

PD-AAW-587

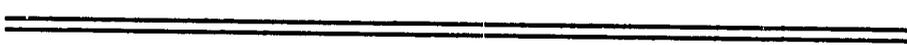
ISN 52749

EXECUTIVE SUMMARY
AID FY82 META-EVALUATION

Contract No. AID/SOD/PPC-C-0391
Work Order No. 2

Conducted for the Program Evaluation Systems Divisor
of the Office of Evaluation (PPC/E/PES)
of the Agency for International Development

April 1984

TRITON 

The purpose of this Project Summary is to provide a brief review of the results of a contract¹ performed by TRITON to implement a previously developed scoring instrument.² This instrument was designed to quantitatively assess the relative quality and completeness of USAID evaluation reports. Its design was intended to be generic so that it could be applied to evaluation reports addressing all stages of the AID project cycle (i.e., mid-term, end of project, etc.).

The instrument was to be implemented by scoring evaluation reports received in AID/Washington, thus establishing a data base that could be used to analyze the strengths and weaknesses of the reports. This would allow USAID to undertake corrective actions such as training, where appropriate.

This Project Summary briefly reviews:

- Purpose of Project and General Approach;
- Summary of Instrument Development;
- Usage and Utility of the Instrument;
- Data Analysis Plan for the Statistical Report;
- Pattern Analysis of FY82 AID Evaluations;
- Findings Report and Purpose of Study.

PURPOSE OF PROJECT AND GENERAL APPROACH

The primary purpose of this project has been to apply the quality/completeness assessment instrument on a large-scale basis. This involved the following tasks:

¹Work Order No. 2., Contract No. AID/SOD/PDC-0391.

²Work Order No. 1., Contract No. AID/SOD/PDC-0391.

- Scoring all (266)³ FY82 USAID evaluation reports. (The delineation of "FY82 reports" was made by Ms. Molly Hageboeck and Ms. Nena Vreeland of PPC/E, Office of Evaluation.)
- Scoring a random sample of forty evaluations each from FY81 and FY80.
- Scoring two completed Impact Series (Community Water Supplies in Developing Countries, Rural Roads Evaluation Summary Report) in composite fashion. This involved a comparison of all impact evaluations in the series, taken as a whole, based on a certain set of parameters.

Once the scoring was complete, an extensive array of statistical analyses were performed on the resultant data in order to:

- Assess the validity, interrater reliability and potential bias of the results.
- Identify any relationships between a variety of external characteristics and scores.
- Determine if any internal aspects of the evaluation reports clearly accounted for the relative differences among the scores.
- Identify any trends over time or significant differences between the 1980-1982 groups' scores.

The evaluations were scored by five TRITON staff members. Training in the instrument was conducted by Ms. Hageboeck and TRITON staff, with the scorers from the previous work order (instrument development) training the new personnel. Each went through a "testing" period to ascertain interrater reliability and general understanding of the methodology. Training consisted of a detailed explanation of the scoring instrument, followed by application of the instrument to an evaluation which the trainer had previously scored. This process was repeated several times until all parties felt that they were examining the same aspects

³282 evaluation reports were received, but 16 were not scored due to brevity and/or language restrictions.

of the evaluations. At this point, the trainees were given a workload similar to that accomplished by the trainer and proceeded at their own pace. Evaluations were randomly rescored to verify interrater reliability. This also provided the opportunity to determine if retraining was necessary.

For purposes of quality control in the scoring process, the following steps were taken:

- Scorers were not informed of scoring results.
- Two or three scorers were given the same report to score on a random basis to monitor interrater reliability.
- Reports scoring more than ± 2 standard deviations from the overall mean were rescored to validate their "outlier" status.
- Individual rater's scoring patterns were analyzed for bias.

The results of these quality control efforts are discussed in detail in the body of this project's Final Report.

The scoring instrument was used to test nine internal characteristics of a good evaluation. These characteristics were:

- Identification of project and evaluation objectives;
- Focus on the evaluation users and their needs/questions;
- Appropriateness of data collection procedures;
- Separation of facts from interpretations;
- Appropriateness of data analysis procedures;
- Evaluation report as a well-written, self-contained document;
- Answers provided to the full set of evaluation questions;
- Action implications are clearly stated; and

- Appropriateness of evaluation design.⁴

In addition to being scored for quality and completeness, the analysis was extended to nine external (independent) variables in order to examine scoring trends and tendencies. These nine external variables were:

- Geographic bureau;
- When evaluation occurred in the life of the project;
- AID management unit directly involved;
- Technical code;
- Host country participation;
- Level of logical framework actually evaluated;
- Evaluation cost;
- Total cost of project; and
- Contractor evaluation entity.

Limitations Of The Results

While the results of the scoring effort have yielded meaningful insights into USAID's evaluation efforts, there are inherent limitations in the instrument. The scoring instrument is not intended to serve as a "grading" tool: a report with score of 74 might not be "better" than one with a score of 71. The instrument's purpose is to identify general patterns of performance and relative strengths and weaknesses regarding the evaluation reports.

SUMMARY OF INSTRUMENT DEVELOPMENT

The details of the instrument development are contained in Work Order No. 1. Final Report, Development of a Quality

⁴A complete description and ranking of the internal characteristics is found in Appendix A.

ness Scoring Instrument for USAID Evaluation Reports. A brief review of this effort is presented below.

The first phase concentrated on identifying factors reflected in a "quality" evaluation report. These factors were compiled from evaluation literature, interviews with USAID and non-Agency evaluation specialists, and TRITON staff's perceptions as to key "quality" and "completeness" indicators. Thirty-four persons were identified to participate in the ranking of quality factors and subfactors. Twenty-two were employed by USAID and twelve by external organizations/ agencies. (A detailed list appears in the Final Report for Work Order No. 1.) A compilation of these factors was the basis for developing this scoring system for AID evaluation reports. The experts were sent questionnaires to rank the major quality factors. Following this process and the computation of its results, TRITON developed the instrument's forms and weighting factors.

The key findings of this ranking analysis revealed that:

- In general, there was a large degree of consensus among the comparative groupings of respondents as to the rankings, particularly when the rankings were "clustered;" i.e., factors with scores within 10 points of each other were considered as being nominally equal in ranking.
- A general pattern could be identified whereby 1-3 factors or subfactors were clearly the highest ranked, a similar number clearly the lowest ranked, and the remainder clustered in a mid-range.

Quantitative values for the scoring instrument were weighted based on the percent of the total ranking points accounted for by each factor/subfactor. For example, if a factor accounted for 105 out of 895 ranking points, it was assigned a weight of 11%. These weighted values are described in Appendix A.

A first draft of the scoring instrument was prepared once these factors were ranked. All but two of the subfactors were scored in a similar manner on dimensions of completeness, clarity, and/or appropriateness. The scorer simply rated each factor based on his/her perception of the evaluation report's performance.

Two remaining subfactors of the characteristic "The overall design of the evaluation is appropriate for answering the evaluation questions" required a more in depth approach to assessment. These subfactors dealt with: 1) the measurement procedures used by the evaluation and their validity and appropriateness and 2) the evaluation design's procedures for addressing hypothesized cause-and-effect linkages.

Worksheets and supporting materials were developed to rate these subfactors on such quality dimensions as validity, reliability, consistency, replicability and objectivity.

The scores provided by the various factors were then summarized and normalized based on a pre-defined set of formulae which yielded:

- A score of 0-100 for each subfactor;
- A score of 0-100 for each characteristic (derived by weighting the factor scores, as per the results of the modified Delphi survey);
- An overall score for the evaluation report of 0-100 (derived by weighting the characteristic scores as per the survey).

Testing The First Draft Of The Instrument

In order to test the draft instrument, two members of TRITON's staff applied it to five USAID evaluation reports.

The testing was performed to determine:

- Interrater reliability (i.e., the similarity of the same report when scored by the two reviewers);
- Ease of applying the instrument (and in understanding it);
- Appropriateness of the instrument (i.e., were key items not addressed or non-relevant items included);
- Time to review report and complete instrument.

This test indicated that a useful instrument had been developed, but that further refinement was necessary to reduce application time, minimize differences in interpretation and eliminate any potential learning curve bias.

Meetings were held with USAID and TRITON to address these refinements and several outcomes resulted.

The first outcome involved retesting some of the evaluation reports. A general reduction in absolute scores was observed, and the average difference in scores was reduced from 17.5 points to 5.5 points. The "learning curve bias" appeared to dissipate, with scores showing no pattern based on the sequence of review.

The second outcome was to determine the correlation between rater scores for the entire report and selected subfactors. Keeping in mind that a "perfect" positive correlation between two scores would be +1.0, the report level scores indicated a very high correlation between the two raters, typically +.70 or higher.

The final outcome was to rearrange the characteristics so that the one requiring the most detailed analysis of the evaluation report appeared first. Applications of this revised instrument showed enhanced clarity and ease of application.

Testing The Revised Instrument

The revised instrument was now used to score forty evaluation reports selected by USAID staff. The scorers were the same two people who conducted the first round of tests.

After the extensive iterative process employed to develop quality and completeness factors/criteria, weightings, and the scoring instrument, the instrument development process arrived at a usable and appropriate tool. A review of the last round of scores indicated high rater consistency and inter-rater reliability with a pattern of scores normal-like in distribution and concentrated among values of 30-70.

The next logical step was to apply the revised instrument to a large array of USAID evaluation reports and to conduct appropriate analyses of scoring trends and patterns by the internal and external variables.

USAGE AND UTILITY OF THE INSTRUMENT

The first major application of the scoring instrument was the FY82 Meta-Evaluation effort. The meta-evaluation consisted of scoring and analyzing all of the evaluations received by the Office of Evaluation from September 1981 through September 1982. These totaled 282, of which 16 were discarded due to brevity and language (i.e., some were too short for the effective utilization of the instrument, while others were not in English and would have required too much of the coder's time). The final number of FY82 evaluations read was 266. A random sample of 40 evaluation from both FY80 and FY81 were scored after the FY82 evaluations had been examined, providing a universe of 266 FY82 evaluations and 80 additional evaluation scores.

The instrument itself was composed of several attachments. Attachment One assessed the internal factors previously described, and primarily examined evaluation design and methodology.

Attachment Two recorded the logical framework, which was either explicitly (i.e., actually written out in the evaluation) or implicitly derived. The latter case involved the examination of the content of the evaluation, i.e., what was being evaluated and how did it fit together in a linear sequence. Even where an evaluation explicitly stated the logical framework, the scorers did not apply the instrument to levels which were not also evaluated. An interim evaluation, for example, might state all of its logical framework, but could not evaluate the goal or purpose levels because the project was not yet fully implemented.

Attachment Three provided a content analysis of the evaluation. It looked at each of the levels of the logical framework determining the indicators used at each level and how valid and reliable each of the indicators was.

Attachment Five was perhaps the most scientific of all of the attachments, as it examined the hypotheses which linked each level and the types of experimental methodology. It was developed to identify experimental designs in use. Factors such as maturation, selection, etc., were analyzed to determine their effect on the results of the evaluation.

Attachments 4, 6, 7 were used for scoring. Weighted values were entered into the computer and then coded so that the manipulation of either external (independent) or internal (dependent) variables would be possible.

DATA ANALYSIS PLAN FOR THE STATISTICAL REPORT

In order to assess the validity of the scoring instrument the first step in the analysis plan was to check the internal consistency of the variables. This was done by determining whether the distributions of two internal variables (e.g., Characteristic I and III) had the same direction and magnitude.

Our first step in assessing the internal consistency of the dependent characteristics, showed that these nine characteristics were statistically consistent with one another.⁵ The symmetrical gamma showed a strong positive association between each characteristic and the total score variable, which means that each characteristic contributes positively to the total score.

The average scores (means) and standard deviations for the internal characteristics were calculated to show the reader what the average case looks like and how odd-looking cases are distri-

⁵We used a symmetrical gamma, which is a statistic used as a measure of association for ordinal variables. Ordinal variables are those which can be grouped into categories, but which cannot be measured on a score scale. That is, ordinal variables are not categorical (nominal) variables such as geographic bureau which has distinct, mutually exclusive categories. Nor are they interval variables such as test scores which show that an 82 is higher than an 80 and that each point earned on that test has meaning.

The internal characteristics measured by the scoring instrument are interval-appearing rather than pure interval variables, and therefore we are treating them as ordinal variables.

A set of measures of association, such as the symmetrical gamma, was selected according to established criteria as described in Loether and McTavish's Descriptive and Inferential Statistics: An Introduction. These measures, based on proportional reduction of error ratios, vary between -1 and +1 with 0.0 indicating no association between the two variables under study. (+1 indicates a perfect direct association and -1 a perfect inverse association).

buted around that typical case. Comparison of means and standard deviations is possible for this analysis plan because the "total score" variable is interval-appearing.

The second step in the analysis plan was to conduct a frequency distribution of each external variable for the FY82 evaluations. This was done in order to spot unusual trends or patterns and to provide primary categorization of the data for further analysis.

In addition, we examined the means and the standard deviations of each variable and its values in order to describe the shape of each variables's distribution curve.

The third task was to perform crosstabulations on each of the external variables by the total score variable. Measures of association such as Somers D and lambda were used to summarize and compare the crosstabulation tables.⁶

These statistics show the reader 1) whether or not an association exists, 2) the strength of the association, 3) the direction of that association and 4) the nature of that association.

⁶Somers D is a measure of association used with ordinal data and is most appropriate in distinguishing between independent and dependent variables.

The lambda measure of association statistic is used to study categorical variables, (i.e., evaluations examining agriculture or health or education projects).

The square of the lambda measure (and other measures of association) may be interpreted as proportion of variance explained in the dependent variable by the independent variable. The measures of association are normed measures (i.e., they vary only between -1 and +1 and are not subject to distortions associated with the magnitude of numbers in a distribution).

The measures of association technique was selected because 1) the FY82 evaluation scores were a universe, 2) the scores could not be assumed to be normally distributed, 3) no inferences could be made about the shapes of the distribution of the data in past years, and 4) they can be easily compared across tables.

Initial inspection of the crosstabulations of the external variable and internal characteristics indicated that the categories of measurement were too narrowly defined for any meaningful interpretations to be made.

The fourth task of the analysis plan was to collapse these categories where appropriate, and to generate first order partial crosstabulations for some of the more interesting relationships. An example of a first order partial is a crosstabulation of evaluation cost by total score for each bureau.

Categories were collapsed by broadening their definition (e.g., evaluation time became interim evaluations and all other evaluation timings). This captured the maximum effect of one variable upon another. Since most of the data was ordinal, and we were trying to distinguish between independent and dependent variables, the Somers D statistic was used most often during this stage of analysis.

In conclusion, because the data are a universe, no statistical predictions can be made, and a strictly descriptive approach was used.

PATTERN ANALYSIS OF FY 82 AID EVALUATIONS

In addition to scoring 266 FY 82 evaluations by internal characteristics such as evaluation design and data collection,

TRITON studied the effect of the several external variables on the overall score.

An overview of the distribution of the evaluations by each external variable follows:

- The largest number of evaluations came from the Bureau for Africa with 34.6% of the total. Next in importance was the Bureau for Latin America and the Caribbean (18.4%), followed by the Bureau for the Near East with 15%, and the Bureau for Asia with 11.7% of the total. The remaining 20.3% of the total is accounted for by central bureaus such as Science and Technology and Food and Voluntary Assistance, or by Project Impact Evaluations.
- Country projects comprised 44.66% of the evaluations studied. Other evaluation scopes included country sector, country program, and world-wide projects.
- AID missions in the field produced 64% of the evaluations examined. Next was the AID/Washington central bureaus which originated 18.77% of the total.
- Interim evaluations made up 65.53% of the evaluations studied. The remaining evaluations included 45 (17.50%) final evaluations, 27 (10.23%) ex-post evaluations, and 27 (10.23%) combination evaluations.
- Project cost was distributed in the following manner:
 - LOW (\$0-949K): 20.28%
 - RELATIVELY LOW (\$950K-5,049K): 40.5%
 - MEDIUM TO HIGH (\$5,050K-10,149K): 19.34%
 - HIGH (\$10,150K - 22,149K): 19.81%
- Data on the cost of the evaluation was available for only 107 of the 266 evaluations studied. 40.19% of these evaluations had a low evaluation cost (\$350-\$5,075), 20.56% had medium evaluation cost (\$11,025-\$32,450) and 9.35% had a relatively high evaluation cost (\$35,870 and over).
- As expected, most the evaluations were conducted by AID mission staff as the sole or as one of several evaluation entities. Next in importance were consultants, host government entities, and AID/Washington personnel.

Purpose, Major Results and Conclusions of the Statistical Report

The purpose of the statistical study was to elicit useful results and practical guidance concerning the quality of AID evaluations from FY80 to FY82. We found that:

- Projects which had high evaluation budgets (most noticeable for those with the health technical focus) tended to have higher scores.
- Evaluations conducted by fewer evaluation entities tended to have higher scores. This was probably because they had fewer coordination problems while conducting the evaluation.
- Evaluation reports examining only the lowest levels of the logframe (input and outputs) scored relatively lower than those examining the highest logframe levels (purpose and goal).
- Very few evaluators looked at the needs of the users. If the evaluators were outside consultants, this was probably because the users did not work closely with the outside contractor to design the evaluation plan.
- Major weaknesses in the evaluations studied concerned data collection and data analysis procedures. Problems included secondary data which was not verified or insufficient baseline data. Specific examples can be found in the findings report.

Recommendations Based on the Statistical Report

Since our knowledge of what comprises a "good" evaluation is not complete (i.e., more evaluations need to be studied in order to predict how AID can specifically improve the quality of its evaluations), we propose that AID focus on low scoring evaluations and then use the following recommendations to improve those scores. These recommendations concentrate on characteristics which fell below the average, such as appropriate evaluation

design, data collection procedures, and data analysis procedures.⁷

Evaluation Design Recommendations

- Evaluators and AID mission personnel should more closely collaborate on the purpose and goals of the evaluation. This includes:
 - Defining the population or universe to be studied;
 - Determining who should be contacted or what villages should be studied;
 - Determining whether interviews, records, or other techniques such as rapid rural appraisal are appropriate for the collection of evaluation data;
 - Designing a timeframe and level of effort; and
 - Making sure that reporting requirements are clear.

Data Collection Recommendations

- Close examination of the evaluation design plan by both the evaluators and the evaluation users must be done in order to determine whether evaluation data is up-to-date and correct.
- Proxy or other innovative measures could be used to verify secondary data whenever appropriate and easily applicable.

Data Analysis Recommendations

- Very few evaluations considered the users of the evaluations. The evaluation should be written with the user in mind, and answers to the users' questions should be presented. Concise evaluations with illustrative analysis

⁷More specific recommendations can be found in the Office of Evaluation's Manager's Guide to Data Collection.

(i.e., pie graphs). or narrative examples are more useful than crosstabulation tables.

- The implications of missing data should be examined carefully.
- In many cases, all the data necessary for analysis will not be available. An analysis plan should identify procedures to be followed if some of the data is not usable.

Overall, the highest scoring evaluation reports avoid global generalities and come up with specific findings and recommendations. They also cite evidence to support favorable or unfavorable opinions.

FINDINGS REPORT & PURPOSE OF STUDY

The findings report and compendium were intended to provide a more qualitative component in the analysis of FY82 evaluations. Findings were defined as terse, pithy statements generated from the evaluation's major recommendations and conclusions. The coders had a short training session and were thereafter required to identify and record the findings. This enabled the coders to record what the evaluation had found, rather than only rating how well the evaluation conformed to the internal characteristics presented in Appendix A.

The findings were put into inductively derived categories based on their content and frequency. These categories were then defined and assigned a value to assist in analysis. The findings are listed below within generic categories encompassing the entire project cycle.

Design

- Overly ambitious objectives
- Conflicting objectives
- Failed assumptions
- Missing inputs and outputs
- Scheduling and budget
- Recommendations and planned changes

Implementation

Contractors:

- Problems finding U.S. contractors and personnel
- Problems finding host country contractors and personnel
- Commitment and performance of U.S. contractors and personnel
- Commitment and performance of host country contractors, government and personnel
- Commitment and performance of both U.S. and host country contractors and personnel

AID-Related Mechanisms:

- AID reporting requirements
- Contracting and funding procedures
- Coordination between AID and host countries
- Procurement of commodities
- Delay litanies
- Coordination between AID and contractor

Institution-Building

Progress:

- At the central level
- With decentralization.
- At the community level
- With training

Problems with:

- Self-sufficiency and recurring costs
- Strategies and structures
- Training

Data Management

- Collection and analysis
- Plans developed via that analysis
- Disseminating information

Impact

- Production impact
- Economic impact
- Social impact
- Spread/limitation effects.

The findings were categorized and subsequently inputted into an Apple computer together with identification numbers for project, bureau, technical code, and coder. This provided a listing of the findings. This numerical information was also inputted into the AID computer, using the SAS package. These two data bases were then used to analyze the distribution of findings.

The findings report examined what effect several key external variables had on the distribution of the qualitative statements, presented the analyses of that distribution, and included examples of the findings themselves. The external variables, such as bureau and coders, addressed the following series of questions:

- Is there one major factor which determines the types of findings that are written? If so, what is it?
- Do the findings of one type of project (i.e., road construction)⁸ form a consistent pattern world-wide, or do they vary from bureau to bureau?
- If there are different bureau distributions for the findings, what might cause the differences?
- Is there a discernible coder bias? If so, where is it most prevalent?
- What correlations are possible between the scores of the evaluations and the numbers/content of the findings?

⁸Type of project is defined as the technical scope of the project and is listed as 'technical codes' on the AID computer.

The conclusions of the report are as follows:

- The most dominant factor appears to be the type of project. For example, a maternal/child health project anywhere in the world would generate similar findings regarding the management and commodities.
- Evaluations initiated by the Central Bureaus produced types of conclusions which were specific to a particular focus of that Office. For example Science and Technology projects contain an output related to data management, and that office provided the most findings in the data categories.
- The types of findings do not vary from bureau to bureau, except where there is a particular bureau-wide focus which permeates even a blue-printed project, such as the Latin-American Bureau's concentration on institution-building.
- The evaluation scores showed no statistically significant coder bias, but there are tendencies in the findings for a coder to concentrate on one aspect of the projects.
- This application of the instrument cannot show a relationship between evaluation/project cost and the number/quantity of findings because of a misconception on the part of the coders as to what constituted a finding.

The findings covered a wide variety of topics within the categorization scheme. Since the scheme did not differentiate between positive and negative findings, no judgements can be made as to what types of findings are found in "good" or "bad" evaluations.

Four categories formed the majority of the findings. These were failed assumptions, recommendations/planned changes and the commitment and performance of host country contractors, government and personnel, as well as U.S. contractors.

Failed assumptions can be broken down into two basic categories: external variables, those things over which a project had no control (i.e., political or economic instability), and

internal, those things over which a project should exercise control, (i.e., country-specific cultural constraints). Some examples include:

- Only a fraction of U.S.-based and third country training occurred and no in-country non-formal training was initiated due to faulty project design which did not account for the difficulty in releasing institution staff for even relatively short training periods. Project No. 2790028.
- Certain key project assumptions proved faulty including favorable environmental conditions, favorable economic conditions at the national level, high adoption of improved agricultural methods by farmers and favorable crop prices. Project No. 4930280.
- Project design assumed that women would receive loans for income-generating activities: women did not receive loans because project managers did not focus attention on this component of project implementation. Project No. 6860212.

Recommendations/planned changes is more a prescriptive than a descriptive category. It provides a way for the project management to redirect the project, either by emphasizing an aspect which was particularly successful, or by redesigning less successful portions. Some examples include:

- Establishment of a design unit at the state level would expedite the project and approval process for the medium irrigation subprojects. Project No. 3880467.
- Host government community development objectives would be better served by a selective rather than blanket coverage approach to extending community development activities. Project No. 6310017.
- More care has to be taken to define the beneficiary population more precisely: project will have to determine in what type of lending the organization has most need of and has a comparative advantage in, especially in terms of the service it can offer, given its limited resources. Project No. 9380131.

- Establishment of a design unit at the state level would expedite the project and approval process for the medium irrigation subprojects. Project No. 3880467.
- Host government community development objectives would be better served by a selective rather than blanket coverage approach to extending community development activities. Project No. 6310017.
- More care has to be taken to define the beneficiary population more precisely: a project will have to determine in what type of lending the organization has most need of and has a comparative advantage in, especially in terms of the service it can offer, given its limited resources. Project No. 9380131.

The other two categories which occurred most often were the commitment and performance of contractors. Most of these had a negative aspect, that is, the contractor had not properly executed his tasks, but some of them were positive. Some examples of both kinds follow:

U.S. Contractors

- The systems approach led to the design of a complicated program which was difficult to manage. Project No. 5220265.
- Although the PVO's demonstrate unusual cultural sensitivity, they do need to systematize their training programs, as well as their evaluation techniques.

Host Country Contractors

- Project has been successful in mobilizing host country scientists to participate in project activities, including training ones. Project No. 2630041.
- Furthermore, bureaucratic conflict has created an atmosphere in which much research done at the center is rejected out of hand by the central Ministry of Agriculture and often has to be redone to be acceptable. Project No. 7005034.

All of the findings are presented in the forthcoming findings compendium. Each evaluation's findings are preceded by a DIU project design abstract, and the compendium serves as a catalogue for the FY 82 evaluations. It is indexed by bureau, technical code and findings categories, which makes it a ready reference tool for managers in charge of the design or evaluation of AID projects.

APPENDIX A

INTERNAL QUALITY/COMPLETENESS CHARACTERISTICS AND SUBFACTORS OF EVALUATION REPORTS

CHARACTERISTIC I: The overall design of the evaluation is appropriate for answering the evaluation questions. (11.00%)*

SUBFACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC:

1. The units of analysis are appropriate given the evaluation questions. (1.43%)
2. As appropriate, given the stage of the evaluation, the evaluation design contains procedures for measuring project efficiency, effectiveness (e.g., the provision of goods/services to intended beneficiaries of the goods/services provided by a project or program). All measurement approaches in the design are conceptually valid. To the degree appropriate, the measurement approaches consider such factors as the timeliness with which goods/services are delivered, the duration of services, etc. (2.75%)
3. As appropriate, given the stage of the evaluation, the evaluation design contains procedures for examining the strength and validity of hypothesized cause and effect linkages. These procedures are appropriate for making determinations concerning the probability that a particular cause or means (provided by the project or program) explains the effects/ outcomes/impacts (of the project or program). The procedures for examining cause and effect relationships are strong enough to give reasonable assurance that major "rival" explanations will be considered and eliminated before claims of a relationship between a project or program and a set of effects/outcomes/impacts are made. (1.65%)
4. Assumptions made by the design are clearly and completely stated. (1.65%)
5. If the design is adapted from another evaluation or research study, it is customized for the situation in which it is to be used, if required. (1.10%)
6. The evaluation design is fully and clearly described by the evaluation report. (1.32%)
7. The design includes procedures for recording any changes in the methodology made during the course of the evaluation and where such changes occur, the evaluation report discusses them. (1.10%)

*Indicates what percent of the maximum quality/completeness score possible (100) is attributable to this characteristic/subfactor, based on priority weightings of review panel.

APPENDIX A (Cont'd)

CHARACTERISTIC II: The evaluation clearly and completely identifies the objectives of the project or program which is being evaluated as well as the evaluation objectives and questions. (15.00%)

SUBFACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC:

1. Project or program objectives are clearly and completely stated. (6.45%)
 2. The objectives of the evaluation are clearly and completely stated; priorities among objectives and reasons for some are clear. (4.80%)
 3. The evaluation questions are clearly and completely stated; priorities among questions are clear. (3.75%)
-

CHARACTERISTIC III: The evaluation focuses on the evaluation users and their needs/questions. (15.00%)

SUBFACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC:

1. Evaluation clients/users are clearly and completely identified. (5.85%)
 2. User needs/expectations are clearly and completely identified. (5.85%)
 3. Areas of "public interest"/broad concern covered by the evaluation are clearly identified. (3.30%)
-

APPENDIX A (Cont'd)

CHARACTERISTIC IV: The data collection procedures/secondary data are appropriate and adequate, not excessive or inadequate. (9.00%)

SUBFACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC:

1. Instruments/approaches for collecting data are valid and reliable; validity and reliability of any secondary data is checked and found acceptable. (1.89%)
 2. Sources of error/biases in the instruments or data collection procedures are described as fully as possible. (1.71%)
 3. Where there is a need to generalize from the data to a larger population, either sampling procedures which allow such generalization are properly used or the limits on generalizing from the data are fully stated. (1.71%)
 4. Neither too much or too little data is secured. (1.35%)
 5. Where cross-cultural sensitivity, language, etc. are potential issues, they are properly handled (e.g. local data collectors used, female data collectors, etc.) (.90%)
 6. Where data must be collected and it is important to do this in a non-disruptive manner, the data collection procedures are as non-disruptive as possible. (.54%)
 7. Instruments used to collect raw data, such as questionnaires, are included as exhibits to evaluation reports. (.90%)
-

CHARACTERISTIC V: Findings, conclusions and recommendations are presented in a way that clearly separates facts from interpretations. (11.00%)

SUBFACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC:

1. Facts are separated from interpretations. (1.76%)
2. Alternative interpretations are discussed and the reason for selecting a specific interpretation or conclusion is made clear. (1.76%)
3. Conclusions are separated from recommendations. (1.10%)
4. Alternative recommendations are discussed and the reason for selecting a specific recommendation is made clear. (1.76%)

APPENDIX A (Cont'd)

5. The study findings, conclusions and recommendations are well organized and presented in a fashion that is understandable to a busy reader/decision-maker who may not be familiar with how studies are conducted. (1.76%)
6. The material on findings, conclusions and recommendations is presented clearly and objectively, in the sense that it neither "hides" data nor makes assertions without adequate facts. (1.76%)
7. The evaluators come a "bottom line" where the evaluation questions and purposes require that some firm conclusions be drawn in the course of the evaluation; i.e., did the project succeed in achieving its objectives or not? (.76%)

CHARACTERISTIC VI: The data analysis procedures are appropriate and adequate. (11.00%)

SUBFACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC:

1. The analysis procedures are clearly presented, match the purposes of the evaluation and fit the evaluation questions and data collected to answer those questions. (2.53%)
2. The analysis procedures are appropriate; they are neither weak nor excessive. (1.43%)
3. Where appropriate, the confidence level of findings is given; e.g., statistical significances of comparisons of quantitative data on two groups, descriptive statements about the confidence that should be placed in answers arrived at through non-quantitative data and analysis. (1.43%)
4. Both quantitative and qualitative data are analyzed if both were secured. (1.43%)
5. Where possible, the evaluation examines how realistic were the project's original estimates of cost, economic return, etc., as well as data on project/program effectiveness and impact. (1.76%)
6. The strength and weaknesses of the data analysis aspects of the evaluation are clearly and completely stated. (1.76%)
7. Where appropriate, the raw data from the study are included, or their availability made known, should it be necessary/appropriate to re-analyze all or part of the study data. (.66%)

APPENDIX A (Cont'd)

CHARACTERISTIC VII: The evaluation report is a well-written, self-contained document. (11.00%)

CHARACTERISTIC VIII: The evaluation produces the types of information it was expected to produce; i.e., in so far as possible, the full set of evaluation questions are answered. (11.00%)

CHARACTERISTIC IX: Action implications of the evaluation are clearly stated and are annotated to indicate who or what unit should act. (9.00%)