked this question directly just at the time applicants became aware of the Department's difficulties in delivering on its confidentiality pledge. Eighty percent of the respondents indicated they would have applied without the confidentiality pledge, though many complained about the Department's making a promise on which it could not deliver. *This evidence suggests that the Department may be able to proceed in the future without the confidentiality pledge, without risking a major deterioration in applications.*

The FACA rules present a more troubling case. Although the pilot implementation avoided the rules altogether, and no decision has been made by the Department about the future, it seems unlikely to knowledgeable Department and evaluation project staff that these rules can be avoided in the future. As a general matter, the rules impose more formality with respect to public notice and the conduct of meetings, including what is open and what is not. For the most part, these rules add bureaucracy, time and expense, but do not interfere with committee operations. The exception concerns matters that may not be dealt with in closed executive sessions, and this may include deliberations about designated programs. It would present a serious problem for the panels to have to deal in open session with deliberations over which applications should receive designations and which should not. Such sessions should be candid assessments of applications, their evidence and professional colleagues. Open sessions are likely to discourage candor. When faced by a court-imposed application of FACA to the National Academy of Science and the National Academy of Public Administration, those organizations successfully sought a Congressional exemption from FACA. They did so with an agreement to follow FACA rules in general, but continue to hold judgmental meetings in executive session.

Full FACA rules would likely be damaging to the expert panel review system in a similar way, and should be avoided. *While applying most FACA rules in the future, the Department should seek an administrative or, if necessary, a legislative exemption for panel decision making sessions on applications.*

4. HOW VALID WERE THE PROCESSES?

As described earlier, the processes used in the initial implementation were consistent in their tiered structure with processes routinely used in large-scale reviews of applications or program evaluations. Thus, the structural features of the processes raise no particular questions about validity. The quality and defensibility of the processes beyond the basic structure also require attention. Some of these other dimensions, improved and made more sophisticated as the initial implementation proceeded, can be further improved in the future.

Some of the *criteria* were quite general, leaving uncertainty as to what was being assessed and how, particularly with respect to the effectiveness criteria. This meant applicants and initial reviewers were unclear as to what was wanted or acceptable. Respondents to the applicant survey indicated they generally understood the criteria, but indicated uncertainty about how much detailed description and evidence was wanted and even more uncertainty about the panel and Departmental processes in reviewing the applications. For example, few programs supplied a “bad news” evaluation or information about limitations of their programs. *It is interesting that 80 percent of respondents said they would provide such information if it was requested.*

The clarity of the application of criteria improved during the course of the initial implementation and can reasonably be expected to continue to improve in future rounds with more work on the specificity of criteria. *An important dimension of this clarification should include communication with the program developer and applicant community about panel assessments of criteria applications at the end of a review round and again at the beginning of the next, as some panels are seeking to do.*

Another crucial ingredient of valid processes is the training and organizing of panelists and reviewers. We will return to this topic in the next section on staffing, but note here the procedural validity aspects of such training. Other review systems pay particular attention to this dimension, especially when there are a large number of reviewers involved in the initial screening. Only careful training and organization can produce sufficient consistency to assure criteria are uniformly achieved. High levels of consistency are particularly important to a sense of fairness in competitive situations.

Two techniques are often used by other review systems to strengthen consistency. First, in the training of reviewers (and panelists), consistent application of criteria in ranking is strengthened by the use of examples, often real applications, in simulated scoring exercises. As a new review system in this case, there was no substantial supply of real examples. One panel did, however, make use of an early application, and another panel created illustrative examples. *The routine use of this procedure in the future will contribute to stronger validity.* Second, the use of experienced reviewers from previous cycles, the matching of experienced and inexperienced reviewers, and the rotation of review assignments during a review cycle are other techniques that can strengthen consistency.

A number of panelists and reviewers, particularly on the crosscutting panels, indicated a strong interest in *on-site visits* to the applicants as part of the process to validate applicant claims. Particularly where the criteria are qualitative in nature, on-site observation is considered an important component of establishing validity. This addition is particularly important where independent evaluation is weak or absent. Resource limitations prevented the implementation of this practice. *For the future, the Department should provide for the use of such review techniques where warranted by the topic under review, and budget the necessary resources.*

Finally, this analysis included an earlier discussion of the problems associated with the variation within and among the panels, some of it related to different topics and criteria, but some related solely to panel choices. We urged minimizing the latter. *Uniformity is no guarantor of validity, but it clearly increases the defensibility of panel results as a matter of fairness.*

5. HOW HAS COMMUNICATION OF RESULTS BEEN CONTEMPLATED AND DELIVERED?

Earlier sections mentioned the considerable emphasis placed on communication of results in both the statute and early Departmental staff thinking. Through time, the expansiveness of these conceptions eroded due mainly to budget limitations and the controversy surrounding some designations. Further, communications between the Department and the applicants during the longer-than-expected duration of the process has been scant.

The mathematics designations were released with a briefing and a glossy publication, along with information in the Department's electronic systems. Since the controversy arose about the math selections, the release material has been more modest, usually a brief description of the designees released at a professional meeting, though the other panels plan more elaborate publications. The results were also posted on the Department's Web site. *While not at odds with the letter of the law, these modest steps are far less than what was originally contemplated.*

Particularly in the Math and Science panel, where the designations were aimed at curricula, the prominence of materials by commercial publishers raised a further question as to what Federal support was necessary or appropriate to communicate the results.

The interviewing in this project raised the question of communication strategy and the level of Federal support for commercial products with panelists, reviewers and support staff. Opinion varied widely on most questions, except the appropriate targets for communication activities. There was substantial agreement that teachers and school system administrators at the local and State level were the appropriate primary targets, though other concerned constituencies—the research community, policy makers, academic departments, professional organizations and the public—were also mentioned by many as appropriate audiences for the results of the reviews.

When asked about a choice among Federal strategies for communication, opinion was much more divided. The majority of panelists, reviewers and staff supported an ongoing, long-term and comprehensive Federal role in communication, but had different views about the appropriate way to disseminate review findings. Their views on what was appropriate ranged from one-time announcements with electronic information to a more proactive effort with professional organizations and the research community. With respect to designations awarded to commercial publishers, the majority of respondents thought the Federal government should play a moderate, but active role in dissemination.

When we review the communication activity, *it is our judgment that the communication strategy should be ongoing, long-term and comprehensive, as well as linked to other*

Departmental activity as discussed in section II. Resources may need to be moderate in the short run, but will need to grow if goals for this program are to be achieved.

6. HOW HAS THE PANEL PROCESS BEEN MANAGED?

The panel process was assigned for administration to the Office of Reform Assistance and Dissemination (ORAD) within OERI. ORAD in turn appointed a general manager for the overall program and a set of staff coordinators for each of the four panels. These coordinators and the panels were in turn supported by external contractors, who are doing most of the facilitation work. After several different support contractors in the early stages of implementation, RMC Research Corporation of Portsmouth, New Hampshire has provided the contractual support for the bulk of the implementation period.

By an overwhelming margin, panelists, reviewers and Departmental staff have given RMC good to very high marks for the quality of the work. Panelists have likewise been complimentary about Departmental staff support, though less complimentary about Departmental processes at the close out of panel work, particularly with respect to its duration and status information.

The most serious problems from the support staff perspective concerned the unfettered procedural discretion given the panels to determine their own processes and the overall panel budget constraints. The staff was uncomfortable with the need to tell panels they could not proceed with desired processes for lack of budget. *This problem can and should be addressed with more clarity at the outset about the range of procedural freedom and the realities of the budgets.*

There is, however, a broader set of concerns about the costs of the panels. The panels cost approximately $250,000 each or a total of about $1,000,000 for the four so far undertaken. If the panels could have undertaken some of the sound ideas that they had to confirm the program claims of clients with on-site reviews, the costs would have been yet higher. These costs do not, in general, exceed the range of reasonableness for what was undertaken. Given the constrained size of the total OERI research budget (approximately $100 million), however, the amount spent on these panels for the return received in program designations raises some real questions about spending priorities.

We earlier discussed the problems of constrained resources for educational knowledge building and the urgent needs to grow that base to support this program and important national needs. *Without that resource growth, this program will not likely withstand serious scrutiny on cost effectiveness grounds.*

7. WHAT CONTINUITY AND FREQUENCY OF REVIEWS HAVE BEEN CONTEMPLATED?

Oral interviewing with some Department staff and panel chairs touched on the question of panel frequency—how often panels should be convened. Panelists also volunteered their expectations about continued operation of their panels. It is clear that panel leaders with whom the question was discussed and the most involved staff prefer to proceed. The official Departmental position statement is that the system and its future are under review to be completed before the existing panels proceed. While the next section deals with broad options for the review system, we deal here with an assessment of the approach to continuity and frequency on the assumption that the system will continue to operate in some fashion.

Those most anxious and prepared to proceed are those associated with core topics of math and science curriculum and educational technology. Those topics are large, with considerable range of content to cover. Each of those panels undertook to limit the scope of their inquiry at the outset, to give their efforts focus and manageability (curriculum in the case of math and science, and systems rather than products in the case of educational technology). Such large and central topics can be treated regularly for an extended period of time within the criteria earlier suggested for topic selection—say, every one to two years. They are areas of more substantial research and activity, thus providing a larger pool of candidate programs to examine.

Other program topics may be appropriate for more periodic rather than regular reviews on the basis of factors for consideration, especially factors related to the scope, quality and rigor of the knowledge base related to the topic. Further, for emerging topics, it may be appropriate, given the knowledge base, to run a cycle aimed exclusively at promising designations, rather than seeking exemplary designations in areas with limited research and rigorous evaluations.

All of these considerations suggest the development of plans for continuity and frequency of topic panels that are closely tied to the linkage of the overall OERI program to the expert panel system. *The continuity and development plan would include two subcategories of continuing and periodic panels with frequencies of panel convening established for each. This plan would be updated every one to two years to reflect new entries and revised frequencies based on assessment of the knowledge base and new needs.*

V. STAFFING

V. STAFFING

In this section, we turn to the staffing and training of the panels. Some aspects of this topic have been covered in the prior section on processes and alluded to in other earlier sections. As is obvious from the description of the panel system and its processes, the character and quality of the panelists and reviewers are critical to the quality of the results achieved, since much depends on their knowledge and judgment.

1. WHO WAS SELECTED?

The statute called upon the Department to select groups of "appropriately qualified experts and practitioners" to staff the panels. The panels, with the help of support staff, then recruited reviewers to assist them.

Eliminating a quite modest attrition from the following count, 78 individuals were recruited for the four panels. Gender Equity had 29 members, while the other three panels ranged between 15 and 18 members. The panelists characterized themselves primarily as researchers, consultants and educational administrators rather then teachers, though many of them had prior teaching experience.

The first four panels together recruited 229 reviewers. The largest group was for the Math and Science panel (94 for math and 36 for science), followed by Safe, Disciplined and Drug Free Schools (60), and smaller groups for the remaining two panels. Of those who responded to the survey, one-third identified themselves as teachers, highly concentrated in math (57% of responding reviewers) and science (60% of responding reviewers). The Math and Science panel had deliberately set out to emphasize teachers among the reviewers. The other panels used reviewers with occupations in research and educational administration much more heavily.

The evaluation project staff and its Technical Working Group believe strongly that panels—particularly those dealing with curriculum—should include members with current teaching experience in the topics under review. We believe that the reviewers should be rigorously screened for their content knowledge and its application to the classroom. We agree with the Math and Science panel's decision to seek such personnel as at least one member of each review team. We recognize that expertise on crosscutting topics may be limited among teachers, particularly in new areas such as education technology. Nonetheless, the presence of the teaching perspective is always important at both the panel and reviewer level, and it will become more so as a given field matures.

2. HOW WERE THE PANELISTS AND REVIEWERS TRAINED?

The training was varied in approach, length and content among the four panels, and it became more sophisticated through time as experience was gained and more resources were applied. In general, the training was aimed at the reviewers, and involved the use of panelists as trainers with the help of the support staff. The content covered the purposes of the review system, the application of criteria and the use of scoring rubrics. As noted earlier, some panels used exercises to teach the review work to be done with simulated, or in one case, real applications. The duration of the training ran from one to three days, and varied in approach from face-to-face sessions to use of the Internet. These variations make it difficult to assess training intensity, as well as its impact on the consistency of program selections.

The scope of this assessment did not include a detailed review of each panel's training regimen, as some the earlier phases are wholly undocumented. We have, however, underscored the great importance of this training for all participants, in order to achieve the highest possible level of validity. The lack of specificity in criteria and the amount of judgment required increase the need for thorough training in the application of criteria and the use of scoring rubrics. Attention to more training should, of course, be accompanied by further work to make criteria more specific and scoring rubrics stronger, as at least one panel has already moved to do.

While training was in no sense inappropriate in later panels, we remain concerned about the intensity of some of the training conducted in the first round, and urge that, in the future, it err on the side of more rather than less intensity and duration. Further, as the inventory of programs considered in review cycles expands, the number of good training examples should increase, and these examples should be used.

VI. OPTIONS, CONCLUSIONS AND RECOMMENDATIONS

VI. OPTIONS, CONCLUSIONS AND RECOMMENDATIONS

Having examined in the foregoing sections the first round implementation of the expert panel system for identifying promising and exemplary programs and the standards they set, this assessment now turns to options, conclusions and recommendations that flow from the analysis.

It is easier to diagnose than to prescribe. We have approached this task from the outset by trying to describe a set of possible options that could be considered for the future of the system, in order to frame the issues and the conclusions and recommendations. We will use those options here with similar purposes in mind. Before laying out the four possible options, however, there are two other components to our analysis that should be summarized. These are a brief discussion of our review of processes in other Federal public agencies that might provide insights about this review system, and also a discussion of the level of detail at which the designations should be made.

1. WHAT CAN BE LEARNED FROM SIMILAR PROCESSES AT THE NATIONAL LEVEL?

Our original assignment contained the thoroughly sensible question about what we could learn from other Federal or national processes with similar goals and objectives. While many of the procedures and the general structure of the expert review system here considered were familiar in many kinds of review and assessment systems (e.g., grant competitions, research assessments), the initial scan revealed no other process that in its goal and objectives was very similar.

We asked Departmental staff who worked with the panels for suggestions about other organizations that used a review process to identify promising and exemplary practices. Suggested review programs included the National Reading Panel's work to assess the scientific bases for best instructional practices, and Project 2061 conducted by the American Association for the Advancement of Science (AAAS) on attaining science literacy, including mathematics. We explored these (and other) organizations' procedures for inputs to the unique aspects of one-time or ongoing identification of promising and exemplary practices. The methodologies used by the National Reading Panel and AAAS were most instructive for this evaluation. Brief descriptions of each of the organizations' review efforts are given below.

The task of the National Reading Panel was to conduct a meta-analysis of the research literature on best practices in reading instruction. The initial screening procedures used by the Panel were of particular interest. First, the Panel identified topics of interest and searched several databases for reports that were appropriate. At this level of review, the Panel specified several criteria the studies had to meet to be considered relevant, including the measurement of reading as an outcome. The resulting studies were then screened a second time for another set of criteria, relating to the rigor of the methodologies used, the experimental (or quasi-experimental) nature of the study design and the source of publication. The studies that remained after both rounds of screening were then subjected to analysis and to a standard set of coding procedures. The data sets from these studies were "converted" so they could be pooled for a larger group from which effects were estimated. Based on the results of this analysis, the Panel concluded that instruction with phonics was more effective than approaches that did not use phonics. Similar processes are now being initiated in a project known as the "Campbell Collaboration" that seeks in a number of areas, including education, to identify the state of the knowledge base derived from rigorous methodologies on a continuous basis.

Project 2061 has been a long-term initiative of the AAAS to provide State, local and national educators with the tools to redesign their curricula in science, including mathematics. Its output has been a family of books and monographs that provide guidance, concepts and benchmarks for reforming science education. While the work done by AAAS was undoubtedly most helpful in getting the Math and Science panel off to a good start on criteria and other matters, Project 2061 and the expert panel review system have different objectives.

From this review, we concluded that there do not appear to be any processes relevant to the U.S. Department of Education's attempt to identify promising and exemplary practices in a wide range of fields. The work in these other projects may supply particular panels with highly useful analyses or useful components to review procedures, but none appears to be an ideal template that could be simply transplanted in whole to this review process.

2. AT WHAT LEVEL SHOULD DESIGNATIONS BE MADE?

Internal discussions within this assessment project, including its Technical Working Group, dealt with an important choice made in connection with establishing the review system. The system was organized, consistent with statute, to produce designations of specific programs that can be implemented, presumably intact, at the school operational level. In commercial terms, the programs are designated at the retail level.

This current approach has the advantage of being a "product" that one can reasonably acquire complete and tailor to local circumstances. Thus, it minimizes the burden on local school districts to learn about key features that are important to selection and search for candidates. This approach has the disadvantages of falling short of providing a comprehensive review of the field and heightening the competitive environment of the process.

An alternative system approach would be to identify the critical and specific components of promising and exemplary programs and then identify a list of programs that possess most or all of those components. This alternative approach would minimize the disadvantages of current practice, but add burden to State and local school officials in making choices of the programs that they wish to adopt and implement.

We did not carry this alternative approach into our options, based on strong arguments within the Technical Working Group that local school officials and teachers needed the specificity of the first approach. We had no way of testing this hypothesis from the information we collected. *If this argument for the current approach is not considered compelling, the alternative approach could be tested and included in the continuation options described below.*

3. WHAT ARE THE BASIC OPTIONS FOR THE EXPERT PANEL REVIEW SYSTEM?

We have identified four theoretical options to examine. Except for the discontinuation option, these options are not mutually exclusive, but can be considered in combination over time. Each of the options is briefly described, and the major advantages and disadvantages are then provided. The conclusions that we draw about these options follow in a later subsection.

3.1 Option 1 would abandon the program entirely on the grounds that it produces higher costs than benefits given existing Departmental capacity and knowledge base, with inadequate resources to improve either.

The Department would advise Congress that the system was a mistake, ahead of its time, and should be withdrawn, at least until the knowledge base is sufficiently strengthened to support the program.

This option has the advantage of off-loading a very difficult and sometimes contentious activity in the context of highly limited resources and staff capacity to support the undertaking. The shock value might provide an unaccustomed "wake-up call" about the inadequacies of the resource and knowledge base.

This option has the disadvantage of leaving widely shared goals for the system without any response, a position likely to further reduce the reputation for competence and capacity of the Department and OERI. It does so without offering the Department or Congress any way out of a dilemma identified in the contextual background.

3.2 Option 2 would make a series of improvements in the current system without major structural change, and continue direct Federal operation of the system.

The Department would make a series of improvements discussed in prior sections within their administrative discretion. These improvements would improve the system short of substantial structural reform requiring legislative action and also maintain the system within direct Federal operation. Such actions would include:

- Strengthening the specificity of the criteria
- Strengthening the linkage of the panel topics with the research priorities in the OERI research agenda and an effective communication system
- Standardizing the review processes not directly tied to idiosyncratic requirements of topic criteria
- Streamlining post-panel review processes
- Careful specification of FOIA and FACA rules so their implications are well understood by all in advance
- Stronger communication links during and after panel reviews with applicants and other interested parties, including panel assessment of the quality of applications and probable improvements in criteria in scoring rubrics
- Announcing a strategy of panel future reviews (some panels would continue and others would not, and be replaced by new entrants)
- Strong screening in selection of panelists and more intense training of reviewers.

The advantages of this option would include the fact that it builds upon the learning from the first round of implementation without posing any significant legislative adjustments that might set off more changes than appropriate. This option would maintain the Department's control of the future evolution of the system. The Department might well see increases in OERI resources in the forthcoming appropriations for FY 2001 that could be devoted in part to strengthening the program. More experience could be gathered as a base for a yet stronger long-run program. While the improved system would not require new legislation, it would be a mistake not to explain improvement plans to Congress in order to assuage any concerns and build support for the changes.

The main disadvantage of this option is that it retains within the Federal government the direct operation of an underfunded system with a still underfunded knowledge building base that can be misinterpreted as a Federal effort to impose a view of appropriate educational content on state and local entities. This option could fail to produce the necessary resource commitments to make the improvements effective.

3.3 Option 3 would make the improvements to the system outlined in Option 2, but outsource the operation to one or more (likely the latter) external parties. These external groups would organize and operate the expert panels, subject to Federal rules and oversight.

This option would recognize external organizations to conduct the quality reviews and designation process pursuant to Federal regulations. This concept is akin to the Department's quality assurance program for postsecondary education institutions, which operates by recognizing the work of accrediting organizations. The Department would decide what topics would be covered and then, through a selection process, seek an appropriate organization to conduct the reviews. Unlike the accreditation process, the Department would bear the cost of the review. *While this approach might be within the scope of present legislation in general, the Department would no doubt want to make clear in legislation that the designations were the product of the panels and the external organizations. Thus, a more general statutory authorization would probably be wise.*

The advantages of this option include the diminution of the direct Federal administrative burden and some of the extra burden imposed by unique Federal procedural requirements. It would also have the advantage of putting some distance between the Federal government and the direct designation of instructional material. It might prove more flexible and expeditious in initiating new panels and in assembling an appropriate staff.

The disadvantages of this option lie in inflexibility and in diminished Departmental control of the developmental process of the system. While initiating new topics may be easier, stopping or sharply altering them is likely to be more difficult. With the rigor and specificity of the criteria still evolving, the issues of inter-topic consistency become more complex and difficult. The selection of competent organizations beyond the mature curricula topics (and related associations) may well be a challenge. While premature at this time, the further development of the Campbell Collaboration and/or the newly established Educational Quality Institute might well provide competent and appropriate organizations. The cost of running the system seems likely to be higher than the present system.

3.4 Option 4 would make all of the improvements outlined in Option 2, and additional structural changes, the latter of which would re quire legislation. It would provide authorization for direct or external operations.

This option would move away from an application-driven process to comprehensive assessments for exemplary program designations. It would recognize the separation of criteria for promising and exemplary programs, and authorize separate review cycles for the two categories. While retaining direct Federal operation for some or all of the topic panels, the legislation would make clear that the technical quality of the judgme nts about the designated programs would be the responsibility of the panels, subject to Secretarial oversight of the rigorous application of appropriate criteria. It would explicitly anticipate a transition period to the restructured system. It would recognize that organizations may well emerge to whom the review and designations of exemplary programs on a comprehensive basis could be comfortably delegated or out-sourced, and authorize such delegations. Efforts like the emerging Campbell Collaboration, if successful, might resemble such an organization.

The advantage of such an approach would be to clarify certain major issues in the existing system, and provide a developmental approach consistent with contextual realities. It is an approach thoroughly consistent with the underlying goals of the review system, leavened with a realistic understanding of where the knowledge base is.

The main disadvantage of this approach is that a transitional period will be required and recognized to be part of the development of the review system (not usually done in federal programs). The Department will thus need to be committed for the longer term to making the review system work, and to the provision of resources necessary not only for the review system, but for the underlying development of the knowledge base needed to make the system work. These conditions are a tall order for an under-resourced organization.

4. WHAT CONCLUSIONS AND RECOMMENDATIONS SHOULD BE DRAWN FROM THE OPTIONS AND THE UNDERLYING ANALYSIS?

In commencing an answer to this bottom-line question, we want to reiterate in a more general judgmental fashion a point made throughout this analysis about resources. *The success of the expert review panel system for the designation of promising and exemplary programs is highly dependent on the resources made available, not only for the operation of the system itself, but for the aggressive development of more rigorous and expansive knowledge base from which designations can come. Without such resources, the expert panel process will not likely be worth the funds it takes to operate it.*

With that understanding, it is now appropriate to assess the options raised in the previous subsection. Option 1 is unattractive in any circumstance other than the absence of adequate resources to make the system effective. If the circumstances of constraint were to continue, then dismantling the system should be considered and undertaken. The widely shared goals that underlie the system appear as enduring as any in education knowledge building, though their pursuit will be greatly enhanced with generous doses of candor and realism.

The content of the improvements identified in Option 2 should be considered, then either accepted, amended or rejected, and then implemented. They represent, in our judgment, a basic threshold of improvements that can and should be made on the basis of experience to date with implementation of the system. As noted, they do not appear to require legislation for implementation, but as a matter of prudence they should be discussed with appropriate Congressional committees.

The outsourcing involved in Option 3 was initially attractive to the project evaluation staff as a possible escape from some the actual and potential bureaucratic issues that seemed to be arising in the implementation of the system. On closer inspection and reflection, *this option as an across-the-board approach is likely to raise as many problems as it solves in the short run.* Further, its adoption is likely to deepen an unfavorable impression of Departmental capacity to the detriment of the program in whatever setting it exists. It is the case, however, that the selective use of external organizations for exemplary reviews may well become appropriate if organizations employing the concepts underlying the Campbell Collaboration evolve into effective entities. These concepts include continuously updated, rigorous and comprehensive reviews of the worldwide knowledge base in a topic area, using consistent criteria. Since it is unlikely that the Campbell Collaborations would be able or want to include promising programs in their reviews, the identification of promising programs would continue as part of the Departmental review system, probably on an application basis, at least until some acceptable external alternative develops.

Option 4 contains a list of attractive long-term ideas that, in all but one case, would require a transitional and developmental period. In combination with Option 2, they provide an attractive package of short-term and longer-range improvements to the program. These improvements could be implemented while serious progress is being made on the strengthening of the research and evaluation program and the knowledge base.

5. RECOMMENDATION

We would adopt Option 4 and initiate a legislative change right away to ensure the technical judgments in the selection of designated programs are those of the panels and not the Secretary. We would begin in the near future to explore and experiment with the separation of the promising and exemplary categories with respect to criteria and review cycles. A revised scheme could be implemented in two to three years. We would set a goal for comprehensive exemplary reviews in a period of three to five years. This, the most difficult component of Option 4, reflects the spirit of the statute, and it is the component that most nearly depends upon significant increases in research and evaluation activity to permit and sustain it.

6. WHAT STRATEGY SHOULD BE ADOPTED FOR PROCEEDING?

The conclusions and recommendation discussed above lay out a sequence of events that forms the basis for a strategy for proceeding. The Department would consider, decide and act upon the short-term improvements identified in the analysis and summarized in Option 2. This would probably include a discussion of general plans with the authorizing committees of Congress, and it would cover components selected from Option 4. This conversation would probably occur in the context of the expected OERI reauthorization hearings and discussions.

The Department should decide and announce over the next four to six months what panels it expects to continue. Educational Technology and Math and Science would be good candidates to consider.

The Department should select the components of Option 4 that it would like to adopt and initiate the staff work to outline the legislative amendments and the transitional plans and schedules appropriate to move those components from concept to reality.