

12/16/2005

Dr. Russell Pimmel  
Division of Undergraduate Education  
National Science Foundation  
Arlington, Virginia 22230

Dear Dr. Pimmel,

Enclosed is our report entitled "Reviewer Database and Evaluation Process." Preliminary work was completed in Worcester, Massachusetts prior to our arrival in Washington, D.C. It was written at the National Science Foundation during the period October 22, 2005 through December 16, 2005. This report is also being submitted to Professors El-Korchi and Servatius for evaluation. Upon faculty review, this report will be catalogued in Gordon Library at Worcester Polytechnic Institute. We appreciate the efforts and time you have devoted to us.

Sincerely,

Jessica Clark  
Brian Guerette  
Victoria Ruhl

# Reviewer Database and Evaluation Process

An Interactive Qualifying Project Report Submitted To:  
Professor Brigitte Servatius  
Professor Tahar El-Korchi  
WORCESTER POLYTECHNIC INSTITUTE  
Washington, D.C. Project Center



By:

---

Jessica Clark

---

Brian Guerette

---

Victoria Ruhl

In Cooperation With:  
Russell L. Pimmel  
National Science Foundation, Arlington, VA



December 16, 2005

Approved By:

---

Professor Brigitte Servatius

---

Professor Tahar El-Korchi

## **Abstract**

Effective reviewer selection at the National Science Foundation can make the review process, the funding decision, and the communication with Principal Investigators more successful. The purpose of our project was to create and validate a process for evaluating panel reviewers along with a searchable and updatable reviewer database. We conducted interviews and surveys of Program Officers to gather information for the evaluation tool and the database, analyzed the information to determine how to create them, and performed validation testing.

## Authorship Page

Jessica Clark	Conducted PO Interviews Developed the Evaluation Process Designed the Database Survey Designed the Database Created the Database in MS Access Interviewed POs about Database usability Major contributions to: Abstract, Introduction, Background, Literature Review, Database Creation, and Discussion
Brian Guerette	Conducted PO Interviews Developed the Evaluation Process Designed the Database Survey Conducted the Database Survey Conducted Evaluation Process Validation Testing Major contributions to: Abstract, Executive Summary, Introduction, Background, Rubric Creation, Appendices
Victoria Ruhl	Conducted PO Interviews Analyzed PO Interviews Developed the Evaluation Process Designed the Database Survey Conducted the Database Survey Made Presentations Major contributions to: Abstract, Introduction, Background, Rubric Creation, Discussion, Recommendations, and Conclusion

## Acknowledgements

Prof. Tahar El-Korchi  
Prof. Brigitte Servatius  
Dr. Russell Pimmel  
Ms. Antoinette T. Allen  
Dr. Barbara Anderegg  
Dr. Myles Boylan  
Dr. Mark Burge  
Dr. Susan Burkett  
Ms. Joyce A. Craig  
Dr. John Haddock  
Dr. R. Corby Hovis  
Dr. David McArthur  
Dr. Duncan McBride  
Dr. Kathleen Parsons  
Ms. LaVerne Paige  
Dr. Joan Prival  
Dr. Nancy Pruitt  
Dr. Herbert Richtol  
Dr. Curtis Sears  
Dr. Jeanne Small  
Ms Melissa Squillaro  
Dr. Keith Sverdrup  
Dr. Harry Ungar  
Dr. Bevlee Watford  
Dr. Terry Woodin  
Dr. Lee Zia

And all the other DUE staff members who have helped us along the way.

## Table of Contents

Executive Summary .....	1
Introduction.....	3
Peer Review .....	6
Peer Review Process.....	6
Peer Review in the DUE.....	7
Panel Selection.....	8
Review Criteria.....	9
Reviewer Responsibilities.....	10
Previous Studies.....	11
Evaluation Process .....	14
Interviews.....	14
Interview Breakdown.....	14
Evaluation Tool Design .....	15
Validation.....	17
Evaluation Tool Results and Validation .....	18
Program Officer Interviews .....	18
Rubric Creation.....	19
Evaluation Tool Validation.....	21
Database.....	24
Database Design.....	24
Database Validation .....	25
Database Creation and Description.....	25
Database Survey.....	25
Database Creation .....	26
Database Modifications and PO Comments .....	28
Societal Implications of Evaluation Tool and Database .....	30
Scoring Reviewers .....	30
Outside Evaluators .....	31
Informing the Reviewers.....	32
Benefits to Program Officers .....	33
Benefits to Principal Investigators .....	34
Benefits to Reviewers .....	35
Benefits to Society .....	36
Recommendations.....	37
Recommendations for the Reviewer Database .....	37
Recommendations for the Evaluation Tool .....	39
Recommendations for Future Projects.....	39
Closing Statements.....	40
References.....	42
Appendix A: Sponsor Description – About the NSF.....	45
DUE and CCLI Statistics .....	45
Appendix B: Interview Questions.....	47
Appendix C: Interviews: Review Qualities .....	48

Appendix D: Interviews: Review Problem Areas.....	50
Appendix E: Interviews: Suggested Evaluations.....	53
Appendix F: Interview Analysis.....	55
Appendix G: First Draft Rubrics.....	56
Rubric #1.....	56
Rubric #2.....	57
Rubric #3.....	59
Appendix H: Six Question Rubric.....	60
Appendix I: Two Question Rubric.....	61
Appendix J: Final Rubric.....	62
Appendix K: Rubric Validation Survey.....	63
Appendix L: Rubric Validation Scores.....	64
Appendix M: Database Survey.....	68
Appendix N: Database Survey Analysis.....	69
Appendix O: Database Screen Shots.....	70
Switchboard:.....	70
Add Reviewer:.....	70
Search.....	71
Search Results.....	71
Details.....	72
Review History.....	72
Update Reviewer Information.....	73
Add Score.....	73
Appendix P: Database Instructions and Test.....	74
Appendix Q: Project Description.....	76
Glossary.....	78

## Executive Summary

The goal of this project was to develop and validate a tool to evaluate panel reviewers along with a searchable and updatable reviewer database for use by the National Science Foundation (NSF) Division of Undergraduate Education (DUE) Program Officers (POs). These tools will assist the DUE POs in the selection of the most helpful reviewers by providing an objective measure of reviewer quality and storing that score along with other reviewer information in a database. Selecting the most qualified reviewers is a significant portion of a PO's job. Reviewer selection affects the funding decision, communication with the Principle Investigator (PI), and the review process in general. A process to evaluate written reviews and a searchable database will make the reviewer selection more effective.

We gathered information from seventeen [DUE POs](#) through interviews, analyzed the interview comments, and were able to identify seven criteria of an effective review. The seven criteria were length, grammar, addressing strengths and weaknesses, constructive criticism for the [PI](#), giving justification for the rating, addressing the program solicitation, and appropriateness of reviewer comments. We combined these criteria into three categories each having a three-point scale.

Validation testing assessed the inter-rater reliability and validity of the evaluation rubric. A total of ten sample reviews were strategically selected to ensure reviews of varying quality. Ten [DUE POs](#) each scored five of the ten



sample reviews and completed a survey. A majority of the participating [POs](#) indicated that the evaluation tool was easy to use and half indicated that they would use it on a regular basis.

A survey allowed us to compile the functions and reviewer information fields that the [POs](#) wanted us to include in the database. We received a total of nine responses, and, at the [POs](#)' requests, decided to include thirty information types and four database functions. We incorporated the evaluation rubric into the database so that [POs](#) could store reviewer evaluation records. To validate the database, we presented it to two [POs](#) and had them comment on the ease of use and suggest improvements. We made modifications to the database according to the [POs](#)' comments and developed the final version of the reviewer database.

We designed the evaluation process and database using input from [POs](#) in every [DUE](#) program. They provide general information that is a useful supplementation to current reviewer selection methods. We recommend that the [DUE](#) put the tools in use and make periodic refinements and updates to maximize their effectiveness.

## Introduction

The National Science Foundation ([NSF](#)) is the funding source for more than twenty percent of all federally funded research projects and the major funding source for all science, engineering, and mathematics research projects completed in research facilities across the United States [1]. The current budget of the [NSF](#) is \$5.5 billion a year [1].

The Division of Undergraduate Education ([DUE](#)) is a subdivision of the Education and Human Resources directorate of the [NSF](#). The mission of the [DUE](#) is to promote excellence in undergraduate education in science, technology, engineering, and mathematics ([STEM](#)) for all students [2]. In 2003, the [DUE](#) received \$172.55 million in funding for undergraduate education and awarded it competitively to various organizations through grant proposals [3]. (See [Appendix A](#) for more NSF background).

Since only 15-20 percent of all proposals can be funded, careful proposal selection is vital [4]. The proposals are evaluated using a merit-review process, carried out by a panel of volunteer peer reviewers from learning institutions, private industries, and professional organizations across the country [5]. Program Officers ([POs](#)) are employed by the [NSF](#) to choose panel members and award or decline project proposals. Their decisions are heavily influenced by the recommendations of the peer reviewers [5]. Ensuring the selection of capable reviewers is the most cost effective approach for use of [DUE](#) funding.

Peer review in the [DUE](#) has two important purposes: to help the [DUE](#) select the best proposed projects to fund, and to provide Principal Investigators ([PIs](#)) with constructive criticism on their proposals [\[6\]](#). Therefore, it is necessary to select the reviewers who can most effectively realize these tasks. Some proposals may require reviewers with specific technical knowledge while others may require work experience such as research or assessment. [\[12\]](#).

Currently, the Program Officers all have individual methods for choosing peer reviewers. There are a number of informal approaches that [POs](#) use. [POs](#) may select a peer reviewer based on personal recollections or use suggestions made by [PIs](#). They may also search among authors who publish in scientific and engineering journals, online science and engineering abstracts, or professors of relevant subjects at universities [\[6\]](#). Roughly half of the [POs](#) are rotators that are at the [NSF](#) for 1-2 years. Many of the rotators start out having no experience in selecting reviewers [\[12\]](#). Having a centralized database with information about the quality of reviewers will be a useful resource for the [POs](#).

Utilizing the advice of [POs](#), the goal of this project is to automate and simplify the reviewer selection process. The project will have two components. The first is to establish a scoring rubric and standard evaluation process so [POs](#) can evaluate individual reviewers each time they serve on a panel. The rubric should have three or four areas of ratings that will allow the [POs](#) to score the reviewers. The second portion of the project is to create a searchable database that can be easily updated with new reviewers and new evaluations. This

database will identify and rank the potential reviewers to be selected for each particular panel.

## Peer Review

This section describes background on peer review and the process that the [NSF](#) uses to select proposals for funding. The proposal review process is successful and well established. The goal of this project is to create a tool that will assist the [DUE](#) in selecting the best-qualified reviewers to match the submitted proposals.

### Peer Review Process

Peer review, as it applies to scientific research, is defined as having independent experts in the field judge a research article or grant proposal [\[20\]](#). There are two purposes for which peer review is standard: to assist a journal editor in making a decision to accept or decline a submitted article, and to advise a decision maker for an institution which distributes money through grants as to which proposals should receive funding.

Journal articles are reviewed after the research is completed. People not linked to the research review the articles to verify that the results are valid and therefore will be respected by others in the field. Peer review takes place in journals because editors need to rely on reviewers with specific expertise to evaluate highly technical material. The peer reviewers function as assistants, making a recommendation about publication [\[23\]](#).

Grant application review requires the reviewers to be able to predict the success of a proposal by evaluating the research plan. Peer reviewers provide additional technical expertise to broaden the range of areas in which a decision

maker can select research to fund. Many funding organizations must be held accountable to society because they are federally funded. Peer review helps to ensure that the best projects are funded [21].

This project focuses on peer review for grant proposals instead of journal article submission; however, the essential qualities that a reviewer must possess are similar for either type of review. Therefore, previous studies done about peer review for journal review can provide insights about peer review for grant proposals.

## Peer Review in the DUE

To understand the difference between poor and excellent reviewers, one must first understand the basics of peer review as it pertains to the [DUE](#). The [NSF](#) tries to accept or decline all proposals within six months of receiving them [8]. The panel members first individually review each proposal, and then they come together as a panel to discuss the proposals as a group.

Prior to sending the proposals to the reviewers, the [POs](#) scan them to ensure that each proposal adequately fulfills all of the mandatory requirements such as appropriate format and correct sections (such as cover page and project summary). This ensures that no reviewer will spend time reading a proposal that is ineligible for funding. Next, the reviewers receive the set of 10-15 proposals via “Fastlane,” the [NSF](#)’s electronic submission and review website [12].

Reviewers are ideally given three weekends to rate each proposal on a scale from “poor” to “excellent,” and also record their comments. These comments should reflect the rating given to the proposal, and should also list strengths,

weaknesses, and ideas to better develop the proposal. Finally, the reviewers meet for a panel discussion and make recommendations to pass on to the [PO](#). The written comments of the individual and panel reviews are sent back to the [PIs](#) along with the [PO](#)'s funding decision [\[6\]](#).

## Panel Selection

The [DUE](#) uses panel review, in which a group of people give both individual and collaborative reviews for a proposal [\[11\]](#). At least 3 reviewers “outside [NSF](#) who are experts in the particular field” are required to review each proposal [\[8\]](#). Panels are comprised of reviewers who are all reviewing the same set of proposals. Typically, each panel consists of 5-7 reviewers who will review 12-15 proposals. The reviewers can be found from a variety of sources. Typical possibilities include authors of works cited in the proposal, members of professional societies, published authors from relevant journals, online listings of abstracts of recent research, the recommendations of the [PI](#), and recommendations of other reviewers [\[5\]](#). The [DUE](#) tries to compose each panel of 1/3 new reviewers, 1/3 past reviewers, and 1/3 current or former [PIs](#) [\[35\]](#).

There are strict regulations barring the selection of certain reviewers. Specifically, the review candidate must decline an invitation to review if the candidate, relatives of the candidate, or the institution that the candidate is affiliated with has a financial interest in the outcome of the proposal. Also, the candidate is not eligible if he or she has a personal relationship with the [PI](#) or Co-[PI](#) of the proposed grant. Personal relationships that would recuse a reviewer include former employers or employees, collaborative colleagues, and relatives

[10]. For panel reviews, the panelist is not allowed to listen to or participate in discussions of proposals from [PIs](#) with whom the panelist has a conflict of interest [6].

Reviewers are not allowed to identify themselves in the reviews, nor are they allowed to discuss the specific proposals with others not on the panel in order to maintain confidentiality and privacy [6].

## Review Criteria

Uniformly across all directorates, the two key criteria to be satisfied by each proposal are “intellectual merit,” and “broader impacts,” each of which consist of several components. For intellectual merit, reviewers judge whether or not the research is original, whether the proposed research advances “knowledge and understanding within its own field or across different fields,” [3] and whether there is a clearly structured research plan. When judging the potential broader impacts of a proposal research, reviewers should evaluate how the research will help the general population, if the research will increase the involvement of underrepresented groups in the sciences, and if the results of the research will be understandable and readily available to the general population [3].

Proposal criteria specific to the [CCLI](#) include: realistic expectations for research outcome, applications to [STEM](#) education, and a well-designed assessment strategy. Projects that will have quantifiable results, continue past research, or focus on communication and cooperation among members of diverse scholarly fields are especially preferred [6].



Other criteria that focus less on the technical outcome of the research and more on the potential for success include the probability of help from the researcher's institution, the outcome of past research from the Principal Investigator ([PI](#)), and the geographic location of the [PI](#), to "avoid concentration of such research and education" [\[13\]](#).

## Reviewer Responsibilities

A panel reviewer must give the proposal a clearly defined rating, and the rating should reflect the written comments. For instance, if the reviewer's comments are a list of concerns about the proposal, the rating should be "fair" or "poor." The comments should describe the strengths and weaknesses of the proposal and the research ideas. The reviewer should not simply describe the proposal, as the [PO](#) is already familiar with its content. The reviewer should also include suggestions for bettering the proposal and research ideas. Finally, the reviewer should consider "the results of any prior support," meaning that he or she should take the [PI](#)'s past successes or failures at the [NSF](#) into account [\[14\]](#).

## Previous Studies

“Strengthening Peer Review in Federal Agencies That Support Education Research” is the result of a workshop sponsored by the National Research Council. The report identifies the purpose of peer review for education research, difficulties with peer review, and ways to improve peer review in general. This report concentrates on peer review for grant proposals [22].

The concluded purposes of peer review from this report are “the identification and support of high-quality education research and the professional development of the field” [22]. Peer review in federal agencies is used to determine where federal tax dollars go, so the process and results must be satisfactory to the public. It is also a tool used to appropriately fund new research, which implies that there is a need for political figures to be intermediaries between the organizations using peer review and the public [22]. The two separate and possibly conflicting goals of peer review can create tensions and difficulties.

There have been recommendations made in effort to ameliorate some of these difficulties. One such recommendation is that federal organizations need to be organized in such a way that the peer review system is well supported. This report also directly states “Additional logistical supports could include well-managed databases” [22]. Databases can help in selecting reviewers that not only have deep knowledge of the field but also a range of different experiences. Keeping track of reviewers and their attributes can explain how well the peer

review system is working [22]. A goal of our project is to create an easily searchable and updatable database of reviewer characteristics. This will help the [DUE](#) to satisfy the recommendation of this paper.

Another recommendation from this panel is that organizations focused on funding educational research should clearly define the results they want to come out of the peer review process. These organizations should also have a routine method of evaluating the peer review process [22]. One such evaluation method could be giving each reviewer a score after a panel session is completed.

There has been some research done into evaluation tools for peer reviewers. One project sought to identify qualities of a helpful journal review, and then condense those qualities into several questions. The product that came out of this research was a seven-question rubric, where each question was graded on a scale of 1-5, called the “Review Quality Instrument.” The questions on the Review Quality Instrument are:

1. Did the reviewer discuss the importance of the research question?
2. Did the reviewer discuss the originality of the paper?
3. Did the reviewer clearly identify the strengths and weaknesses of the method (study design, data collection, and data analysis)?
4. Did the reviewer make specific useful comments on the writing, organization, tables, and figures of the manuscript?
5. Were the reviewer’s comments constructive?
6. Did the reviewer supply appropriate evidence using examples from the paper to substantiate their comments?
7. Did the reviewer comment on the author’s interpretation of the results? [21]

While our project deals with grant proposals instead of journal articles, we hope to accomplish the same goal as the writers of this paper. Therefore, our

evaluation tool should answer similar questions. A stated limitation of the Review Quality Instrument is that “although it can measure the quality of the comments the reviewer has made, it cannot assess the accuracy of those comments” [21]. This problem will be addressed by the fact that the [PO](#)'s select multiple reviewers who are experts in the area, a task that will be made easier by the creation of a database.

## Evaluation Process

One goal of this project was to create an evaluation process that Program Officers will use to evaluate a review. This section describes the methods in reaching these goals, how we conducted the procedures, and the resulting evaluation tool. We used Microsoft software for all procedures undertaken because the [NSF](#) uses Microsoft, and it was readily available to us.

### Interviews

The initial research phase consists of interviewing the [POs](#). Interviews (rather than surveys) allow a better response rate and a chance to clarify any confusion about the interview questions. The interviews are semi-formal [\[15\]](#). Since there are about twenty interviews to conduct, the questions can be somewhat open-ended in order to foster discussion or debate. The major research areas about which each [PO](#) can provide information are: the current reviewer selection process, good qualities of a written review, and problem areas with reviews (See [Appendix B](#) for Interview Questions).

### Interview Breakdown

To analyze our data we first clarify any questions on which there was confusion. We check for missing data, incorrect recording, and inconsistent data. If there are issues within any of the interviews, we try to resolve these issues by going back to the [PO](#) and re-asking or re-phrasing the question. Since the interviews are not formal, there is also a possibility that the [PO](#) will later remember something and contact us with more information [\[15\]](#).

To analyze the interviews, we use our written notes to create a coding system appropriate for our information. This analysis will allow us to process qualitative information scientifically [17].

Our coding system for the scoring rubric interviews consists of three MS Excel spreadsheets: good review qualities, review problem areas, and suggested evaluation areas. Each column represents a different [PO](#) and each row contains answers for the different answers in each interview area. In the spreadsheets the answers are abbreviated to certain codes, and a coding key is updated as answers are entered into excel. Down the columns, boxes are checked off according to the answers given by each [PO](#) in the interviews.

## **Evaluation Tool Design**

We use a scoring rubric as our evaluation tool. The purpose of our evaluation rubric is to quantify a [PO](#)'s opinions, attitudes, and feelings about a reviewer by having them answer a question or series of questions. A general type of question useful for our purposes is called the Likert-scale [29]. This type of question typically has a series of answers each expressing a category numerically labeled along a range. For example, a 1 may represent "very poor" and a 5 "very good" with 2, 3, and 4 representing specified categories within this range. The Likert-scale is commonly used and widely accepted in practice [29].

There are various ways to create a range that will not hinder the questions' effectiveness. One issue is the number of possible answer options in the question scale. Too few options can make it difficult to identify smaller differences in quality. Larger scales provide greater resolution at the cost of

inter-rater reliability [31]. The 1-5 scale is widely accepted since it provides enough resolution to obtain detailed information yet does not provide so many options that a lack of inter-rater reliability obscures the data [29]. Inter-rater reliability is the degree to which questions are answered consistently among different people. However, there is another problem that occurs with odd numbered scales. Evaluators generally show a reluctance to form an opinion and have preference towards neutral options, such as a 3 representing “average” in a 1-5 scale. An even scale confronts this issue by eliminating any neutral option, forcing the rater to express an opinion one way or the other [29]. However, this may be detrimental as it may force the rater to lean towards one side, when they really have no opinion. Depending on the question, it may be appropriate to add an additional choice of “no opinion.” Designing the questions appropriately depends on variables such as intended audience and desired information. The rubric is best refined through pilot testing.

The rubric structure is also an important factor that determines inter-rater reliability and resolution of information it provides. We are examining three different types of scoring rubrics used to judge writing: holistic, analytic, and combined. A holistic scoring rubric asks someone to give an opinion in a single score that reflects all aspects of the written piece. An analytic scoring rubric breaks the aspects of the writing into different categories, and the scorer judges in each category. This type of rubric has the highest inter-rater reliability and yields the most information, but is the most time consuming to use [30]. The combined type of scoring rubric weights the score in each category in an analytic

scoring rubric and combines them into one final score. This method generally has the lowest inter-rater reliability [30].

To create the scoring rubric, we compile all of the [POs](#)' answers to our interview questions, and determine which areas are the most important. For the initial design, each of us creates our own scoring rubric. We test each of these scoring rubrics on several reviews, taking notes on the ease of use and the functionality of each question on the rubrics. When we have all tested the rubrics, we compare the evaluation scores and notes. This comparison will help us to see what type of rubric is the easiest to use and most useful.

For the purposes of our project, it may be useful to create several versions of reviewer evaluation rubrics for pilot testing. Pilot testing can be done by having current [POs](#) evaluate sample reviews. The questions and rubric structure can be changed as necessary, and tested again. The final scoring rubric should consist of 3-5 areas that the [POs](#) feel are most important in evaluating a review.

## **Validation**

After creating the scoring rubric, we conduct a test run with the Program Officers to validate it. The test run consists of handing out 5-10 reviews and the evaluation tool to the program officers. The [POs](#) have a week to evaluate the reviews. Included with the reviews and evaluation is a comments sheet, so the [POs](#) can tell us what they like about the evaluation tool and what they think should be changed. We assess the inter-reliability and validity (the degree which we accurately measured what we wanted to).



## Evaluation Tool Results and Validation

### **Program Officer Interviews**

We conducted a total of seventeen Program Officer interviews, each lasting between twenty minutes and one hour. There are a total of twenty-three [POs](#) within the [DUE](#); however, they were not all were present at the NSF or available for interview. During these interviews, we took notes in bulleted form and immediately transcribed them to a MS Word document. We separated the notes for each interview into three general categories: review qualities, problem areas, and suggested evaluation areas. We created three MS Excel spreadsheets to group the comments made for each category so that we could make comparisons (See Appendices [C](#), [D](#), and [E](#) for the content analysis charts of the [PO](#) interviews).

The most popular answer for the “problem areas” was, “there was not enough detail in the review” (14 responses). For the “qualities” category, the most popular answer was, “the reviewer gives the [PI](#) constructive criticism” (15 responses). The “quality of writing in the review” was the most suggested evaluation area (14 responses).

Other areas such as “focusing on only one aspect of the proposal” (problem areas), “addressing the program solicitation” (good qualities), and “ability to understand the proposal” (suggested evaluation areas) were less commonly stated. Although these were not common responses, we still considered including them in the scoring rubric.

Next, we combined the different answer categories into general areas, and calculated the sum total of responses for each one. The areas with the most responses were “Detailed Justification” (42 responses), “Style” (43 responses), and “Solicitation/ Broader Impacts/ Intellectual Merit/ Strengths and Weaknesses” (45 responses). Other common response areas were “Proposal or Solicitation Understanding” (26 responses), and “Constructive Criticism” (32 responses). Breaking down the content analysis charts indicated that we should include seven aspects in our final evaluation tool: length, grammar, constructive criticism for the [PI](#), addressing strengths and weaknesses, giving justification for the rating, addressing the program solicitation, and addressing broader impacts and intellectual merits. We did not specifically include “Proposal or Solicitation Understanding” because satisfying all the other criteria implies that the reviewer understood them (See [Appendix F](#) for Interview Analysis).

## **Rubric Creation**

Initially, we each created our own evaluation rubric incorporating the most important aspects from the content analysis into three questions. We each tested all three evaluation rubrics on twelve reviews, informally compared results, and took note of difficulties we encountered using the rubrics. We also informally assessed the inter-rater reliability of our scores (See [Appendix G](#) for the initial rubrics). The results of this exercise indicated that questions having answers with a high degree of detail or overlapping questions are confusing and take too long to complete.

Utilizing the insights of this experience, we created another rubric, consisting of six short questions. The questions were: length, grammar, strengths and weaknesses, program solicitation, rating justification, and constructive criticism. Again, we each completed this rubric to test for difficulties and inter-rater reliability of scores (See [Appendix H](#) for rubric). The six short questions were much easier to read through and took a shorter time to score a review than our initial rubrics.

Next we combined related criteria from the six-question rubric into two questions and eliminated one criterion. The two questions were “Style” and “Content”. “Style” combined grammar and length and “Content” combined rating justification, strengths and weaknesses, and program solicitation. We eliminated constructive criticism from the evaluation because we found that it was too similar to addressing weaknesses (See [Appendix I](#) for rubric). We showed the two-question rubric to two [POs](#) and asked their opinions on the questions. Both agreed that constructive criticism should be included in the rubric and that there should be three questions instead of two.

The final version of our rubric is three questions. The first question, called “Style”, includes length, grammar, and the appropriateness of reviewer comments. The second question, “Rating Justification,” involves justifying the rating given to a proposal and addressing strengths and weaknesses. The third question, “Program Solicitation,” includes addressing the program solicitation and giving constructive criticism for the [PI](#) (See [Appendix J](#) for final rubric). After creating the final rubric we validated it by testing it with the [POs](#). We chose a

three-point scale because we found that more possibilities caused not enough distinction between the value descriptions.

## **Evaluation Tool Validation**

We conducted a validation test of the final rubric to assess its inter-rater reliability and validity. We selected ten review samples from the [CCLI](#) program. These reviews came from a variety of different reviewers and individual proposals, and were chosen to have varying style and quality. Ten [DUE POs](#) volunteered for the validation testing. They were each given a unique set of 5 reviews along with our scoring rubric, an excel spreadsheet to record scores, and a 3-question survey via email. Our validation test results consist of 5 sets of scores for each of the ten reviews, along with PO comments about the rubric (See [Appendix L](#) for Evaluation Rubric Scores and Comments).

There were a number of factors limiting our ability to perform tests of inter-rater reliability. The length of time required to score reviews using our rubric and the small number of [POs](#) who volunteered limited our ability to obtain a large number of scores. There were also several factors that potentially influenced [POs'](#) scores. Review quality tends to vary between disciplines, so [POs](#) from different disciplines would view “Acceptable” reviews differently. As one Program Officer expressed, “social science panels are even more helpful and complete in their comments [than engineers] so rating panelists would be a very discipline specific activity” [\[32\]](#). Because all of our reviews came from engineering panels, [POs](#) from non-engineering disciplines could have scored them differently than engineers. Other potential confounding factors included differences in the sets of

reviews that they received and tendency to go with an initial feeling about the review.

The most appropriate measure of inter-reliability for a small dataset such as ours is the average deviation. In this analysis we use mean average deviation to determine the error. The mean average deviation is the average absolute difference between each value and the mean of the values [36]. If  $x_i$  represents each score,  $\bar{x}$  represents the mean of the score, and  $n$  represents the number of scores, the mean average deviation,  $\mu$ , can be calculated by using the following formula:

$$\mu = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

The mean average deviation values for style, justification of rating, and program solicitation for all of the reviews collectively are .27, .39, and .34 respectively. These numbers represent the degree of error. Converting to percent error ( $\lambda$ ) yields the following formula:

$$\lambda = \frac{\mu}{n} * 100$$

The percent errors for style, justification of rating, and program solicitation are, respectively, 11.7%, 22.7%, and 20.6%. A larger dataset would allow us to use standard deviation, which is commonly accepted as a more accurate measure. A better test of the rubric's success would be for the POs to use it over time, and compare ability to select reviewers based on the scores provided from

other POs. This test can only be performed over time (for at least several months).

The intent of the survey was to determine if the rubric would be used on a regular basis, the simplicity and accuracy of how it recorded a [POs](#) opinion of the reviews, and to collect suggestions for rubric improvement. Eight in ten [POs](#) responded that the rubric was easy to use, but half responded they would not regularly use the rubric because it took too long to complete. The amount of time needed to rate a reviewer would be significantly reduced during actual use because the [POs](#) would become accustomed to the rubric. Of all ten [POs](#), one indicated that in some instances the score did not reflect their opinion of the reviews.

People gave us suggestions to improve the rubric. Many of the suggestions covered topics we had already considered. For example, one individual preferred a 1-5 scale. Through our pilot testing we dismissed a 1-4 scale in favor of 1-3 because the criteria for the answers in larger scales were too difficult to define distinctly and simply. For our purposes, a 1-5 scale would have been far too “fine grained.” Other suggestions included to make more descriptive headers for each question, to eliminate the “OR,” and “AND” operators, or to include a question with a non-calculated reviewer rating. Analyzing the responses we received, we would also suggest making minor modifications to the criteria for specific programs. For example, since more extensive elaboration is expected for some programs, each program could create a standard length for each rating on the scale.

## Database

The other goal of this project was to create a searchable and updatable database that Program Officers will use to store reviewer information and scores from the evaluation tool. This section describes the methods in reaching these goals, how we conducted the procedures, and the resulting product.

### Database Design

The key to our searchable database is that a search returns all eligible reviewers who satisfy the user-specified requirements. The eligible reviewers are ranked according to their evaluation scores.

There are several key steps to database creation. The first is to decide the function of the database, which is to search for possible reviewers. The second step is to figure out what information we want to enter into the database, and to decide the tables used to store that information. Next we have to identify primary keys that uniquely classify the records of reviewers, such as their names. The [POs](#) use the primary keys to search for certain reviewers.

Currently, Program Officers keep their own records or databases of reviewers. Often, one [PO](#) will make an excel spreadsheet of all reviewers for a specific discipline (for instance, engineering or math). All of the [POs](#) in this program will then use the excel spreadsheet to find reviewers. [POs](#) can also store reviewer information in an informal collection of notes or business cards. We obtain information about the individual [POs](#)' reviewer storage methods

informally through emails and meetings. By collecting the fields of each of these databases, we create a universal field list that includes all potential fields from each program. [\[19\]](#)

## **Database Validation**

The database also needs to be validated by the Program Officers. The [POs](#) make sure that the database is easy enough to use and is easily updatable. There is also a comment sheet handed out when testing the database. Modifications of the database are made according to the feedback of the [POs](#).

## **Database Creation and Description**

### **Database Survey**

We conducted a survey about the database to determine the tasks that a [PO](#) should be able to perform by using the database and what reviewer information the database should store. The survey was emailed to the [DUE POs](#), who were given a week to respond (See [Appendix M](#) for Survey).

The fields our liaison told us must be included in the database were: name, email address, phone number, institution, department, review history (such as dates of panels and the programs that the panels were for), and review scores (from our evaluation tool). Other fields our liaison thought might be important were: gender, ethnicity, and institution type. The tasks that we intended the database to perform were: add a new reviewer to the database, add an additional score to an existing reviewer, add or change information in existing



reviewer entries, add an entry into the reviewer history, and search for a reviewer. We asked the [POs](#) if they wanted to include any other fields or be able to perform other tasks with the database.

We obtained a total of nine responses. The most frequently requested features were additional subfields to identify area of expertise, specific institution type (for instance, community college or minority serving) and review history. We then created the database (See [Appendix N](#) for Survey Analysis).

## **Database Creation**

The database has two main functions: add new reviewers and search for reviewers. The [initial screen](#) that users see when opening the database is a “switchboard,” with buttons linking to both of these functions. Each function also has a “back” button, linking to the switchboard screen.

Behind the screens that the user interacts with are a number of tables. The information tracked in the database is separated into four categories: Personal Information, Professional Information, [NSF](#) Information, and Review History. “Personal Information” includes first name, last name, email, phone number, gender, state, ethnicity, web site, and notes. “Professional Information” includes last name, highest degree, primary field of expertise, four subfields, institution, department, and institution type. “[NSF](#) Information” includes last name, the number of times the reviewer has reviewed for [DUE](#), whether the person is a current or former [PI/Co-PI](#) or not, and the reviewer’s average score. An identification number is automatically assigned to each reviewer. The

identification number links the “Personal Information”, “Professional Information”, and “[NSF](#) Information” tables together and it allows the computer and the user to distinguish between reviewers with similar information. The “Review History” table consists of panel information for all reviewers. The specific fields included in this table are: last name of reviewer, month and year of panel review, program reviewed for, score for style, score for rating justification, score for program solicitation, and average score for that panel.

To [add a reviewer](#) to the database, the user enters information into fields, several of which are drop-down boxes. Drop-down boxes ensure that the user enters information correctly and consistently into fields with a limited number of choices, such as state, gender, or ethnicity. Other fields, such as name or institution, cannot be drop down boxes, since they have infinite possibilities. The user does not need to enter information into every field; some may be left blank. The information is automatically added to the Personal Information, Professional Information, and [NSF](#) Information tables. After entering one reviewer, the “Clear All Fields” button clears the information from all of the fields to facilitate entering multiple reviewers quickly and easily.

To [search the database](#), the user enters criteria into the fields. The user can search on any of the fields tracked in the database. Some of the fields may be left blank. The database is searched with “AND” between each of the criteria. For instance, if the user inputs “math” for the primary field of research and “female” for the gender, the result will be all reviewers who satisfy “math AND female,” as opposed to “math OR female.” The database search yields [summary](#)

[information](#) about all reviewers who satisfy the criteria, ranked by their score average. The user can click on the “[details](#)” button to get more information about an individual reviewer, including [review history](#).

To [update reviewer information](#), the user searches for the reviewer in the search function. When the results from the search come up, selecting the “Details” button will allow the user to view or update that reviewer’s information. [Adding a score](#) to a reviewer’s information is also done through the search function. When the details of a reviewer are shown, there is an option to “Add a score”. When adding a score, the user fills in the score for each of the three rubric questions, the program reviewed for, and the month and date of the panel. The database will calculate the total panel score and update the overall average score. (See [Appendix O](#) for all Database Screen Shots)

## **Database Modifications and PO Comments**

Our initial database design used a reviewer’s last name as the primary identification. Also, “Add Reviewer”, “Update Reviewer Information,” “Add Scores to Reviewer,” and “Search for a Reviewer,” were all separate functions. This creates problems because it is almost certain that two or more reviewers will have the same last name. Also, in order to update information or add scores, the [PO](#) would have to know the exact last name of that particular reviewer. We had assumed that an [NSF](#) identification number could be used, however, these numbers are not easily accessible or regularly used. It would be even more inconvenient to have to type in an identification number than it would be to type

in a last name. Also, Dr. Pimmel indicated that he preferred that we not need the identification number at all.

The design solution to this problem is for Access to automatically assign an identification number when the user enters a new reviewer, and to include the “Update Reviewer Information,” and “Add Scores to Reviewer” functions under the Search function, as described in the database creation section. This eliminates the user’s need to know a reviewer’s last name before updating or adding scores, and also allows multiple people with the same last name to be stored in the database.

When the design was finalized, we created a set of instructions for the database, followed by a test for a [PO](#) to evaluate how easy and useful it is (See [Appendix P](#) for instructions and test). Dr. Lee Zia agreed to test the database for us. Dr. Zia made several suggestions as to how to improve the database, such as adding further details to the button names to clarify their functions, adding more back buttons to take someone to the switchboard on every screen, and changing the options in some of the drop-down boxes. He particularly liked that our database has the capability to show a reviewer’s history at NSF, as well as track their performance over time.

## Societal Implications of Evaluation Tool and Database

This section will discuss the relevance of our project within the National Science Foundation and society.

### Scoring Reviewers

A possible conflict of our project is determining if reviewers should be evaluated while volunteering their time to the [NSF](#). This tool would evaluate reviewers for services they voluntarily perform.

As a federally funded organization the [NSF](#) has a duty to make the most effective use of its funding. The primary purpose of peer review is to spend tax dollars on the most promising research proposals. The [NSF](#) must efficiently utilize its resources.

The [NSF](#) also has a duty to the research community to select the most promising proposals for funding. Being a reviewer at the [NSF](#) is seen as a privilege and a professional obligation. Reviewers who are not doing their job in the capacity they are expected to reflect poorly on the reputation of the [NSF](#). Our evaluation tool and database will help the [DUE POs](#) avoid reviewers who repeatedly have issues with their performance.

Evaluating reviewers is not a judgment of their professional abilities. Instead, they provide information on that person's usefulness as a reviewer, which makes no negative implications about their skill as a researcher or their technical abilities in general.

## Outside Evaluators

The [DUE](#) needs a formalized method of evaluating reviewers and a central location to store [POs](#)' evaluations and reviewer information. The [POs](#) each evaluate the reviewers in some way; however, they should not be the people who create the evaluation process. [DUE](#) needs an independent source to evaluate the reviewers.

Every [PO](#) has a different personal method of evaluating reviewers. If a [PO](#) or group of [POs](#) created the process, then they would have preconceived notions of how to appropriately evaluate a reviewer that would not necessarily be representative of all [POs](#).

[POs](#) may also have some reluctance to evaluate people who they know personally or are very well known and respected researchers. This is especially true if the [PO](#) is assigning the reviewer a poor rating. Because we, as an objective third party, created the method of evaluation, we removed the degree of personal feeling from the evaluation process. We gathered information and opinions from all of the Program Officers as to what an appropriate method and criteria would be for the evaluation process, so that each [PO](#) would have an equal say in the creation. We also made the criteria in the rubric as objective as possible to eliminate any conflicts of interest when the [POs](#) perform the evaluations. For instance, while a question such as "Is the reviewer knowledgeable about the field that the proposal deals with," might be useful to score on a rubric, it is not only difficult to judge from a one-page review, but a negative score can also feel like a personal attack to the person performing the

evaluation, especially if they know the reviewer. Questions such as, “How long is the review?” “Did the reviewer address strengths and weaknesses,” or “Did the reviewer supply constructive criticism for the [PI](#),” are questions that are useful and informative, are easy to answer by reading a review, and also do not question the reviewer’s technical competence.

## Informing the Reviewers

Our group and [DUE](#) had to make the decision whether or not to inform reviewers that they were being evaluated. One problem that could occur if reviewers knew they would be receiving a score is that many would want to know how well they did. [DUE POs](#) do not have the time to tell every reviewer the scores from the last panel. Also, if a reviewer, for any reason, did not get asked back to the [NSF](#), that reviewer could have undue concern about the evaluation scores.

The [DUE](#) has decided not to inform reviewers of the evaluation process. [DUE](#) plans to show reviewers the criteria from the scoring rubric as a guideline for how to write an excellent review. They will not tell the reviewers that it is a rubric or that they will be receiving scores based on it to avoid potential problems and complications. These criteria will inform the reviewers the best way to do a review without telling them they are being scored.

## Benefits to Program Officers

Our evaluation tool and database project will most directly affect the [DUE](#) Program Officers. A critical part of a [PO](#)'s job is to select reviewers to serve on panels. Good reviewer selection can improve the effectiveness of the peer review process.

Currently, the [DUE POs](#) all have their own methods of finding reviewers. Some are *ad hoc* (such as a stack of business cards) and others share a spreadsheet of reviewer names among the [POs](#) in specific disciplines. There is already an [NSF](#)-wide database in use; however, it is difficult to use. It is not [DUE](#) specific, and [DUE](#) users cannot perform searches based on criteria that would be useful to them (such as the number of times the person has reviewed at [NSF](#)). Thus, there is no [DUE](#)-specific location where reviewer information is easily accessible.

There is also no formal way to evaluate reviewers and track that information. Currently, [POs](#) informally assess a reviewer's performance while reading a review, but the evaluation criteria can vary between [POs](#), as we saw in our interviews. Storing standardized reviewer evaluations for future reference is especially crucial because about half of the [POs](#) are rotators who usually serve for 1-2 years and take their individual knowledge of reviewers and personal memory of reviewers' capabilities with them when they leave. Our project, a centralized database combined with a standardized method of evaluating reviewers, will facilitate communication among the [POs](#) about specific reviewers, making it easier to find the most useful reviewers. It will also preserve



knowledge about reviewers after a rotator has left, helping both the [POs](#) who stay behind and the rotator's successor. This will yield a more effective and efficient reviewer selection process.

## Benefits to Principal Investigators

There are two direct benefits to the [PI](#) of an improved reviewer selection process. First, with improved reviewer selection, it will be more likely that reviewers provide written reviews that include careful and detailed justification of their opinions and statements. Second, the [PI](#) receives all of the reviews from the panel. The [PO](#) can “redact” parts of the review, or remove them from the selection process, “because they contain irrelevant, non-substantive, or otherwise unusable statements, show evidence of bias, or contain intemperate personal attacks” [33]. The redacted selections, however, are not completely erased. Instead, they are marked with a strikethrough (~~strikethrough~~), so that the [PI](#) can still read them. Such reviews are not only insulting to the [PI](#), but also can make a [PI](#) doubt the legitimacy of the review process used on their proposal and damage their opinion of [NSF](#) in general. The database and evaluation process inform the [POs](#) which reviewers tend to not adequately justify their reviews or make remarks that are redacted. The [POs](#) acting on that information will improve the likelihood that the [PIs](#) get helpful and respectful feedback.

## Benefits to Reviewers

Since being a reviewer is regarded as not only an honor, but also a professional obligation, professionals who work in the sciences want to be chosen as reviewers. An evaluation tool that helps them know how to be a better reviewer means that they can improve the quality of their work.

The most useful reviewers and those who improve their performance according to the criteria on the evaluation rubric will benefit from the existence of the centralized reviewer database because their usefulness will be noted in the database. The review history shows all of the scores that a reviewer has received and the dates of the panel that they served on. It will be evident in the database if a reviewer starts off with poor review skills, but improves them.

Program Officers may have a potential bias in reviewer selection, in that they will tend to choose people who they know personally, are better known in their field of expertise, or who come from a large research institution. They minimize this bias, however, by balancing the panels by gender, ethnicity, and institution type. One [PO](#) responded in our database survey that those identifiers are, “absolutely essential. We have to balance our panels according to these attributes” [\[34\]](#). Maintaining diversity within the panels is a crucial obligation the [POs](#) have in their reviewer selection that will be made easier with the use of this database. The gender, ethnicity, and institution type information that is stored in the database will help the [POs](#) ensure diversity within the panels and expose them to a more diverse set of potential reviewers.

## Benefits to Society

The \$5.65 billion budget, of which about \$170 million goes to the [DUE](#), comes from United States tax dollars. Selecting the most qualified reviewers who would recommend funding for only the best proposals will facilitate cost-effective spending of American tax dollars.

The [POs](#) can spend more time making informed funding decisions because the reviewer selection process will be more. The reviewers chosen by the [POs](#) will be the most useful (shown by the scores) who will help the [POs](#) make the best funding decisions.

Underrepresented groups of the population will be used more often as reviewers through the [DUE](#) because of our reviewer database. The database keeps track of gender, ethnicity, and institution type (such as two year college or minority serving). These fields of information will assist the [POs](#) in choosing more diverse panels to review proposals.

The [DUE](#) solely funds proposals that look to improve undergraduate education. Funding the most promising proposals will result in the best education research being conducted. The general population of America will benefit by this research because undergraduates will be better educated.

## Recommendations

Selection and evaluation of reviewers will always be a large part of a Program Officer's job. These tasks can be made much more efficient and effective if the [DUE POs](#) follow a few of our recommendations.

### Recommendations for the Reviewer Database

#### **Recommendation 1: That someone in the DUE populates the reviewer database.**

We created the reviewer database as a place to store reviewer information with an easy-to-use interface, but it has no reviewer information stored in it. We recommend that the [POs](#) of the [DUE](#) gather all the reviewers they personally have kept track of, format the data so that it matches the format of the database, and input the information into the database. This will allow all the [POs](#) to use the database to its full potential, and search through reviewers from everyone's personal records, not just their own.

#### **Recommendation 2: That the DUE adds new reviewers to the database after each panel and updates their information regularly.**

Each [DUE](#) panel is made up of approximately 1/3 new reviewers. After each panel is done and the reviewers have been evaluated, the reviewer should be entered into the database with scores and all of the information that the [PO](#) knows. Also, the [POs](#) should update reviewer information that is already in the database. Every time someone reviews for the [DUE](#), that reviewer's score should be entered. If a reviewer's personal or professional information changes (for instance, they switch to a different institution), a [PO](#) should update the database with that change.

**Recommendation 3: That the Program Officers update the fields in the database to keep them current with DUE structure.**

The [DUE](#) undergoes change regularly. For instance, the [DUE](#) programs change names, and are added, eliminated, or combined into different programs. Currently, [POs](#) can search for reviewers or add new reviewers' program service information using a drop-down box with a list of the current programs; however, when the programs change, someone will need to update the choices in the drop-down box to reflect the change. Also, after using the database for some time, the [POs](#) might decide that they want to keep track of additional information, such as membership in professional societies. If the database is not regularly updated, it will become out of date in a few years. [DUE](#) will go back to looking for reviewers from their own records with limited communication between Program Officers or knowledge of reviewer usefulness.

**Recommendation 4: That the Program Officers add potential reviewers to the database.**

Program Officers are always searching for people who can be reviewers. They meet people at conventions, through membership in professional societies, or from journal articles. If the [PO](#) meets a person that they can not use in a panel immediately, though, that person's contact information may be lost. If the [POs](#) use the database to store all new potential reviewers, they will be able to find the potential reviewers' information easily.

## Recommendations for the Evaluation Tool

### **Recommendation 5: That the DUE Program Officers use our evaluation tool to score all reviewers.**

Our evaluation tool was designed to help the [POs](#) select the most qualified reviewers for the job. The [POs](#) should evaluate each reviewer and enter the scores into the database to inform everyone within the [DUE](#) how well each reviewer scored. The score history will indicate to the [POs](#) which reviewers should not get invited back and who should serve on another panel. If the [POs](#) use our evaluation tool after each panel, the information about reviewers can be shared across the [DUE](#).

## Recommendations for Future Projects

### **Recommendation 6: That future project groups establish friendly relationships with the POs right away.**

When we arrived at [NSF](#), we immediately met all of the Program Officers. They were very friendly and willing to help us with our project. Throughout the term, the [POs](#) contributed valuable input to our project, with the interviews, survey, and validation testing. It would benefit future project groups to develop the same relationship with the [POs](#) as we have.

### **Recommendation 7: That the NSF continues to be a sponsor for WPI IQPs.**

Overall, our group had a positive experience working on our project at the National Science Foundation. The people of the [DUE](#) were very friendly and helped our project in any way they could. Also, we learned a lot while working within the [DUE](#). Our particular project probably would not be able to be

continued because there is not much further work to be done on the database or evaluation tool, but there are many other things that students could help with in the [DUE](#). We recommend that the [NSF](#) continue to do projects with WPI students because other students would most likely have as good of an experience as we had.

## **Closing Statements**

Using qualitative research methods, we developed a standardized evaluation process for written reviews and a database of reviewers information, including evaluation scores. The rubric's effectiveness has been illustrated through validation testing that assessed its validity and inter-rater reliability. After several revisions a searchable and updatable database design has been created and will be ready for use after reviewers have been entered into it.

The potential impact of this project on the [DUE](#) or even the [NSF](#) is significant. [POs](#) now have a way of objectively assessing written reviews and storing that information for future reference. They can easily implement the evaluation of reviews, which will significantly improve the quality of reviews in a number of ways. [POs](#) usually try to have panels consisting of 1/3 experienced reviewers. Using the evaluation rubric, they can select experienced reviewers taking previous performance into account while minimizing personal bias. This rubric has the potential to help standardize selection of reviewers. The DUE-specific searchable and updatable database will be a useful tool for storing reviewer information. When [POs](#) retire or rotate out of DUE, their reviewer

records will be kept. After reviewers are entered into the database, the selection process will become more convenient and effective.

DUE Program Officers have expressed excitement and interest to start using the database and evaluation process, and want to know when they can get started. The database and evaluation process have also sparked interest from POs in other divisions of EHR, such as REC, or Research, Evaluation and Communication. One REC PO asked for a copy of our software, saying that something similar to our project has been planned for years. Now that we have created it, his division will have something to work from. A final result of our project will be to motivate other divisions to improve on their reviewer selection processes, causing change and improvement in the NSF.



## References

1. NSF at a Glance. NSF. 26 Sept. 2005  
<<http://www.nsf.gov/about/glance.jsp>>.
2. About Undergraduate Education (DUE). NSF. 14 Sept. 2005  
<<http://www.nsf.gov/ehr/du/e/about.jsp>>.
3. Education and Human Resources. NSF. 14 Sept. 2005  
<[http://www.nsf.gov/about/budget/fy2005/pdf/fy2005\\_14.pdf](http://www.nsf.gov/about/budget/fy2005/pdf/fy2005_14.pdf)>.
4. United States. National Science Foundation. NSF Funding By Account. 2006. 5 Oct. 2005  
<<http://www.nsf.gov/about/budget/fy2006/tables/OVERVIEW/Overview-01.xls>>.
5. Chase, Jody. United States. Education and Human Resources. National Science Foundation. NSF Merit Review Process. 4 Apr. 2005. 5 Oct. 2005  
<<http://www.nsf.gov/bfa/dias/policy/docs/meritrevoakland.pdf>>.
6. "Instructions for Reviewers: Course, Curriculum, and Laboratory Improvement (CCLI) Program." NSF, 2005.
7. CCLI Program. 15 Feb. 2005. National Science Foundation. 5 Oct. 2005  
<[http://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=5741&org=DUE&from=home](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5741&org=DUE&from=home)>.
8. United States. Course, Curriculum, and Laboratory Improvement Program. Department of Undergraduate Education. CCLI Program Solicitation. 9 June 2005. 5 Oct. 2005  
<<http://www.nsf.gov/pubs/2005/nsf05559/nsf05559.htm>>.
9. Petruccelli, Joseph D. Personal Interview. 19 Sept. 2005.
10. Grant Proposal Guide. 01 Sept. 2004. NSF. 14 Sept. 2005  
<[http://www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=gpg#top](http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg#top)>.
11. Pimmel, Dr. Russell L. "RE: WPI Student Project." 29 Sept. 2005.
12. Pimmel, Dr. Russell L. Telephone interview. 11 Oct. 2005.
13. Colwell, Rita R. "NSF Merit Review Criteria." NSF. 20 Sept. 1999. 14 Sept. 2005  
<<http://www.nsf.gov/pubs/1999/nsf99172/nsf99172.htm>>.

14. Cole, Stephen, Leonard Rubin, and Jonathan R. Cole. Peer Review in the National Science Foundation: Phase One of a Study. Washington, D.C.: National Academy of Sciences, 1978. 1-19.
15. Rissmiller, Kent. "Interviewing." ID 2050 class, WPI. 26 Sept. 2005.
16. Staff Directory. NSF. 05 Oct. 2005  
<[http://nsf.gov/staff/staff\\_list.jsp?org=DUE](http://nsf.gov/staff/staff_list.jsp?org=DUE)>.
17. Rissmiller, Kent. "Question Design and Surveys." ID 2050 class, WPI. 21 Sept. 2005
18. Jolliffe, F R. Survey Design and Analysis. Chichester, England: Ellis Horwood Limited, 1986.
19. Rissmiller, Kent. "More Methods- focus groups, history, and content analysis." ID 2050 class, WPI. 28 Sept. 2005.
20. Mooney, Chis. "The Politics of Peer Review." 8 Jan. 2004. 15 Dec. 2005  
<http://www.csicop.org/doubtandabout/peerreview/>.
21. Van Rooyen, Susan. "The Evaluation of Peer Reviewer Quality." Learned Publishing 14 (2001). 15 Dec. 2005.
22. Towne, Lisa, Jack M. Fletcher, and Laress L. Wise. Strengthening Peer Review In Federal Agencies That Support Education Research. Washington, D.C.: The National Academies P, 2004. 15 Dec. 2005  
<<http://www.nap.edu/books/0309090997/html/R1.html>>.
23. Davidoff, Frank. "Improving peer review: who's responsible?" Editorial. Biomedical Journal 20 Mar. 2004.
24. Pimmel, Dr. Russell L. Telephone interview. 11 Oct. 2005.
25. Rissmiller, Kent. "Interviewing." ID 2050 class, WPI. 26 Sept. 2005.
26. Staff Directory. NSF. 05 Oct. 2005  
<[http://nsf.gov/staff/staff\\_list.jsp?org=DUE](http://nsf.gov/staff/staff_list.jsp?org=DUE)>.
27. Jolliffe, F R. Survey Design and Analysis. Chichester, England: Ellis Horwood Limited, 1986.
28. Rissmiller, Kent. "More Methods- focus groups, history, and content analysis." ID 2050 class, WPI. 28 Sept. 2005.

29. Brown, James. "What issues affect Likert-scale questionnaire formats?" Shiken:JALT Testing & Evaluation SIG Newsletter. Apr. 2000. University of Hawaii. 1 Nov. 2005 <[http://www.jalt.org/test/bro\\_7.htm](http://www.jalt.org/test/bro_7.htm)>.
30. Gibson, Shanan. Holistic versus Decomposed Ratings. Diss. East Carolina Univ., 1999. 1 Nov. 2005 <<http://harvey.psyc.vt.edu/Documents/SIOP-8-Gibson.pdf>>.
31. "Developing a Scoring Rubric." Instructional Intranet: Chicago Public Schools. 2000. Chicago Board of Education. 6 Nov. 2005 <[http://intranet.cps.k12.il.us/Assessments/Ideas\\_and\\_Rubrics/Create Rubric/Step\\_5/step\\_5.html](http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Create_Rubric/Step_5/step_5.html)>.
32. Woodin, Terry. Personal Communication, 12/5/05.
33. NSF. Proposal and Award Manual. 12/12/2005. <<http://www.inside.nsf.gov/pubs/2002/pam/pam.pdf>>
34. Sears, Curtis. Personal Communication, 11/16/05.
35. Pimmel, Russell. Personal Communication, 10/24/05.
36. Jones, Larry. "Accuracy, Precision, Average Deviation, Error." 7 Nov. 2005. 15 Dec. 2005 <[http://www.sciencebyjones.com/average\\_deviation.htm](http://www.sciencebyjones.com/average_deviation.htm)>.

## **Appendix A: Sponsor Description – About the NSF**

The National Science Foundation is a funding organization that caters to diverse types of research. The NSF is divided into six offices and seven directorates; Office of Director, Office of Budget, Finance, and Management, Office of Information and Resource Management, Office of Polar Programs, Office of International Science and Engineering, Office of Cyber Infrastructure, Directorate for Biological Sciences, Directorate for Computer and Information Sciences and Engineering, Directorate for Engineering, Directorate for Mathematical and Physical Sciences, Directorate for Geosciences, Directorate for Social, Behavioral, and Economic Sciences, and Directorate for Education and Human Resources. Each of these Directorates is further divided into Divisions, and divided again into Programs [1].

### **DUE and CCLI Statistics**

The Division of Undergraduate Education is a division of the Education and Human Resources directorate of the [NSF](#). The [DUE](#) encompasses all aspects of undergraduate education in the [NSF](#). In FY2004, Congress approved \$841.42 million for this directorate. Dr. Russell Pimmel, the sponsor for this IQP project at the NSF, is a Program Director for the Course Curriculum and Laboratory Improvement ([CCLI](#)) Program, so knowledge in this area is particularly relevant [7]. The [CCLI](#) is a program within the [DUE](#), and the goals of this program include: Conducting research on undergraduate [STEM](#) teaching

and learning; Creating learning materials and teaching strategies; Developing faculty expertise; Implementing educational innovations; and Assessing learning and evaluating innovations [7]. The amount of money given to the [CCLI](#) in 2004 was \$93.2 million [4].

## **Appendix B: Interview Questions**

### **Interview Introduction**

- Confidentiality/quoting
- Explain project goals
- Creating a rubric to generate reviewer scores in a few questions with a clearly defined numerical answer
- Our hope is to eventually put the scores into a database that will rank reviewers

### **Interview Questions**

- What qualities make a good review when you are trying to make a decision?
  - Content
  - Readability
  - Broader impact/ intellectual merit
  - Strengths/weaknesses
  - All categories
- Problem areas?
- PI point of view (improvement for PI)
- If you were going to score a review, what features would you use to evaluate the review?
  - Readability
  - Ask description of scoring for one category
  - Do you feel any of these areas are more or less important?

**Give notice of future database survey.**

## **Appendix C: Interviews: Review Qualities**

**S/W** - address strengths and weaknesses

**BI/IM** - address broader impacts and intellectual merit

**Grammar** - includes grammar and spelling, any technical problems w/reviews

**CC** - constructive criticism for the PI, no rude comments

**Justify** - review should be consistent with rating and give justification as to why they assigned that rating, give reasoning for everything they say about the proposal, more than just a summary

**Proposal** - should show they took time to read and understand the proposal  
**Reference** - no personal references in the review

**POC** - provides information to use in the PO comments.

**Expertise** - should show some technical knowledge/expertise in the field of the proposal

**SOA** - should show they know the “state of art”, new research/assessment ideas

**Solic.** - should refer to NSF solicitation criteria

**Panel** - interact well with the panel, keeps conversation moving, asks questions, no “rigid” personalities

**Clarity** - clear statements about proposal

**Scribe** - acts as a good scribe during panel session; panel summary should not just be a copy of his/her review

**Fund info** - tells PI how they can get funded, tells what PI should improve on

**Diverse** - diversity of expertise, can give good reviews on all the proposals in the panel, not just one or two

### Review Qualities

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
S/W	1		3	4	5		7		9	10	11	12	13	14	15	16	17
BI/IM	1	2	3	4	5		7		9			12	13	14		16	
grammar	1		3			6			9		11	12					
CC	1	2	3	4	5	6	7		9	10	11		13	14	15	16	17
justify	1	2	3	4	5	6			9	10				14	15	16	17
proposal	1	2			5	6	7	8		10			13				
reference	1																
POC		2	3				7										
expertise		2					7	8	9		11		13		15		
SOA		2						8			11						
solic		2											13		15		17
panel		2	3			6											
clarity				4	5		7										17
scribe						6			9								
fund info			3	4	5	6		8	9	10	11						
diverse								8	9		11						



## Appendix D: Interviews: Review Problem Areas

**N/E info** - not enough information to back their statements about the proposal, just giving a rating, not enough detail

**Rude** - comments on proposal are too harsh, doesn't respect the PI

**Misunder** - reviewer does not understand the proposal; reviewer has no knowledge of subject matter

**Inapp** - inappropriate comments, prejudiced comments

**Missing** - review is missing comments that are supposed to be included about every proposal, or focuses on only one aspect of the proposal

**Panel** - problems on the panel, ranging from "rigid" personalities, to getting into a group think mode

**Read** - did not take enough time to read the proposal thoroughly

**Reference** - reviewers did not exclude reference to themselves; review would have to be thrown out

**Grammar** - problems with the grammar or style of the review

**Misrep** - broader impacts has 4 areas, some reviewers only think BI includes under represented minorities

**Criteria** - quote the GPG, comment on criteria that the PO has already looked through the proposal for

**Summary** - just gives a summary of the proposal, gives no opinion on the proposal

**Quantity** - the review says a lot but does not say anything worthwhile, long but not good quality

**Check** - sometimes reviewers make a checklist of what they think should be included in their proposal, and do their review based solely on the checklist, sometimes basing checklist on everything included in the solicitation

### Review Problem Areas

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
N/E info	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
rude	1			4			7								15	16	17
misunder	1		3				7	8									
inapp	1		3	4												16	
missing	1							8			11		13				17
panel		2						8									
read		2									11						
reference				4													
grammar					5		7				11	12	13	14	15	16	
misrep					5				9					14			
criteria						6											
summary						6		8	9	10		12		14			
quantity										10						16	
check									9								

## Appendix E: Interviews: Suggested Evaluations

**S/W** - rate how well they address the strengths and weaknesses of the proposal and give details to why they are strength and weaknesses

**Criteria** - rate if they address BI and IM and the other criteria included in the NSF solicitation

**Quality** - the quality of the review, grammar, you can tell the reviewer took time to read the proposal, make argument clear

**PI feed** - rate the feedback the reviewer provides for the PI, should give suggestions on how to improve

**Reason** - provide quality reasoning for all statements about the proposal

**Expertise** - can tell the reviewer has expertise in the subject area, they understood the proposal

**Rating** - review matches the rating given by the reviewer

**Panel** - interaction with other panel members

**PO** - helpful in making PO comments and review analysis

**Holistic** - thinks we should give one holistic rating of reviewers and POs would be able to make additional comments if they wanted to

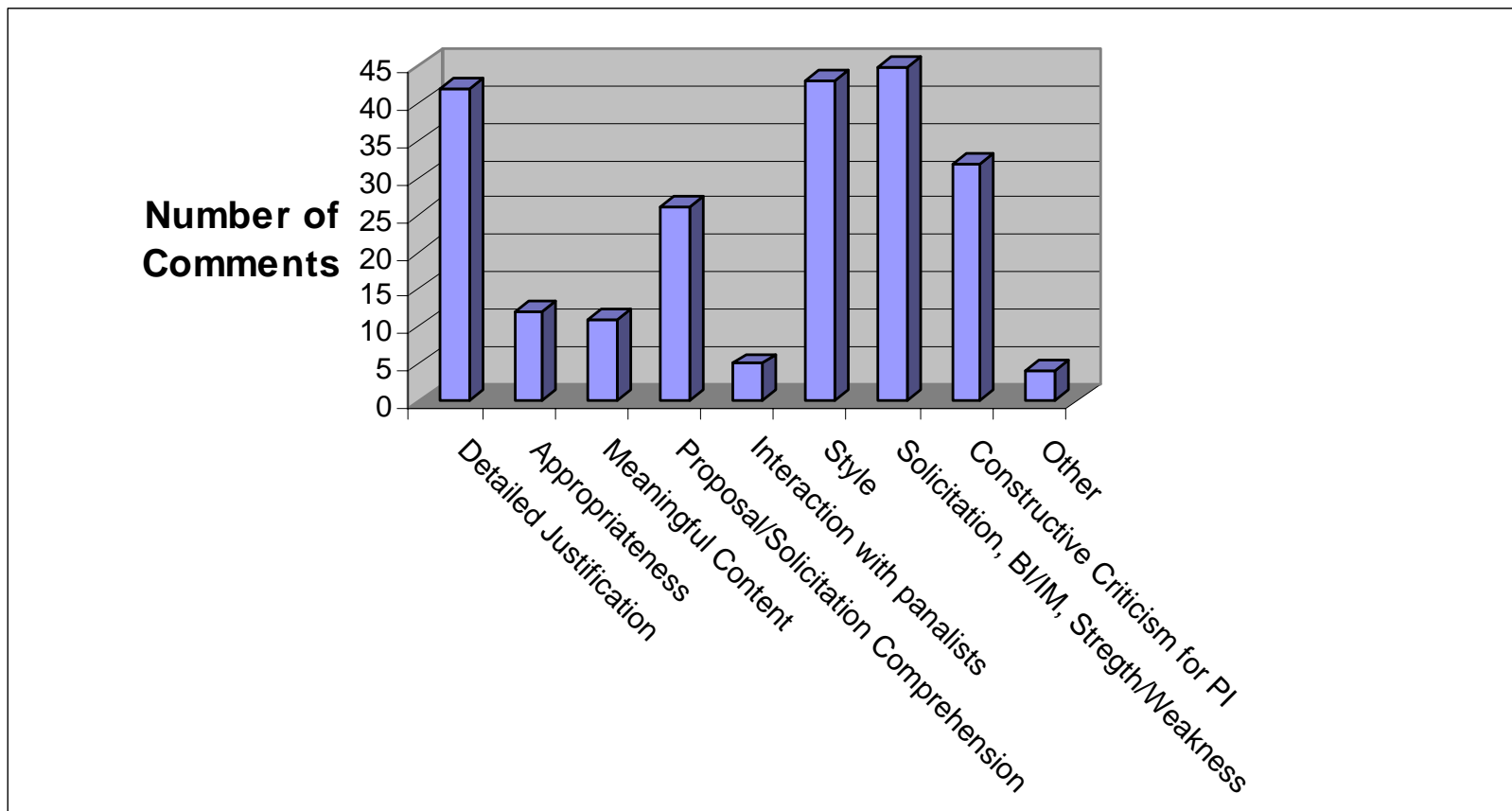
**Length** - should rate on the length of the review

**Profess** - professionalism in the review

### Suggested Evaluations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
S/W																	
criteria																	
quality																	
PI feed																	
reason																	
expertise																	
rating																	
panel																	
PO																	
holistic																	
length																	
profess.																	

## Appendix F: Interview Analysis



## **Appendix G: First Draft Rubrics**

### **Rubric #1**

#### **Quality of the Review**

**1** - does not give sufficient information on the proposal, does not properly address criteria of the NSF solicitation, shows the reviewer did not take time to read and understand the proposal, not a long review, one or two lines at most, does not at all justify statements made about the proposal

**2** - shows some understanding of the proposal, addresses parts of the solicitation, puts more than one line for sections of the review, statements made are not clearly justified

**3** - shows a good understanding of the proposal, gives justification of statements made, has paragraphs in each section of the review, addresses each part of the NSF solicitation

**4** - gives detailed information on all NSF solicitation criteria, gives proper justification for every statement made about the proposal, long review, good length paragraphs for each section

#### **Addressing Strengths and Weaknesses**

**1** - does not at all address strengths or weaknesses, PO cannot tell why the reviewer gave the review the rating, does not clearly address broader impacts and intellectual merit

**2** - lists some strengths and weaknesses but does not give reasoning for the statements made, addresses broader impacts and intellectual merit but does not include all aspects of both

**3** - gives strengths and weaknesses and some reasoning for each strength and weakness, PO can understand why a certain rating was given, includes both broader impacts and intellectual merit

**4** - gives strengths and weaknesses with clearly stated reasons why the reviewer thinks they are strengths or weaknesses, helps PO make decision, clearly addresses all aspects of broader impacts and intellectual merit

#### **Helpfulness to the PI**

**1** - reviewer does not give constructive criticism on the proposal, gives no information on how to improve the proposal, does not help PI understand why the proposal was given a certain rating

**2** - provides PI with little criticism; PI can identify broad areas of improvement,

**3** - gives constructive criticism to the PI; the PI will be able to understand exactly what areas need to be improved in the proposal

**4** - PI will know why the decision was made, gives PI polite, constructive criticism on how to improve the proposal, clearly states what the PI needs to do to get funded

Rubric #2

1) Satisfies Reviewer Directions	
<b>1</b>	<b>2</b>
Barely addresses strengths and weaknesses, if at all. Does not use any specifics from the proposal to justify rating. Length is a few sentences at most. Some sections may not be completed.	Mentions only strengths or only weaknesses relative to each criterion. Minimal use of specifics to justify rating. All sections are completed with at least a few sentences.
<b>3</b>	<b>4</b>
Discusses strengths and weaknesses relative to each criterion. Uses specifics to justify rating. Length is at least a few short paragraphs	Detailed discussion of strengths and weaknesses. Rating is completely justified by the specifics given. Length is at least several well-developed paragraphs.

2) Preparation for Review	
<b>1</b>	<b>2</b>
No mention of program solicitation or discussion of proposal's relevance to proposal solicitation. Misses major points of proposal. Generally characterized as a summary of the proposal.	Mentions program solicitation, but little to no discussion of proposal's relevance to proposal solicitation. Discusses major points of proposal but no discussion of details. More than just a summary.
<b>3</b>	<b>4</b>
Some discussion of proposal's relevance to proposal solicitation. Discusses entire proposal. Summary information is secondary to review.	Detailed understanding and discussion of proposal's relevance to proposal solicitation. Knowledge of entire proposal. No summary information.

3) Helpfulness to Principal Investigator	
1	2
No constructive criticism. No suggestions for improvement. May include some grammar or spelling errors.	Includes minimal constructive criticism or suggestions for improvement, but not both. May include some grammar or spelling errors.
3	4
Includes constructive criticism and some helpful suggestions for improvement. Very few, if any, errors in grammar or spelling. Politely and clearly worded.	Constructive criticism is insightful and suggestions are helpful. No errors in grammar or spelling. Politely worded. Very clearly written.



## Rubric #3

### **Did the reviews help you make good funding decisions?**

**1 Unacceptable** The reviews did not express an opinion or include details. The reviews did not follow the NSF solicitation. They were generally characterized as one sentence or a summarization of the proposal. Not helpful to you as a PO.

**2 Acceptable** The reviews usually expressed an opinion and included details. Following the NSF solicitation, the review attempted to address Broader Impacts or Intellectual Merit. As a PO, you can generally identify good or bad proposals.

**3 Good** The reviews expressed an opinion and supported it with details. Following the NSF solicitation, the review appropriately addressed Broader Impacts and Intellectual Merit. As a PO, you can almost always identify good or bad proposals with minimal difficulties.

**4 Excellent** The reviews expressed a very clear opinion and always supported it with elaborate details. Following the NSF solicitation, the review appropriately addressed Broader Impacts and Intellectual Merit. As a PO, you can easily identify good or bad proposals without any problems.

### **Were the reviews helpful for the PI?**

**1 Unacceptable** The review failed to provide any criticism or justification for the proposal rating. The review does not help the PI know what the strengths or weaknesses of the proposal were.

**2 Acceptable** The review provided the PI criticism. Strengths or weaknesses of the proposal where were sometimes supported with specifics that explained the rating. The PI should be able to identify areas of potential improvement in their proposal.

**3 Good** The review provided the PI with constructive criticism. Both strengths and weaknesses of the proposal where defined, and were supported with specifics that explained the rating. The PI will be able to understand exactly what can be improved with their proposal.

**4 Excellent** The review provided the PI polite, constructive criticism. Both strengths and weaknesses of the proposal where clearly defined, and were supported with specifics that explained the rating. Suggestions were made such that the PI may improve the proposal. The PI will be able to understand exactly what can be improved with their proposal and how to accomplish this improvement.

## **Appendix H: Six Question Rubric**

### **Length**

1. One sentence in each section
2. A few sentences, not detailed
3. Paragraph for each section
4. Several detailed paragraphs

### **Addressing Strengths and Weaknesses**

1. Does not address strengths or weaknesses
2. Only addresses strengths or only addresses weaknesses
3. Addresses both strengths and weaknesses

### **Constructive Criticism**

1. Does not give any criticism to the PI, or makes rude comments
2. Gives some criticism but is not specific about suggestions for improvement
3. Gives criticism with detailed suggestions for improvement

### **Grammar/spelling**

1. Makes major grammar and spelling mistakes
2. Makes few or no grammar and spelling mistakes

### **Addressing the Program Solicitation**

1. Does not address any criteria of the solicitation
2. Addresses some parts of the program solicitation
3. Addresses all applicable parts of the program solicitation

### **Justification for rating**

1. Rating not supported by specifics from the proposal
2. Minimal use of specifics to justify ratings
3. Rating completely supported by specifics from the proposal

## Appendix I: Two Question Rubric

<b>1) Style</b>	
<b>1</b>	<b>2</b>
One sentence per section OR Contains noticeable grammar or spelling mistakes	A few sentences; not detailed OR A few bullet points; not detailed
<b>3</b>	<b>4</b>
A paragraph for each section OR A more detailed bulleted lists	Several detailed paragraphs OR Very detailed bulleted lists

Note: Reviews receiving scores 2,3, or 4 contain NO noticeable grammatical mistakes

<b>2) Content</b>	
<b>1</b>	<b>2</b>
Does not address strengths or weaknesses OR Does not address criteria from the program solicitation OR Rating not supported by the review	Addresses only strengths or only weaknesses  OR Minimal use of specifics to justify rating
<b>3</b>	<b>4</b>
Addresses both strengths and weaknesses AND Minimal use of specifics to justify rating	Addresses both strengths and weaknesses AND Gives specifics from the proposal and solicitation AND Rating is completely supported in the review

## Appendix J: Final Rubric

### Style

1- Poor	One sentence per section OR Contains annoying grammar or spelling mistakes OR Makes inappropriate comments
2- Acceptable	(No annoying grammatical mistakes) A few sentences; not detailed OR A few bulleted points; not detailed
3- Excellent	(No annoying grammatical mistakes) Several detailed paragraphs OR Very detailed bulleted lists

### Justification of Rating

1- Poor	Does not address strengths or weaknesses OR Rating not supported by specifics from the proposal
2- Acceptable	Addresses only strengths or only weaknesses OR Minimal use of specifics to justify rating
3- Excellent	Addresses both strengths and weaknesses AND Rating completely supported by specifics from the proposal

### Program Solicitation

1- Poor	Insufficiently addresses criteria from the program solicitation Or Does not address broader impacts or intellectual merit
2- Acceptable	Addresses the majority of the program solicitation AND Addresses broader impacts and intellectual merit with minimal detail AND Gives areas of improvement that the PI may use
3- Excellent	Addresses all applicable parts of the program solicitation AND Addresses broader impacts and intellectual merit in detail AND Gives criticism to the PI with detailed suggestions for improvement

## **Appendix K: Rubric Validation Survey**

Thank you for participating in the validation testing of our evaluation rubric.

Attached is a word document with five sample reviews selected from different reviewers and proposals. We would like you to score each review on the attached spreadsheet and answer the brief survey below. Keep in mind that in the future, scoring will be done on a computer.

NOTE: If you have any questions, please ask us and do not discuss your answers or the rubric with anyone else.

To do list:

Read the attached reviews and score them in the excel spreadsheet.

After the spreadsheet is complete, please answer and return the survey below with the excel file attached.

Evaluation Rubric Survey

Is the rubric easy to use? Would you use it on a regular basis?

Does the rubric provide a simple and accurate measure reviewer quality?

Do you have any recommendations to improve the rubric?

Thanks again, we appreciate your help.

## Appendix L: Rubric Validation Scores and Survey Responses

R1Q1	2	2	2	2	1
R1Q2	2	2	2	1	2
R1Q3	1	2	1	2	1
R2Q1	3	3	3	3	3
R2Q2	3	3	3	3	2
R2Q3	3	3	3	3	3
R3Q1	1	1	1	1	2
R3Q2	1	1	1	1	1
R3Q3	1	1	1	1	1
R4Q1	1	1	1	1	2
R4Q2	1	2	1	2	2
R4Q3	2	2	1	2	1
R5Q1	2	3	2	1	3
R5Q2	1	2	2	1	3
R5Q3	1	2	2	2	2
R6Q1	1	2	1	1	1
R6Q2	2	1	1	2	1
R6Q3	2	2	1	2	1
R7Q1	2	2	3	2	2
R7Q2	1	2	2	3	2
R7Q3	2	2	2	3	2
R8Q1	1	2	1	1	2
R8Q2	1	2	1	2	1
R8Q3	2	2	1	1	1
R9Q1	1	1	1	1	1
R9Q2	1	2	1	1	1
R9Q3	1	2	1	1	1
R10Q1	3	3	3	3	3
R10Q2	3	2	2	3	2
R10Q3	3	3	2	2	3

## Rubric Survey Responses

1.

Interesting exercise.

2.

The rubric was somewhat difficult to use because of all the logical ANDs and ORs. I found myself just wanting to use the idea, "on a scale of 1 to 3, how well did the reviewer justify his or her rating [or other feature]?" Indeed, if I were doing this on a regular basis, I would just simply score the reviews 1, 2 or 3. This may be because I'm so used to looking at reviews and can quickly establish how useful they are to me as I do my work.

The last category of the rubric, Program Solicitation, was the hardest to use. In reality, I'm happy if a reviewer simply addresses intellectual merit and broader impacts. If they go farther and show deep knowledge of the solicitation, that's great, but I feel like it is more my job as a program officer to "have the solicitation memorized" (so to speak) and see how the proposal fits the solicitation.

On how to improve the rubric--I've always liked the idea of a visually-presented continuum, such as:

Length

1----- 2----- 3----- 4----- 5  
at least five complete paragraphs  
only three sentences

Grammar and Spelling

1----- 2----- 3----- 4----- 5  
highly readable with no grammatical and spelling  
errors  
many grammar and spelling errors

etc.

The user would circle the appropriate number and add up the tally. Descriptors can (and should) be given for the intermediate numbers, too.

3.

1. The rubric is easy to use.

2. Yes. All the reviewers, in my judgment were of poor quality. They did not provide sufficient reasons for their ratings and the reviews were of little help to program officers.

3. I like the rubric and it is well done.

4.

These were generally pretty minimal reviews and not very articulate.

5.

Is the rubric easy to use? Reasonably

Would you use it on a regular basis? No It is time consuming and doesn't really seem to serve a purpose.

Does the rubric provide a simple and accurate measure reviewer quality? I am not sure as much of a reviewer's purpose is the wisdom they supply during panel discussions. Their written remarks often do not capture this component of their contribution to the review process.

Do you have any recommendations to improve the rubric? No I think given your assignment you did an excellent job. I tried to judge this on the basis of the reviews I have seen from engineering panels. Judged on the basis of what I see in biology, these appear to be rather short on useful information to the PI concerning how their proposal might be improved, specific strengths and weaknesses and attention to some of the components listed in the Program Solicitation. The social science panels are even more helpful and complete in their comments so rating panelists would be a very discipline specific activity. It would also be very time consuming as it would be necessary to rate many reviews by a panelist before arriving at an even vague idea of their skills as a reviewer.

6.

Here are my results. It was easy to use, but I found myself being annoyed at the time it took. It would just be easier to say, "Don't invite this reviewer again." I realize there are problems with that approach. I think once I used the rubrics alot, they would not be annoying. Thanks for working on this.

7.

Rubric is easy to use although I found myself just looking for my gut reaction to the review as poor, acceptable, or excellent rather than looking through the rubric.

Rubric is useful but title headings may be just important and therefore, maybe a little more precision could be added.

Recommendations: I think it is hard to come up with examples of things that make up the rubric, maybe more important to concentrate on the headings. Also I like rating the style and justification. I'm not sure about rating the reviewers ability to address the program solicitation.



8.

Yes, it was easy to use.

Yes, kind of a go-no go type of evaluation[usefulness]

Nope[no recommendations]

9.

I thought the process went pretty smoothly.

10.

Easy to use; helps direct the evaluation.

I found it difficult to assess the quality of the reviewers – sometimes I gave a “2” because they met the criteria for a “2” but I still felt the review was in general poor.

I would like an additional, non-calculated, “overall” reviewer rating field.

## Appendix M: Database Survey

The following survey was sent via email on 11/0/2005.

Hi everyone,

Thank you for talking to us about the evaluation tool we are creating. Now we have some questions about our database design. If you could take some time to fill out this survey and get it back to us by Wednesday November 16<sup>th</sup> 2005 that would be greatly appreciated.

The database we are creating will be a resource for the program officers to use to track and find reviewers. For all the reviewers in the database you will be able to find contact information, expertise information, and reviewer scores from the evaluation tool if the reviewer served on a panel.

1. Listed below are some of the fields that will definitely be included in our reviewer database. Please list all additional fields you would use in the database.
  - a. Name
  - b. Email address
  - c. Phone number
  - d. Institution
  - e. Department
  - f. Review history
  - g. Review scores (from the evaluation tool we are creating)
  - h. Other areas we may include are: gender, ethnicity, and type of institution (community college, minority serving, etc.)

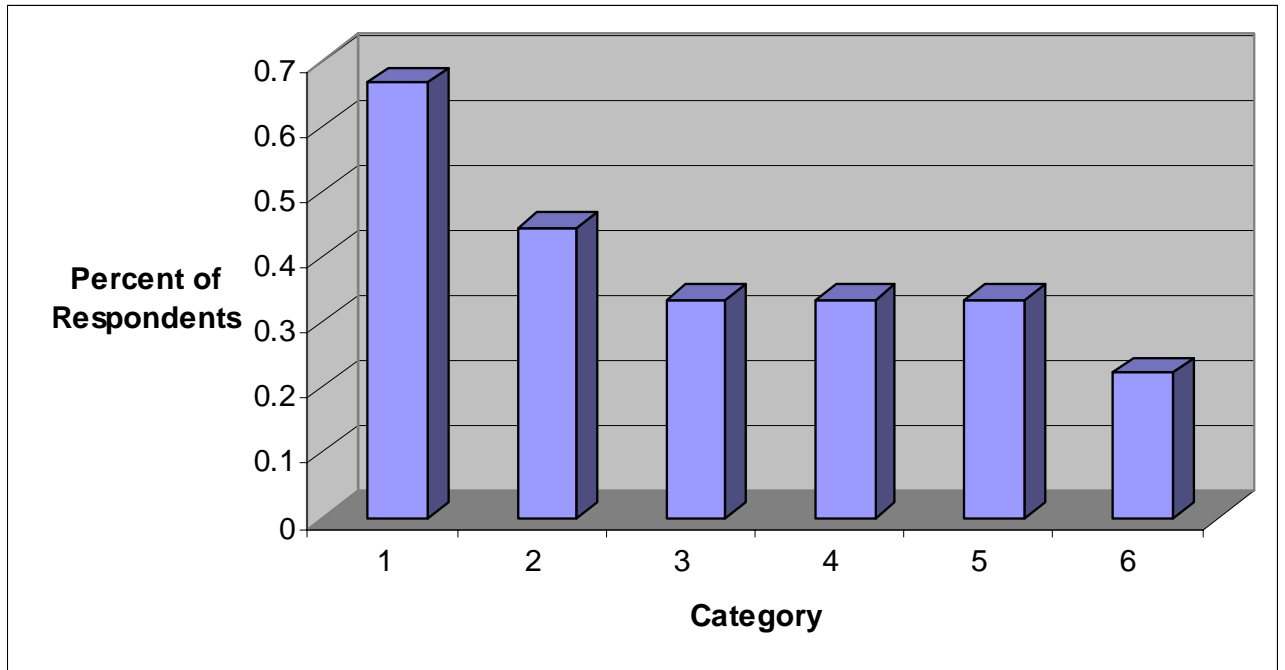
Please also list how many expertise descriptors you would like to see included and what kind of descriptors.

2. When creating the database, we plan to make it easily updatable and searchable. Our plan is to make it so that the program officers can easily...
  - a. Add a reviewer
  - b. Add an additional score to an existing reviewer
  - c. Add or change information in existing reviewer entries
  - d. Add an entry into the reviewer history
  - e. Search for a reviewer by certain characteristics

If there is anything else you would like to be able to do with the database, please list here.

3. Please list any other comments you have about creating a searchable and updatable reviewer database.

## Appendix N: Database Survey Analysis

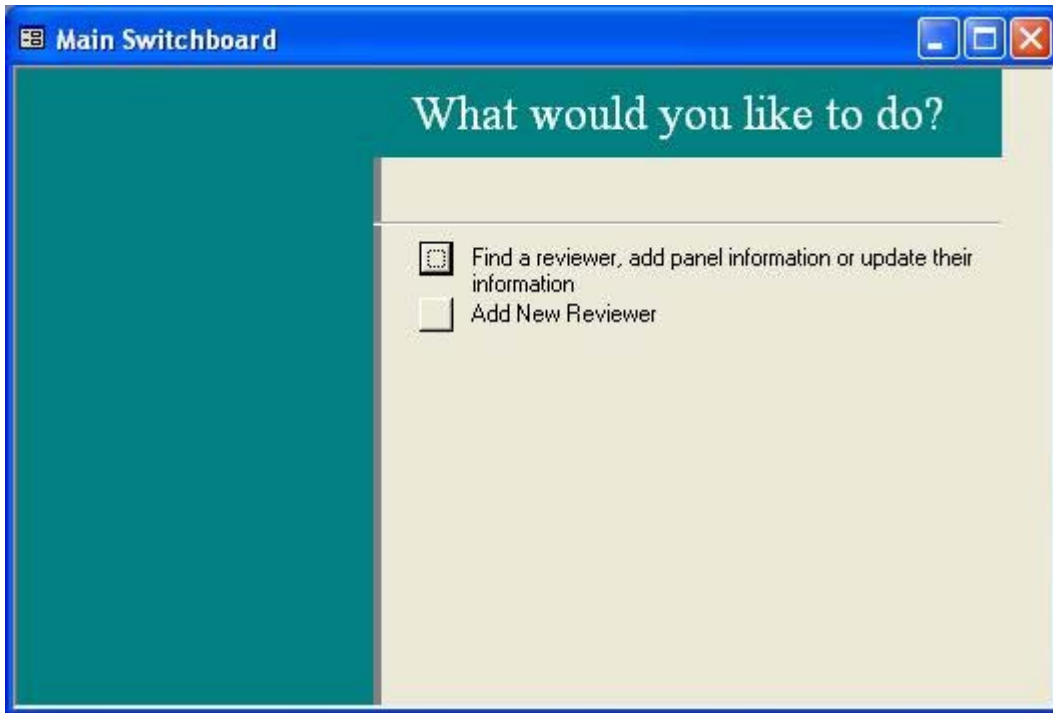


Key:

- 1) Subdiscipline
- 2) Additional Descriptors (ethnicity, gender, institution type)
- 3) Review History
- 4) Evaluation score (especially length)
- 5) Comments
- 6) Contact Information

## Appendix O: Database Screen Shots

### Switchboard:

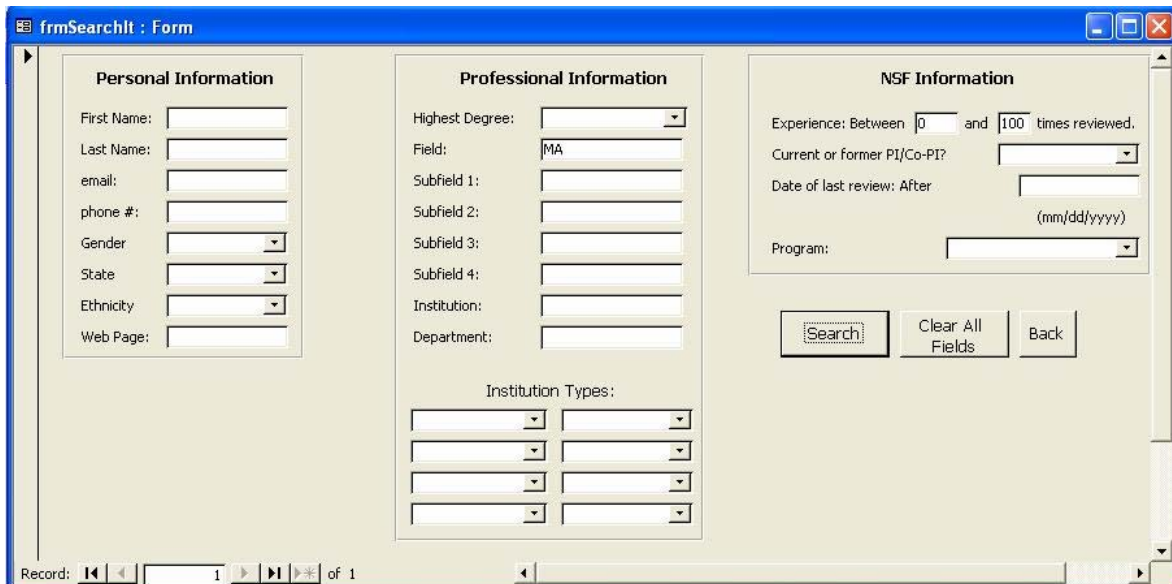


**Main Switchboard**

What would you like to do?

- Find a reviewer, add panel information or update their information
- Add New Reviewer

### Add Reviewer:



**frmSearchIlt : Form**

**Personal Information**

First Name:   
Last Name:   
email:   
phone #:   
Gender:   
State:   
Ethnicity:   
Web Page:

**Professional Information**

Highest Degree:   
Field:   
Subfield 1:   
Subfield 2:   
Subfield 3:   
Subfield 4:   
Institution:   
Department:   
  
Institution Types:

**NSF Information**

Experience: Between  and  times reviewed.  
Current or former PI/Co-PI?   
Date of last review: After   
(mm/dd/yyyy)  
Program:

Record:  of 1

## Search

frmSearchIt : Form

Personal Information	Professional Information	NSF Information
First Name: <input type="text"/> Last Name: <input type="text"/> email: <input type="text"/> phone #: <input type="text"/> Gender: <input type="text"/> State: <input type="text"/> Ethnicity: <input type="text"/> Web Page: <input type="text"/>	Highest Degree: <input type="text"/> Field: <input type="text" value="MA"/> Subfield 1: <input type="text"/> Subfield 2: <input type="text"/> Subfield 3: <input type="text"/> Subfield 4: <input type="text"/> Institution: <input type="text"/> Department: <input type="text"/>  Institution Types: <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>	Experience: Between <input type="text" value="0"/> and <input type="text" value="100"/> times reviewed. Current or former PI/Co-PI?: <input type="text"/> Date of last review: After <input type="text"/> (mm/dd/yyyy) Program: <input type="text"/>
<input type="button" value="Search"/> <input type="button" value="Clear All Fields"/> <input type="button" value="Back"/>		

Record: 1 of 1

## Search Results

qry\_SearchIt

Last Name	First Name	Email	Primary Field	Institution	Score Average	
Adcock	Merrick	madc	Chemistry		0	<a href="#">Details</a>
Cadoret	Jacki		Chemistry		0	<a href="#">Details</a>
Hills	Bonnie	bhil	Chemistry		0	<a href="#">Details</a>
Porter	Carola	cpor	Chemistry		0	<a href="#">Details</a>
Smith	Bob	bsmith@gmail.co	Chemistry	WPI	0	<a href="#">Details</a>
*						<a href="#">Details</a>

Record: 1 of 5

## Details

qryDetails

Personal Information	Professional Information	NSF Information
Last Name: <input type="text" value="Smith"/>	Highest Degree: <input type="text"/>	Number of times Reviewed: <input type="text" value="3"/>
First Name: <input type="text" value="Robert"/>	Primary Field: <input type="text" value="Chemistry"/>	Current or former PI/Co-PI?: <input type="text" value="No"/>
State: <input type="text"/>	Subfield 1: <input type="text" value="Molecular Design ar"/>	Program Last Reviewed For: <input type="text" value="CCLI"/>
Email: <input type="text" value="bsmith@wpi.edu"/>	Subfield 2: <input type="text"/>	Score Average: <input type="text" value="0.83333333333333"/>
Phone Number: <input type="text" value="555-555-5555"/>	Subfield 3: <input type="text"/>	Date of Last Review: <input type="text" value="02/05/2005"/>
Gender: <input type="text" value="Male"/>	Subfield 4: <input type="text"/>	
Ethnicity: <input type="text" value="Black"/>	Institution: <input type="text" value="WPI"/>	
Web Page: <input type="text" value="www.bobsmith.com"/>	Department: <input type="text" value="Chemistry"/>	
Notes: <input type="text"/>		
	Institution Type:	
	<input type="text" value="University"/> <input type="text" value="Co-ed"/>	
	<input type="text" value="Engineering"/> <input type="text" value="No Religious Affiliat"/>	
	<input type="text" value="Private"/> <input type="text" value="No Ethnic Affiliation"/>	
	<input type="text" value="Small"/> <input type="text" value="Not Industry"/>	
		<input type="button" value="Get Review History"/>
		<input type="button" value="Update Reviewer Information"/>
		<input type="button" value="Add Panel Info to Reviewer"/>
		<input type="button" value="Back"/>

Record:   1     of 1

## Review History

qryHistory

Date	Program	Style	Content	Average
01/02/1933	ATE	2	2	2.5
*		0	0	0

Record:   1     of 1

## Update Reviewer Information

The screenshot shows a web form titled "frmUpdate : Form" with three main sections:

- Personal Information:** Fields for First Name (Robert), email (bsmith@wpi.edu), phone #, Gender, State, Ethnicity, and Web Page.
- Professional Information:** Fields for Highest Degree, Field, Subfield 1, Subfield 2, Subfield 3, Subfield 4, Institution, and Department.
- Institution Type:** A grid of four dropdown menus.

At the bottom right, there are three buttons: "Update", "Clear All Fields", and "Back".

Record: 1 of 1

## Add Score

The screenshot shows a web form titled "qryGetID" with the following fields:

- Date: Text input field
- Program: Dropdown menu
- Style Score: Dropdown menu
- Content Score: Dropdown menu

Below the fields is a button labeled "Add Panel Information".

Record: 1 of 1

## Appendix P: Database Instructions and Test

Open the database. It is on the 'Q' drive, under the folder, "WPI Interns 2006."

Under the "Objects" tab on the left side of the screen, open the "Forms" window. Open the "SWITCHBOARD – START HERE" form. This screen links you to the different tasks that the database can perform. You can either (1) find a reviewer or (2) add a new reviewer.

- I. The "Find a reviewer" function can be used to select members for a panel, to update information about an individual reviewer, or to add panel information (date of panel, program, and evaluation scores) to an individual reviewer. To use this function:
  1. Click the *[Find a reviewer]* button.
  2. The Search Criteria form will appear.
    - i. If you are looking for a specific reviewer, type that reviewer's last name into the *[Last Name]* box.
    - ii. If you are searching for reviewers to fill a panel, type criteria into the various text boxes (or select appropriate values from the drop-down boxes) to search for reviewers who satisfy those criteria.

**Example:** To search for a Mechanical Engineer from the state of North Dakota, type "ME" into the *[Primary Field]* box, and select "ND" from the *[State]* drop-down box.
    - iii. If you don't know exactly what you're looking for, or if you don't know the exact spelling of the individual's last name, leave all of the search fields blank to display all of the reviewers in the database.
    - iv. Choose whether to sort your results by "Last Name" or by "Evaluation Score" in the *[Sort By]* box. The default is "Last Name."
  3. The results of your search, in summarized form, will appear in the order that you specified. To select a reviewer, click the *[Details]* button to the right.
  4. The "Details" screen shows all information about the reviewer and allows you to modify that information. You can:
    - i. *[View Review History]*: See the dates, programs, and evaluation scores received for all of the panels on which the reviewer has served.
    - ii. *[Update Reviewer Information]*: Change any information about the reviewer that you are looking at by typing the new information into the textboxes, or selecting the appropriate values from the combo boxes. Click *[Update]* to update the information.



- iii. *[Add Panel Info to Reviewer]*: Add the date, program, and evaluation scores for a panel on which the reviewer has served.
- II. The “Add New Reviewer” function is used to insert a new reviewer into the database. To use this function:
1. Perform a search on the reviewer that you are entering to make sure that the reviewer is not already in the database.
  2. Click *[Add New Reviewer]*.
  3. Type all known information about the reviewer into the text boxes, or select the appropriate values from combo boxes. NOTE: The more information you can enter, the more useful the database is.
  4. Click *[Add Reviewer]* to save this person into the database.
  5. To immediately add a panel’s date, program, and evaluation scores to this reviewer, click *[Add Panel Information]*.

**Tasks to test the database:**

1. Add yourself to the database. Check the personal information, professional information, and NSF information tables to check that you’re really there. (You’ll be added at the bottom.)
2. Pretend that you have switched home institutions and update your information. Check the tables to see what happened.
3. Add at least two panels to your history. Fill in at least your name and make up some scores in each category.
4. Close the tables and search for yourself based on different criteria, such as state, field of expertise, or gender. Try different combinations of criteria as well.
5. Find yourself in the results of a search. Click “Details” to see more of your information. Click “Get review history” to view your review history.

**Questions:**

1. Is the database easy to use? Are there any usability problems that should be fixed?
2. Are there any fields that should be added to the information? Are there any that you wouldn’t use and you think should be eliminated?
3. Are the functions sufficiently helpful, or should they be added to or changed in any way? How?
4. Will use of this database make the reviewer selection process any easier? Why or why not?

Any other comments?

## **Appendix Q: Project Description**

National Science Foundation  
Division of Undergraduate Education  
Proposal Reviewer Evaluation Process

The mission of NSF's Division of Undergraduate Education (DUE) is to promote excellence in undergraduate science, technology, engineering, and mathematics (STEM) education for all students. The division accomplishes its mission through several strategies including supporting curriculum development, stimulating and funding research on learning, and promoting development of exemplary materials and strategies for education. The primary mechanism is the funding of educational research and development projects at universities throughout the US. Faculty members submit proposals, usually in response to one of DUE's several solicitations, describing their project, and NSF project directors select the most promising proposals for funding.

A critical aspect in DUE's funding decisions is a peer review process where individuals from academia, industry, and other government agencies are asked to sit on a review panel to evaluate and rank a set of proposals. Usually, the reviewers have several weeks to read assigned proposals and write a formal review for each. They then meet as a panel to discuss each proposal and generate a panel summary capturing the group's discussion of each proposal. These individual reviews and panel summaries provide crucial information as the DUE program offices decide which proposal to fund. Thus, the effectiveness of the whole decision process depends very strongly on the effectiveness of the individual reviewers and, as with most processes involving volunteers; this varies considerably from one individual to the next.

Optimizing reviewer use is critical to efficient grant review. Some of the programs are more complicated than others and require experienced reviewers with specialized knowledge to deal with the complexity and sophistication of the proposals, while other programs are simpler and allow the use of less experienced reviewers. Similarly with many programs, domain specific content knowledge is essential for understanding the proposed projects and so it is critical to match the reviewer's expertise with panel needs while in other cases, general knowledge of a field is adequate.

Currently, reviewers are tracked in an informal process that relies on the program officers' ad hoc evaluations and personal recollections, leading to suboptimal selection of reviewers for a given panel. Since several of the program directors are rotators that join NSF for a year or two and then return to their permanent jobs, personal information about reviewers is lost with each rotator transition. Certainly, a formal process that evaluates each reviewer after every panel and tracks this information along with the reviewer's experience and expertise in a computerized database would make the panel formation process more robust. The WPI project would develop and implement a process for characterizing and evaluating individual reviewers, for storing this information in a convenient database and for searching the database to identify reviewers with specific characteristics.

## **Glossary**

**CCLI** - Course, Curriculum, and Laboratory Improvement

**DUE** - Division of Undergraduate Education

**EHR** - Education and Human Resources

**NSF** - National Science Foundation

**PI** - Principal Investigator

**PO** - Program Officer

**STEM** - Science, Technology, Engineering, and Mathematics