

FINAL REPORT:

ANALYSIS OF THE DISTRIBUTION OF QUALITY/COMPLETENESS SCORES  
OF FY83 AID EVALUATION REPORTS

Contract No. OTR-0000-C-00-3482-00

Conducted for PPC/E/PES  
of the Agency for International Development

March 1985

March 4, 1985

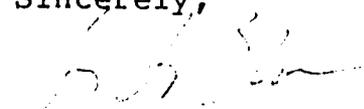
Ms. Nena Vreeland  
Technical Officer  
PPC/CDIE  
Room 3659, New State  
Agency for International  
Development  
Washington, DC 20523

Dear Ms. Vreeland:

TRITON Corporation is pleased to submit the report entitled, "Analysis of the Distribution of Quality/Completeness Scores of FY83 AID Evaluation Reports." The analyses presented in this report pertain to the distribution of scores of the FY83 AID evaluations as generated by the application of TRITON's scoring instrument to measure characteristics of quality and completeness. A series of thirteen major questions relating to these scores to various external attributes of the evaluation reports are answered herein.

Thank you again for your cooperation and assistance in developing this report.

Sincerely,



Sonny S. Bloom  
Vice President

Enclosure

SSB:691

**TRITON**

1255 Twenty-third Street, N.W. Suite 275, Washington, D.C. 20037 • (202) 296-9610

TABLE OF CONTENTS

<u>Section</u>	<u>Page</u>
I. Conclusions from Data	
A. Overview.....	I-1
B. Differentiation of Bureau Scores.....	I-2
C. Characteristic Scores of the Bureaus.....	I-9
D. Cost/Score Correlation.....	I-12
E. Correlations Between Types of Evaluators and Scores.....	I-14
F. Host Country Participation and Managing Unit.....	I-19
G. Host Country Participation and Contractors.....	I-20
H. Evaluation Entities and Scores.....	I-21
I. Technical Foci and Scores.....	I-24
J. Time Taken to Complete Evaluation Versus Scores.....	I-25
K. Logframe Levels and Evaluation Time.....	I-27
L. Levels Related to Bureau, Technical Code and Scorer.....	I-29
M. Resource Reallocation and Scores.....	I-31
N. Time Series Analysis.....	I-32
II. Recommendations	
A. General Recommendations.....	II-1
III. Methodology	
A. Development of Project.....	III-1
B. Description of Instrument.....	III-2
C. Application of Instrument.....	III-4
D. Scoring Process.....	III-4
E. Data Procedures.....	III-4
F. Statistical Procedures.....	III-5
G. Analysis Plan.....	III-6

## I. CONCLUSIONS FROM DATA

### A. OVERVIEW

As discussed in detail in Chapter III, "Methodology," the analyses presented in this report and in this chapter, in particular, represent the quantitative results of scoring all FY1983 USAID Evaluation Reports on various factors and characteristics regarding "quality" and "completeness," by applying a "metaevaluation" instrument developed by TRITON for AID. (A slightly different version of this instrument had previously been utilized by TRITON to score all USAID FY82 Evaluation Reports, with analyses similar to those contained herein reported to AID). The scoring process yields total values of 0-100 for each report scored, as well as the individual factors and characteristics evaluated within each report: conceptually, the higher the score, the higher the "quality" and "completeness" of the report. The "Quality Score" for each evaluation was calculated using a subset of three of the nine characteristics by which the reports were scored. These three factors were deemed to most directly address the "quality" of data collection and analysis and evaluation design.

The analyses presented in this chapter attempt to address a series of questions about the external attributes of the reports (e.g., bureau associated with a given report) versus their internal characteristics of quality and completeness. Briefly, the questions addressed include:

1. How do the report scores differ based on bureau affiliation?
2. What are the characteristic profiles of the reports based on bureau affiliation?
3. What is the correlation between the cost of an evaluation and its quality/completeness score?

4. What is the relationship between evaluation contractor type(s) and report scores?
5. What is the extent of host country participation in the evaluations' conduct, based on the "managing unit" (sponsoring bureau) for the reports, as well as by mission/bureau?
6. What is the extent of host country participation in the evaluations' conduct based on the contractor type involved in the reports?
7. What is the correlation between report scores and the number of contractors involved in a given report?
8. What is the relationship of report scores versus "technical code" addressed by the evaluation?
9. What is the relationship of report scores to the length of time taken to perform the evaluation?
10. What is the relationship between the number of "logframe levels" examined by a report versus the point in time in the project's life the evaluation occurred?
11. What is the relationship between the number of logframe levels examined by reports versus technical code and bureau?
12. What is the relationship between evaluation reports that address the need for project "resource reallocation," based on the evaluations's findings, versus report scores, bureau, and technical code?
13. What are the differences/similarities between the scoring patterns of the FY83 evaluation reports versus those of the FY82 reports?

The remainder of this chapter presents tables and related hypotheses and conclusions to address each of the above questions individually.

#### B. DIFFERENTIATION OF BUREAU SCORES

Evaluation scores showed some variation according to the initiating bureau. The overall mean total score was 53.8, with a

standard deviation of 15.6, and the overall mean quality score was 12.5, with a standard deviation of 4.9. By bureau, arranged in order from lowest to highest score, the mean total scores were as follows:

EXHIBIT 1: TOTAL SCORE STATISTICS FOR EACH BUREAU			
<u>BUREAU</u>	NUMBER OF CASES	MEAN TOTAL SCORE	STANDARD DEVIATION
Latin America and Caribbean (R)	38	50.9	17.7
Near East (R)	33	51.1	13.0
Science and Technology (C)	14	52.1	21.3
Africa (R)	103	52.5	14.1
Asia (R)	59	56.9	16.4
Food and Voluntary Assistance (C)	15	59.1	15.5
P.P.C. Impact (C)	8	65.9	10.7
Note: (R) = Regional Bureau, (C) = Central Office Bureau			

Exhibit 1 depicts the distribution of the bureaus across total scores in deciles, while Exhibit 2 segments total scores into low, medium and high groupings.

EXHIBIT 2: BUREAU BY TOTAL SCORE IN DECILES

BUREAU	MEAN TOTAL SCORE (FREQUENCY AND PERCENT WITHIN BUREAU)									
	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
NEAR EAST (R)	0 0.0%	1 3.0	6 24.2	8 24.2	11 33.3	5 15.2	2 6.1	0 0	0 0	33 100%
ASIA (R)	1 1.7	1 1.7	6 10.2	11 18.6	16 27.1	11 18.6	8 13.6	3 5.1	2 3.4	59 100
LAC (R)	1 2.6	4 10.5	7 18.4	6 15.8	5 13.2	10 26.3	3 7.9	2 5.3	0 0	38 100
AFRICA (R)	2 1.9	3 2.9	9 8.7	31 30.1	29 28.2	20 19.4	7 6.8	1 1.0	1 1.0	103 100
IMPACT (C)	0 0	0 0	0 0	0 0	3 37.5	3 37.5	1 12.5	1 12.5	0 0	8 100
SCITECH (C)	2 14.3	0 0	1 7.1	3 21.4	2 14.3	4 28.6	1 7.1	1 7.1	0 0	14 100
FVA (C)	0 0	0 0	2 13.3	3 20.0	2 13.3	3 20.0	4 26.7	1 6.7	0 0	15 100
TOTAL FREQUENCY	6	9	31	62	68	56	26	9	3	270
OVERALL PERCENT	2.2	3.3	11.5	23.0	25.2	20.7	9.6	3.3	1.1	100

EXHIBIT 3: BUREAU BY PERCENTAGE OF EVALUATION REPORTS  
IN LOW, MEDIUM AND HIGH SCORE CATEGORIES

<u>BUREAU</u>	<u>MEAN TOTAL SCORE</u>			<u>TOTAL</u>
	<u>LOW (10-40)</u>	<u>MEDIUM (40-70)</u>	<u>HIGH (70-100)</u>	
NEAR EAST (R)	21.2%	78.2%	6.1%	100.0%
ASIA (R)	13.6	54.5	31.9	100.0
LAC (R)	31.6	55.1	13.3	100.0
AFRICA (R)	12.6	78.6	8.8	100.0
IMPACT (C)	0.0	75.0	25.0	100.0
SCITECH (C)	21.4	64.3	14.3	100.0
FVA (C)	13.3	53.4	33.3	100.0

The Latin American and Caribbean (LAC), Near East (NE), Science and Technology (SciTech), and Africa bureaus' average scores lie within a range of 1.6 points, Africa's being the highest. The Asia, Food and Voluntary Assistance (FVA) and Impact evaluations average 4.4, 6.6, and 9 points higher, respectively, than the Africa bureau's mean. P.P.C. Impact evaluations show a mean total score significantly higher than all the rest of the bureau groupings. The Asia bureau's mean total score of 56.9 is particularly noteworthy, as it is higher than any of the other regional (R) bureaus and is 3.1 points above the overall mean. An analysis of the total score statistics by individual bureau follows.

1. Near East

The mean total score (51.1) for the Near East bureau is below the overall mean (53.8). The distribution of mean total scores

1

for this bureau forms a bell curve with the largest percentage (33.3%) of reports falling within the 40 to 50 scoring range, as it does in most other bureaus. Of all NE reports scored, 72.7% scored in the medium (40 to 70) range. No evaluations in the NE bureau scored extremely high (90-100 score) or low (10-20 score).

## 2. Asia

The Asia bureau scored significantly higher than the other regional bureaus, as mentioned earlier. It is important to note that 31.9% of the Asia evaluations scored in the "high" (70 to 100) range. In comparison, only 25% of the Impact evaluations, which had the highest overall mean of any bureau grouping, scored between 70 and 100, while 3.4% of the Asia evaluations scored very high (90 to 100), which is more than in any other bureau. It appears that the Asia bureau does comparatively well on evaluations, based on a number of measures, particularly in comparison to the other regional bureaus.

## 3. Latin America and the Caribbean

The LAC bureau had a mean total score (50.9) that is lower than the overall average. This bureau had the highest concentration (26.3%) of its total scores in the 60 to 70 range. While most other bureaus have their highest concentration of report scores in the 50 to 60 range, the LAC evaluation report cannot be judged in general based on this relatively positive outcome alone, since 31.6% of its evaluations scored in the "low" (10 to 40) range, more than any other bureau. The next highest concentration of low scores was 21.4%, found in the SciTech Bureau. This demonstrates that the percentage of low-scoring evaluations in LAC is much greater than in any other bureau. One LAC evaluation scored very low (10 to 20) and none scored extremely high (90 to 100). The remainder of the LAC report scores are relatively spread out in their distribution, as compared to the other bureaus. This

suggests that there may be comparatively less standardization or uniform "quality control" in the bureau as regards to evaluation reports.

#### 4. Africa

The Africa bureau distribution of scores most closely follows the overall pattern in total scores. This is probably due to the fact that the largest percentage (38.1%) of all evaluations scored were from the Africa bureau. The absolute size of this sample makes it more likely to correlate strongly with the aggregate patterns. The Africa bureau evaluations on the whole, however, scored slightly lower than the overall mean (52.5 versus 53.8).

#### 5. Impact

Impact evaluations had the highest mean total score (65.9) of all the bureaus by a large margin. The lowest score for an Impact evaluation was 54.1, which is still higher than the overall average of 53.8. No Impact evaluations had total scores classified as "low" (10 to 40). On the other extreme, none scored extremely high (90 to 100), as might have been expected. Two out of eight scored high (70 to 100), and the remaining 75% scored between 50 and 70.

Impact evaluations, by their very nature, are able to examine a large number of "E" levels; i.e., input-output-purpose-goal levels in a project, based on its logical framework. "Impact" explicitly means going beyond outputs to examine results in terms of purposes and goals. It was found in TRITON's metaevaluation project addressing AID's FY82 reports, which utilized statistical methods to analyze correlation, that there was a high correlation between total score and number of levels examined. Since Impact evaluations examine, on the average, more project levels than do other types of evaluations, they would be expected to score higher

overall. Taking this into consideration, it is possible that the Impact evaluation reports' quality could have actually been closer to the rest of the evaluation reports, holding the variable of "number of E-levels" constant.

Impact evaluation scores cluster around the 50-70 point range and taper off in the 80-90 range. This concentration may reflect the standardization of the Impact evaluation procedures and format.

#### 6. Science & Technology

Science & Technology scored lower than the other Central bureaus on the mean total score. The largest portion (28.6%) of this bureau's reports scored in the 60-70 decile. What is of particular note is that 14.3% scored in the 10-20 decile. While this represents only two actual evaluations, the only other bureau which had two extremely low report scores was Africa, for which two evaluations represent only 1.9% of that bureau's total reports. This bureau also had no "extremely high" evaluation scores. The data indicate that the bureau may need to improve its control mechanism to ensure that evaluations meet minimum standards of quality and completeness.

#### 7. Food and Voluntary Assistance

This bureau's mean total score (59.1) is second only to that of Impact evaluations. It is important to note that the highest concentration (26.7%) of the FVA evaluations scored in the 70 to 80 decile. No other bureau has its largest concentration of total scores in such a high range. There are no extremely high (90 to 100) scores, but a full one-third of the FVA scores are in the high (70 to 100) range. No other bureau has that high a percentage of its scores in the high range. FVA, as compared to S&T, for example, seems to be able to maintain a relatively high minimum standard for its evaluations.

C. CHARACTERISTIC SCORES OF THE BUREAUS

The nine individual characteristics utilized in the metaevaluation instrument to assess evaluation report quality and completeness are listed below, in ascending order of their mean scores, rounded off to whole numbers. The maximum score for each characteristic is 11.1.

EXHIBIT 4: SCORE STATISTICS ON INDIVIDUAL CHARACTERISTICS				
			<u>MEAN</u> <u>SCORE</u>	<u>STANDARD</u> <u>DEVIATION</u>
Characteristic No:	4	Data Collection	3.50	2.2
	1	Evaluation design	4.20	1.6
	6	Data analysis	4.80	1.8
	9	Action implications	5.60	1.9
	7	Writing, completeness	5.90	2.0
	5	Separation of fact from interpretation	6.00	1.7
	8	Production of intended specifics	6.40	1.9
	3	Focus on user needs	8.40	3.0
	2	Identification of project objectives and evaluation questions and objectives	9.10	2.9

This information indicates that for the most part, the evaluations identified "project objectives" and "evaluation questions and objectives" as well as, to a lesser extent, produced what they had intended. However, the "appropriateness" of the evaluation design was often questionable, and the "data collection and analysis" was rated particularly low. Each of these characteristics is discussed below.

### Characteristic 1: Evaluation design

Impact evaluations scored particularly well on this characteristic. This is not surprising, as it should be clear from the outset of such evaluations what they are expected to accomplish -- assessment of impacts. Therefore, those who design these evaluations should have specific goals in mind and planned accordingly. In contrast, the goals of interim evaluations, which make up the bulk of most scored, are not always as clear, so evaluation design may be more problematical.

### Characteristic 2: Identification of project objectives and evaluation questions and objectives

Again, the bureau which scored significantly higher than the rest on Characteristic 2 was Impact. The fact that Impact evaluations are a relatively "new" and "special" type of evaluation may account for many of the differences observed in the scores for Impact reports on Characteristics 1 and 2 in relation to the other bureaus. On the whole, however, most bureaus performed fairly well on Characteristic 2 in comparison with the other characteristics. No single bureau scored particularly low on this characteristic.

### Characteristic 3: Focus on user needs

The bureaus which scored considerably higher than the others on this characteristic were Impact and FVA. Unfortunately, it may not be appropriate to examine the other bureaus-- especially the regional bureaus -- against these two bureaus, as their evaluations and users are quite different in nature.

Characteristic 4: Data collection procedures

Impact evaluations once again scored relatively higher on Characteristic 4.

Characteristic 5: Clear separation of fact from interpretation

No bureau scored particularly high or low on this characteristic.

Characteristic 6: Data analysis

This characteristic was not scored particularly high or low on any one evaluation or bureau grouping of evaluations.

Characteristic 7: Writing, completeness

Like Characteristic 6, this characteristic, did not receive any significantly high or low scores.

Characteristic 8: Answers to evaluation questions  
(Production of intended specifics)

The Asia bureau scored highest in this category by a significant margin.

Characteristic 9: Clarity of the action implications

Although no bureau scored particularly high or low on this characteristic, it is interesting to note converse outcomes regarding two of the bureaus. The Africa bureau, which scored below average on all other characteristics, scored significantly above average on Characteristic 9. This above-average performance theoretically indicates that the Africa bureau's reports better

clarify what the evaluators have found, thereby making it easier for their evaluation users to employ the reports' results. (For example, while not within the scope of this study, it would be interesting to examine whether the Africa bureau is acting on its evaluations more than the other bureaus are.)

Conversely, Impact evaluations, which scored well above average on Characteristics 1 through 8, scored below average on Characteristic 9. This may be due to the fact that action implications in most evaluations are directed toward the missions and the project implementation organizational units, in terms of what actions should be taken on the projects; this is not the case with Impact evaluations. Action implications can and should, however, also come out of expost facto evaluations such as Impact evaluations. The possible issues to be addressed by future Impact reports, given the relatively low scores of Impact evaluations on this characteristic, include: 1) whether such reports should be recommending actions to be taken in similar ongoing projects; and (2) the potential replication of the project(s) under evaluation.

#### D. COST/SCORE CORRELATION

Figures on evaluation cost and total scores were available for 92 of the evaluations scored. With the assumption that this is a representative group, the findings do not support the hypothesis that more money spent on an evaluation yields a better evaluation. The following tables present the evidence.

EXHIBIT 5: UNIVARIATE STATISTICS ON EVALUATION COSTS					
<u>COST OF EVALUATION</u>	<u>NUMBER OF CASES</u>	<u>MEAN TOTAL SCORE</u>	<u>STANDARD DEVIATION</u>	<u>MEAN QUALITY SCORE</u>	<u>STAND DEV.</u>
\$0 to \$9999	40	55.9	13.2	12.6	4.2
\$10,000 to \$19,000	27	53.6	14.9	13.0	5.1
\$20,000 to \$44,999	14	55.6	16.4	12.6	4.5
\$45,000 to \$200,000	11	54.3	12.3	13.3	3.7

EXHIBIT 6: COST OF EVALUATION BY TOTAL SCORE										
<u>COST OF EVALUATION</u>	<u>MEAN TOTAL SCORE (for 92 cases)</u>									
	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
\$0 to \$9999	0%	2.5	10.0	17.5	40.0	15.0	10.0	5.0	0	100%
\$10,000 to \$19,000	0	3.7	11.1	18.5	37.0	14.8	3.7	11.1	0	100%
\$20,000 to \$44,999	0	7.1	7.1	21.4	28.6	14.3	14.3	7.1	0	100%
\$45,000 to \$200,000	0	0	0	54.6	18.2	18.2	9.1	0	0	100%

An intervening variable may be that certain types of projects cost relatively more to evaluate. This would help to explain the lack of association between evaluation cost and scores; however, such an analysis is beyond the scope of this study.

15

E. CORRELATIONS BETWEEN TYPES OF EVALUATORS AND SCORE

The evaluation reports were analyzed according to the types of evaluators preparing the reports. Five separate categories were defined: 1) AID and non-AID evaluators; 2) mission and non-mission; 3) consultants and non-consultants; 4) university and non-university; and 5) implementor-evaluators and external evaluators.

1. AID vs. Non-AID Evaluations

Of those which could be delineated by contractor type, 72 evaluations were conducted by AID personnel and 167 by non-AID individuals or teams. Evaluations conducted by non-AID personnel scored slightly higher, on the average, than those conducted by AID personnel (55.2 versus 50.4). The significance of this difference is diminished, however, by looking at the distribution of scores for the two groups, which are quite similar. Both groups have the highest concentration of total scores in the 50 to 60 decile, as does the group of all evaluations. These data suggest that the difference in terms of evaluation quality between AID and non-AID evaluators is minimal.

EXHIBIT 7: TOTAL SCORE STATISTICS FOR AID/NON-AID EVALUATORS			
<u>TYPE OF EVALUATOR(S)</u>	<u>NO. OF CASES</u>	<u>MEAN TOTAL SCORE</u>	<u>STANDARD DEVIATION</u>
AID PERSONNEL	72	50.4	14.5
NON-AID	168	55.2	15.9

EXHIBIT 8: QUALITY SCORE STATISTICS FOR AID/NON-AID EVALUATORS			
<u>TYPE OF EVALUATOR(S)</u>	<u>NO. OF CASES</u>	<u>MEAN QUALITY SCORE</u>	<u>STANDARD DEVIATION</u>
AID PERSONNEL	44	11.1	4.8
NON-AID	73	14.1	4.9

EXHIBIT 9: AID/NON-AID EVALUATORS BY TOTAL SCORES											
<u>TYPE OF EVALUATOR(S)</u>	<u>TOTAL SCORE</u>										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
AID	0%	2.8	4.2	15.3	22.2	29.2	19.4	6.9	0	0	100%
NON-AID	0.6	2.4	3.0	8.3	23.2	25.6	19.6	10.7	5.4	1.2	100%

EXHIBIT 10: AID/NON-AID EVALUATORS BY QUALITY SCORE					
<u>TYPE OF EVALUATOR(S)</u>	<u>QUALITY SCORE</u>				
	0-10	10-20	20-30	30-40	TOTAL
AID	44.4%	52.8%	2.8	0	100%
NON-AID	22.0	70.3	7.1	0.6	100%

17

## 2. Consultant vs. Non-Consultant Evaluations

The data for consultant versus non-consultant evaluators reveal an even smaller difference in evaluation quality. While the mean total score for the consultant group (55.5) is higher than that for non-consultants (53.0), there is evidence that consultant-conducted evaluations range widely in scores from very low to very high. The relatively large standard deviation for the consultant group (17.3) shows that the evaluation scores are widely scattered. This is also shown in the distribution of scores. Note that the the consultant group had 8.7% of its total scores in the 0 to 30 range, while the non-consultant group had 5.2% in the same range.

EXHIBIT 11: UNIVARIATE STATISTICS FOR CONSULTANT/ NON-CONSULTANT EVALUATORS					
TYPE OF EVALUATOR(S)	NO. OF CASES	MEAN TOTAL SCORE	STANDARD DEVIATION	MEAN QUALITY SCORE	STANDARD DEVIATION
CONSULTANT	69	55.5	17.3	13.3	5.8
NON-CONSULTANT	170	53.0	14.8	12.1	4.5

EXHIBIT 12: CONSULTANT/NON-CONSULTANT EVALUATORS BY QUALITY SCORES					
TYPE OF EVALUATOR(S)	QUALITY SCORE				
	0-10	10-20	20-30	30-40	TOTAL
CONSULTANT	23.2%	66.7	8.7	1.4	100%
NON-CONSULTANT	31.0	64.3	4.7	0	100%

EXHIBIT 13: CONSULTANT/NON-CONSULTANT BY TOTAL SCORES											
TYPE OF EVALUA- TOR(S)	<u>TOTAL SCORE</u>										
	0- 10	10- 20	20- 30	30- 40	40- 50	50- 60	60- 70	70- 80	80- 90	90- 100	TOTAL
CONSUL- TANT	0%	2.9	5.8	5.8	23.2	23.2	18.8	11.6	7.3	1.5	100%
NON- CONSUL- TANT	0.6	2.3	2.3	12.3	22.8	28.1	19.9	8.8	2.3	0.6	100%

Regarding the above results, it should be kept in mind that outside evaluators are likely to spend more time than AID personnel in explaining their evaluation design and data collection analysis in order to present it to the appropriate AID personnel.

### 3. University vs. Non-University Evaluations

The evaluation reports were divided between those conducted by university personnel and those which were not. Only ten evaluations were identified as having been conducted by university personnel. While this small number of cases renders generalizations invalid, it is of interest to compare average total scores and quality scores for the two groups, as shown below:

EXHIBIT 14: UNIVERSITY/NON-UNIVERSITY EVALUATOR STATISTICS					
<u>EVALUATORS</u>	<u>NO. OF CASES</u>	<u>MEAN TOTAL SCORE</u>	<u>STANDARD DEVIATION</u>	<u>MEAN QUALITY SCORE</u>	<u>CASES STANDARD DEVIATION</u>
UNIVERSITY	10	48.3	13.0	11.3	4.1
NON-UNIVERSITY	230	54.0	15.7	12.5	5.0

The mean total score for the university group of evaluations is 48.3, and the mean for all other evaluations is 54.0. It is evident that the evaluations conducted by university personnel tended to score slightly lower than those which were not, and that none

of that group scored either very high or very low. The following two tables show the percent distribution of total scores and quality scores for university versus non-university evaluators:

EXHIBIT 15: UNIVERSITY/NON-UNIVERSITY EVALUATORS BY TOTAL SCORES											
EVALUATORS	TOTAL SCORE (for 230 reports)										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
UNIVERSITY (4%)	0	0	0	40	20	20	20	0	0	0	100%
NON-UNIVERSITY (96%)	0.4	2.6	3.5	9.1	23.0	27.0	19.6	10.0	3.9	0.9	100%

EXHIBIT 16: UNIVERSITY/NON-UNIVERSITY EVALUATORS BY QUALITY SCORES					
TYPE OF EVALUATOR(S)	QUALITY SCORE				
	0-10	10-20	20-30	30-40	TOTAL
UNIVERSITY	50.0%	50.0	0	0	100%
NON-UNIVERSITY	27.8	65.7	6.1	0.4	100%

This confirms the conclusion that non-university evaluations tend to score higher than university-conducted ones. It is interesting to note that none of the university category reports scored in the highest or lowest three deciles.

F. HOST COUNTRY PARTICIPATION AND MANAGING UNIT

This study examined the extent of host country participation within the group of those reports which could be so designated. Within this group, more mission-conducted evaluations incorporated host country participation than did non-mission evaluations. The table below gives the percentages found in the FY 83 evaluations:

EXHIBIT 17: MANAGING UNIT BY HOST COUNTRY PARTICIPATION		
<u>EVALUATORS</u>	<u>HOST COUNTRY PARTICIPATION</u>	
	Yes	No
MISSION	50.0%	50.0%
NON-MISSION	26.5%	73.5%

Host country nationals participate in about half of the in-house evaluations; a much lower percentage of the non-mission evaluations include host country participation. It should be noted that host country participation is not clearly indicated on all the evaluation reports, and could not be determined at all for approximately one-eighth of them. Judging by those reports in which it was clearly indicated, it was concluded that evaluations conducted by the missions are more likely to incorporate host country participation than are other evaluations. It is likely to be more difficult for outside evaluators to contact nationals to work on evaluations than it is for mission personnel. Therefore, it would be useful for AID to investigate means of facilitating host

21

country participation in evaluations conducted by outside contractors.

Some bureaus included host country participation in evaluations more than others. The table below presents the percentages of evaluations in each bureau which did and did not use host country nationals, for those reports that the information could be determined.

EXHIBIT 18: BUREAU BY HOST COUNTRY PARTICIPATION		
<u>BUREAU</u>	<u>HOST COUNTRY PARTICIPATION</u>	
	YES	NO
ASIA	53.6%	46.4%
NEAR EAST	78.3%	21.7%
LAC	28.0%	72.0%
AFRICA	36.6%	63.4%
IMPACT	42.9%	57.1%
SCITECH	25.0%	75.0%
FVA	20.0%	80.0%

G. HOST COUNTRY PARTICIPATION AND CONTRACTORS

In contrast to the findings of the preceding section, evaluations in which AID personnel participate incorporate host country participation less than those in which AID personnel do not participate:

22

EXHIBIT 19: AID/NON-AID EVALUATORS BY HOST COUNTRY PARTICIPATION		
<u>EVALUATORS</u>	<u>HOST COUNTRY PARTICIPATION</u>	
	Yes	No
AID PERSONNEL	21.4%	78.6%
NON-AID PERSONNEL	52.6%	47.4%

The Asia bureau stands out as having by far the most host country participation. This may help to explain their relatively high scoring evaluations. The only other bureau which has more evaluations with host country participation than Asia is the Near East. Of the central bureaus, Impact evaluations include the highest percentage of host country participation and FVA has the lowest percentage of all the bureaus.

#### H. EVALUATION ENTITIES AND SCORES

"Evaluation entities" refers to the different types of contractors working on any one report. The following list was used to designate the evaluation entities:

AID Mission staff as implementors/evaluators  
AID Mission staff as external evaluators

US university staff as implementors/evaluators  
US university staff as external evaluators  
Host country university staff as implementors/evaluators  
Host country university staff as external evaluators

US consulting firm/private research organization as  
implementors/evaluators

Host country consulting firm/private research organization as  
implementors/evaluators

23

Host country consulting firm/private research organization as external evaluators

Free lance US consultant as implementor/evaluator

Free lance US consultant as external evaluator

Free lance host country consultant as implementor evaluator

Free lance host country consultant as external evaluator

PASA/RSSA personnel (eg. USDA) as implementors/evaluators

PASA/RSSA personnel as external evaluators

Peace Corps Staff or Volunteers as external evaluators

Int'l. agencies (cf. bilateral or multinational) as external evaluators

Host country gov't staff as external evaluators

U.S. based PVO as implementor/evaluators

U.S. based PVO as external evaluators

Host country PVO as implementor/evaluators

Host country PVO as external evaluators

IMPACT: AID personnel

IMPACT: Other than AID

Up to four evaluation entities were recorded for each evaluation report. The tables below show the mean total and quality scores and the percent distribution of total scores for evaluations with different numbers of entities:

EXHIBIT 20: STATISTICS FOR NUMBER OF EVALUATION ENTITIES				
<u>NUMBER OF ENTITIES</u>	<u>NUMBER OF CASES</u>	<u>MEAN TOTAL SCORE</u>	<u>STANDARD DEVIATION</u>	<u>MEAN QUALITY SCORE</u>
1	103	51.0	5.3	11.3
2	58	55.1	4.7	12.9
3	41	55.1	3.9	14.5
4+	41	45.1	4.5	10.5

With the exception of the fourth group, it appears that a larger number of entities correlates with a higher total score and a higher quality score. We can hypothesize, therefore, that it is valuable to contract with a mixed evaluation team in order to produce a higher quality evaluation.

EXHIBIT 21: NUMBER OF EVALUATION ENTITIES BY TOTAL SCORE										
<u>NUMBER OF ENTITIES</u>	<u>TOTAL SCORE (for 208 cases)</u>									
	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
1	3.9%	5.8	13.6	21.4	27.2	16.5	6.8	3.9	1.0	100%
2	1.7	3.5	10.3	20.7	22.4	24.1	15.5	1.7	0	100%
3	0	0	4.9	17.1	34.2	24.4	12.2	7.3	0	100%
4	16.7	0	0	50.0	16.7	16.7	0	0	0	100%

25

I. TECHNICAL FOCI AND SCORES

The evaluations were grouped according to the technical foci of the projects being evaluated in the reports. The ten categories identified were: Agriculture; Rural non-agriculture; Rural multi-function; Nutrition; Population; Health; Education; Human Resource Development; Infrastructure and Housing; and Other. Evaluations classified as "Other" encompass most of SciTech's evaluations. Unfortunately, most of the categories contained only a small number of cases, while Agriculture accounted for the overwhelming share (57.4%):

EXHIBIT 22: UNIVARIATE STATISTICS FOR TECHNICAL FOCUS GROUPS					
<u>TECHNICAL FOCUS</u>	<u>NUMBER OF CASES</u>	<u>MEAN TOTAL SCORE</u>	<u>STANDARD DEVIATION</u>	<u>MEAN QUALITY SCORE</u>	<u>STANDARD DEVIATION</u>
Agriculture	95	52.5	15.6	12.6	5.3
Rural non-agriculture	12	51.5	13.5	10.0	3.3
Rural multi-function	15	52.2	21.4	12.3	5.9
Nutrition	6	52.6	18.5	11.7	4.3
Population	6	51.1	17.5	12.0	5.7
Health	30	53.0	13.7	12.2	4.6
Education	22	56.7	16.7	13.6	5.0
Human Resource Development	17	53.1	13.3	12.3	3.5
Infrastructure and Housing	18	59.2	15.9	13.5	5.7
Other	38	53.6	14.8	12.1	4.6

2/6

No single category scored exceedingly high or low. The five Education project evaluations had the highest mean total score, 68.3. The lowest total score in the Education category was 48.9, which is only 4.9 points lower than the overall mean. There were, therefore, no low-scoring Education evaluations. Had the number of cases been larger, however, there might have been cases of low scoring Education evaluations. The Education category must, therefore, be viewed with caution due to the very small number of cases.

The lowest scoring group was the thirteen Health project evaluations, with an average total score of 46.9. The FY82 evaluations, addressed in TRITON's previous AID metaevaluation project, revealed a tendency for both the Education and Health categories to score higher than the average, so it is interesting to note that the 1983 Health evaluations demonstrated such a difference.

#### J. TIME TAKEN TO COMPLETE EVALUATION VERSUS SCORE

The time taken to do each evaluation, when available, was recorded by the scorers. The tables below show the mean total scores and quality scores for each of five time categories and the percent distribution of total scores:

EXHIBIT 23: TECHNICAL FOCUS OF PROJECT BY TOTAL SCORE											
TECHNICAL FOCUS	TOTAL SCORE (for 260 cases)										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
Agriculture	1.0	1.0	6.3	12.5	25.0	19.8	20.8	10.4	2.1	1.0	100%
Rural non-agriculture	0	0	8.3	8.3	16.7	50.0	16.7	0	0	0	100%
Rural multi-function	0	13.3	0	6.7	26.7	20.0	0	26.7	6.7	0	100%
Nutrition	0	16.7	0	0	0	33.3	50.0	0	0	0	100%
Population	0	16.7	0	0	0	16.7	16.7	0	0	0	100%
Health	0	3.3	0	6.7	33.3	26.7	23.3	6.7	0	0	100%
Education	0	0	4.6	13.6	13.6	27.3	22.7	9.1	4.6	4.6	100%
Human Resource Development	0	0	0	23.5	11.8	35.3	17.6	11.8	0	0	100%
Infrastructure and Housing	0	0	0	11.1	16.7	22.2	27.8	11.1	11.1	0	100%
Other	0	0	0	2.6	13.2	31.6	21.0	15.8	7.9	7.9	100%

EXHIBIT 24: STATISTICS FOR TIME TAKEN TO CONDUCT EVALUATION					
TIME TAKEN TO DO EVALUATION (for 147 reports)	NUMBER OF CASES	MEAN TOTAL SCORE	STANDARD DEVIATION	MEAN QUALITY SCORE	STANDARD DEVIATION
0-3 weeks	105	54.3	15.1	12.3	4.6
4-6 weeks	30	57.9	16.9	13.7	4.4
7-9 weeks	7	49.7	4.9	11.6	1.4
10-12 weeks	0	----	----	----	----
12+ weeks	5	77.0	14.0	23.6	8.4

The large majority of evaluations (105 of 147 records) were conducted in three weeks or less. The only substantial evidence that more time is associated with better evaluations is in the 12+ weeks category, which contains the Impact evaluations. This group of evaluations had mean total and quality scores substantially

28

higher than the other groups. It contains only five cases, however, so conclusions must be made with caution. It should also be noted that the mean total and quality scores in the 7-9 weeks category are lower than for the preceding (4-6 weeks) group. This would put into question the hypothesis that more time spent yields better evaluations.

EXHIBIT 25: TIME TAKEN TO COMPLETE EVALUATION BY TOTAL SCORE										
TIME TAKEN	TOTAL SCORE (for 147 cases)									
	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	TOTAL
0-3 weeks	3.8%	1.9	7.6	20.9	30.4	23.8	7.6	3.8	0	100%
4-6 weeks	0	3.3	13.3	13.3	23.3	16.7	20.0	10.0	0	100%
7-9 weeks	0	0	0	57.1	42.9	0	0	0	0	100%
10-12 weeks	0	0	0	0	0	0	0	0	0	100%
12+ weeks	0	0	0	0	0	20.0	40.0	20.0	20.0	100%

K. LOGFRAME LEVELS AND EVALUATION TIME

It would be expected that the time at which an evaluation is conducted, in relation to the life of the project, would have an influence on the number of logframe levels examined. For example, for a project with the sample logframe below, if the evaluation discussed the "input" and "output" levels, it examined two levels; if it examined the "purpose" level as well, it would have addressed three levels, and so on.

29

GOAL (Level 4)	Students prepared in technical skills.
PURPOSE (Level 3)	More students can attend technical school.
OUTPUTS (Level 2)	1. Schools built 2. Teachers trained
INPUTS (Level 1)	1. Grants and loans 2. Technical assistance 3. Commodities

The later in the life of the project an evaluation is conducted, the more levels it would be expected to examine.

The evaluations were categorized according to evaluation time: interim, final, impact, and other. The percentages of the different numbers of levels examined are shown in the table below for each evaluation time category:

EXHIBIT 26: EVALUATION TIME BY NUMBER OF LEVELS EXAMINED								
EVALUATION TIME	NUMBER OF LOGFRAME LEVELS EXAMINED							
	0	1	2	3	4	5	6	TOTAL
Interim	8.1%	23.0%	35.1%	25.0%	7.4%	0.7%	0.7%	100.0%
Final	1.8	32.1	30.4	25.0	3.6	5.4	1.8	100.0
Impact	0	14.3	14.3	42.9	28.6	0	0	100.0
Other	0	50.0	16.67	16.67	16.67	0	0	100.0

Impact evaluations clearly show a tendency to examine more levels than the other evaluations. What seems unexpected is that the data indicate Final evaluations do not examine more levels, on the

20

whole, than do Interim evaluations. This ought to be of some concern to AID if it is interested in being appropriately informed about the results of their projects and all their phases, based on the point in time of the project that evaluations are conducted.

L. LEVELS RELATED TO BUREAU, TECHNICAL CODE AND SCORER

In examining the table of distribution of logframe project levels by bureau, it is clear that Impact evaluations examine the highest number of levels. This should be self-evident, since Impact evaluations are intended to examine the upper levels of the logframe. Another bureau which merits mention is LAC, in light of the fact that it scored relatively low on the number of logframe levels examined versus all other bureaus. This bureau had, by far, the lowest percentage of evaluations examining only one level (16.1%), with the exception of Impact evaluations. The next lowest bureau was FVA, with 27.3%. It is necessary to keep in mind, however, the difference in sample size between LAC (31) and FVA (11).

The relationship between the number of logframe levels examined versus technical focus is difficult to analyze, because the number of cases in many of the technical focus categories is so small. This renders any conclusions drawn from the distribution as highly speculative. The table below presents technical focus by number of levels examined.

EXHIBIT 27: TECHNICAL FOCUS BY NUMBER OF LEVELS EXAMINED								
TECHNICAL FOCUS	NUMBER OF LEVELS EXAMINED							TOTAL
	0	1	2	3	4	5	6	
Agriculture	7 9.2%	22 28.9	20 26.3	17 22.4	7 9.2	3 3.9	0 0	76 100%
Rural Non-Agriculture	0 0	3 20.0	9 60.0	3 20.0	0 0	0 0	0 0	15 100
Rural Multi-Function	2 20.0	2 20.0	5 50.0	1 10.0	0 0	0 0	0 0	10 100
Nutrition	1 25.0	1 25.0	1 25.0	1 25.0	0 0	0 0	0 0	4 100
Population	0 0	4 50.0	2 25.0	1 12.5	1 12.5	0 0	0 0	8 100
Health	2 7.7	3 11.5	11 42.3	7 26.9	2 7.7	1 3.8	0 0	26 100
Education	1 6.7	3 20.0	3 20.0	7 46.6	0 0	0 0	1 6.7	15 100
Human Resources	0 0	1 8.3	5 41.7	3 25.0	3 25.0	0 0	0 0	12 100
Infrastructure and Housing	0 0	5 33.3	6 40.0	1 6.7	2 13.3	0 0	1 6.7	15 100
Other	0 0	12 33.3	9 25.0	14 38.9	1 2.8	0 0	0 0	36 100
TOTAL FREQUENCY	13	56	71	55	16	4	2	217
OVERALL PERCENT	6.0	25.8	32.7	25.4	7.4	1.8	0.9	100

Regarding scores, an analysis of logframe levels examined versus scorer reveals some differences among the individuals who

2/2

read and scored the evaluations. Scorer 1 found more levels per evaluation, on the average, than the other scores: 54.3% of Scorer No. 1's cases had three levels, while the majority of the other scores' cases had two levels. Scorers No. 3 and 4 had no evaluations with 3 or more levels. What may help to explain this is the disproportionate number of Impact evaluations scored by Scorer No. 1, even though this is a small fraction of the total number of reports scored by that individual.

M. RESOURCE REALLOCATION AND SCORES

While reading the evaluations, the scorers were asked to answer the following questions: "Did the evaluators discover that resources needed to be reallocated among all the inputs to achieve outputs?" and "Did the evaluators discover that resources needed to be reallocated among the outputs to achieve project purpose?" It was found that most evaluations indicated no need for either input or output reallocation. The total scores by deciles are shown below for the evaluations that did and did not identify the need for input and output reallocation.

EXHIBIT 28: REALLOCATION OF RESOURCES BY SOURCES										
	<u>TOTAL SCORE</u> (for 117 reports)									
	10- 20	20- 30	30- 40	40- 50	50- 60	60- 70	70- 80	80- 90	90- 100	TOTAL
INPUT REALLOCATION										
Yes (29.9%)	2.9%	0%	0%	25.7%	34.3%	20.0%	11.4%	5.7%	0%	100.0%
No (70.1%)	3.7%	1.2%	10.9%	21.9%	20.7%	20.7%	12.2%	8.5%	0%	100.0%
OUTPUT REALLOCATION										
Yes (20.5%)	0%	0%	4.2%	33.3%	33.3%	12.5%	8.3%	8.3%	0%	100.0%
No (79.5%)	4.3%	1.1%	8.6%	20.4%	22.6%	22.6%	11.8%	8.6%	0%	100.0%

37

It is clear from this table that there is little correlation between scores and the identification of a need for input or output reallocations by the evaluators.

N. TIME SERIES ANALYSIS

The important relationships between the 1983 and 1984 meta-evaluation results are discussed below.

Total scores were calculated on approximately the same basis both years. (See Section III, "Methodology", for details.) There was an overall higher tendency in the 1984 scores.

There is a notable difference between the outcomes of the 1983 (FY82 reports scored) and 1984 (FY 83 reports scored) meta-evaluation projects in the area of score differentiation by bureaus (see table below). While in the metaevaluation study of FY 82 reports, the distribution for any one bureau did not differ very much from the overall distribution, the differences were much greater for the FY 83 report distributions. The most marked difference were in the Asia and Impact evaluations. For FY 82 reports, the mean total score for Asia bureau evaluations was 8.7% lower than the overall mean; in the current study of FY 83 reports, the bureau's means total score was 5.8% higher than the overall mean. One must view these statistics with caution, however, because the sample of Impact reports in both years was very small.

EXHIBIT 29: TOTAL SCORE STATISTICS FOR EACH BUREAU (FY82/FY83)			
BUREAU	NUMBER OF CASES	MEAN TOTAL SCORE	STANDARD DEVIATION
Near East	40/33	53.5/51.1	13.3/13.0
Asia	31/59	47.3/56.9	17.5/16.4
Latin America	49/38	51.8/50.9	17.1/17.7
Africa	92/103	53.5/52.5	16.0/14.1
Impact	16/8	56.9/65.9	10.5/10.7
SciTech	26/14	45.4/52.1	13.6/21.3
FVA	12/15	52.0/59.1	15.7/15.5

In the area of individual characteristics, the most notable difference between the FY82 and FY83 scores on individual characteristics was in Characteristic 3, "focus on user needs." The wording of the questions was identical both years. This characteristic's mean score was well below average in 1983 and well above average in 1984. The table below shows all characteristics' ascending mean scores, rounded off to whole numbers, for all bureaus in 1983 (FY82 reports) and 1984 (FY83 reports). The scoring systems were different; hence, the difference in scores. The maximum value for each characteristic in the FY82 reports was 100, while the maximum for the FY83 reports was 11.1. Note the position of Characteristic 3.

35

EXHIBIT 30: MEAN CHARACTERISTIC SCORES FOR FY 82 AND FY 83			
FY82 REPORTS		FY83 REPORTS	
<u>CHARACTERISTIC</u>	<u>MEAN SCORE</u>	<u>CHARACTERISTIC</u>	<u>MEAN SCORE</u>
1	48	4	3.5
3	49	1	4.2
6	49	6	4.8
4	52	9	5.6
5	52	7	5.9
7	55	5	6.0
9	57	8	6.4
2	58	3	8.4
8	58	2	9.1

Based on the above data, it cannot necessarily be concluded that AID evaluations have made vast improvements in their attention to user needs," Characteristic 3. Rather, an analysis of the scorers' perceptions of the questions in the scoring instrument was undertaken in order to explain the difference. Interviews with the scorers revealed that of the four 1984 scorers, the two who did not participate in the 1983 metaevaluation may have applied the "user needs" questions less rigorously than did the other two.

The characteristics pertaining to data collection, data analysis, and evaluation design (1, 4 and 6) received low scores relative to the other characteristics in both studies. Identification of evaluation objectives and questions and of project objectives (Characteristic 2) and answers to evaluation questions (Characteristic 8) received relatively high scores both years.

While the 1983 metaevaluation of FY82 reports found evaluation cost to be positively associated with quality, this study did not find conclusive evidence to that effect.

Similar findings resulted for the variable concerning the time taken to conduct evaluations in comparing the 1983 and 1984 results. The association found in 1983 was not found in 1984.

## II. RECOMMENDATIONS

### A. GENERAL RECOMMENDATIONS

The recommendations herein are based on the major findings of this study as regards to the aspects of evaluation reports examined in the metaevaluation process. They are by necessity general in nature, and pertain to the improvement of the quality and completeness of AID evaluations.

1. It would be worthwhile for the benefit of the other bureaus for AID to take a closer examination at the way in which the Impact and Asia evaluations are conducted. Impact evaluation reports scored the best overall by TRITON's standards, and the Asia evaluations scored significantly better than those of the other regional bureaus. In addition, they have accomplished great improvement between FY82 and FY83. Although it is true that Impact evaluations tend to be more expensive than others, it is still worthwhile looking into their content, structure, and process as a means to improve evaluation efforts overall at AID.

2. The areas of data collection, data analysis, and evaluation design need the greatest improvement. They received the lowest scores overall of the nine characteristics assessed. Last year's metaevaluation also supports this conclusion. This is a crucial area of evaluation and merits serious attention for remedial action by all bureaus.

3. AID should examine evaluation costs carefully. Although the 1983 metaevaluation found a tendency for higher costs to associate with high quality, the current study found little association between the two variables. Thus, it must never be assumed that "more money means a better evaluation." Perhaps evaluations can be more efficiently utilized keeping this in mind. It would be worthwhile, however, to look into all major determinants of

evaluation cost in order to economize on evaluations and improve their efficiency.

4. If host country participation is a goal in AID evaluations, work must be done to encourage it in evaluations conducted by non-mission entities. The Asia bureau incorporates host country participation much more frequently than any other bureau. Asia evaluations also scored well in comparison to the other bureaus. Their methods might serve as a model for other bureaus wishing to include host country participation in evaluations.

5. The time taken to complete an evaluation appears to have little to do with quality or completeness. Although it might seem logical that more time spent would yield better quality, the results of this study do not support that hypothesis. Perhaps too much time is being spent on some evaluations as well as not enough on others. (This would be consistent with the "cost vs. quality" finding discussed above).

#### B. RECOMMENDATIONS FOR FUTURE META-EVALUATIONS

A metaevaluation can be a useful tool if it is conducted with specific goals. The information collected and produced during the course of such a study represents a data base which can serve many different functions. The scope of this study was very broad, and comprised many focal points. A more directed approach would be easier to conduct and more useful, and would likely improve scoring and statistical procedures. The recommendations made in this study might serve as the basis for further, more specific studies.

A metaevaluation is an "evaluation" of evaluations" and, therefore, should be conducted under all the requirements for a good evaluation. A more limited and focused scope would be the first step in conducting a meaningful evaluation of evaluations. Lastly, uniformity in metaevaluation procedures and analyses from

year-to-year would enable AID to develop a powerful tool to identify and address deficiencies in its evaluation activities.

### III. METHODOLOGY

#### A. DEVELOPMENT OF PROJECT

During FY 1982, TRITON conducted a "metaevaluation" project to assess the quality and completeness of the Agency for International Development's evaluation reports, under the auspices of AID's Program Evaluation Systems Division.<sup>1/</sup> A scoring instrument was designed to provide that division with a diagnostic tool to support its work in monitoring the Agency's evaluation system. It is based on a series of key issues concerned with quality and completeness for AID evaluation reports.<sup>2/</sup> The evaluations were read and then rated using the instrument developed by TRITON.

The ultimate use of the instrument is to build up a data base derived from the routine review and scoring of AID evaluation reports which will help determine the strengths and weaknesses of the reports, based on sector, geographic focus, and other specific aspects of an evaluation. The instrument can be used on all types of AID evaluation reports, including mid-term evaluations, end-of-project evaluations and impact studies.

The following steps were taken to develop an instrument to assess the quality and quantity aspects of AID evaluations. A report identifying the attributes of a "good" evaluation was developed based on evaluation literature and interviews with relevant personnel from various organizations involved in development projects. This compilation of factors served as the basis for

---

1/ Final Report: Analysis of the Quality of FY80-82 AID Evaluation Reports. Contract No. AID/SOD/PDC-0391, Work Order No. 2.

2/ Final Report: Development of a Quality/Completeness Scoring Instrument for USAID Evaluation Reports. Contract No. AID/SOD/PDC-0391, Work Order No. 1.

developing a scoring system for AID evaluation reports. The Program Evaluation Systems Division of AID performed a content analysis of the factors to identify the major quality and completeness characteristics and to segregate a number of subfactors within each major category. Nine major internal factors which could be measured solely by reviewing the evaluation report were isolated. This list of factors was refined and ranked by relevant individuals within and outside of AID in order to develop the numerical weighting and scoring process. The draft of the scoring instrument was then tested and further revised.

In this final version, nine (9) characteristics--six (6) of which were further broken down into sub-characteristics--of a "good" evaluation were identified. Each evaluation report is rated for each characteristic/subcharacteristic on a scale of 0-4 (low-high), with a "not applicable" possibility for some sub-characteristics. These scores are then summed, weighted and normalized on a 0-100 scale.

The 1984 metaevaluation utilized a scoring instrument which was more qualitative than previous metaevaluations. The new instrument records more specific external information about the evaluation such as host country participation, and measures attribution, sustainability and external influences.

#### B. DESCRIPTION OF INSTRUMENT

Development of the scoring analysis procedure resulted in a six-part scoring instrument which consists of the following:

1. The Facesheet: contains specific information about the project (the list of "findings" prepared by the reviewer is attached to this form);
2. Findings: are short, concise sentences referring to conclusions and recommendations found in the evaluation;

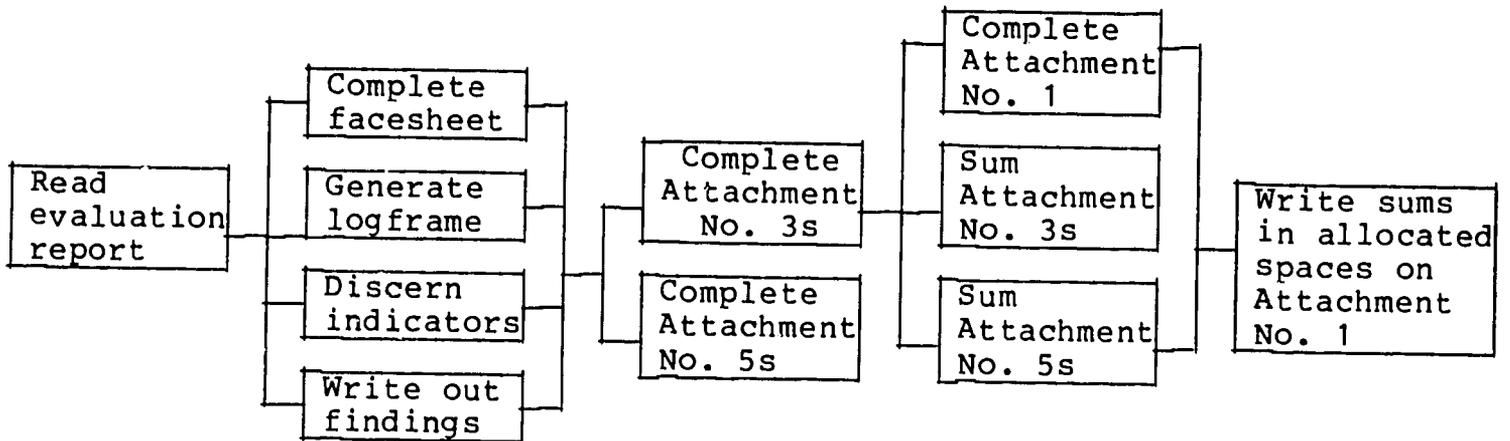
3. The logical framework presents a project or program in terms of its inputs, outputs, purpose, goals, assumptions and unexpected results.
4. Attachment 1 consists of a series of statements pertaining to the evaluation report's quality and completeness, which the reviewer scores according to the extent to which they are true for the evaluation;
5. Attachment 3 reflects how well the evaluation report assessed a project's various components e.g., inputs, outputs; and
6. Attachment 5 records the quality of the evaluation of the management transformation and the hypotheses. It examines the degree to which the transformation from level to level was evaluated and how well.

Both internal and external factors are recorded on the scoring instrument. The internal variables (Attachments 1, 3 and 5) are utilized to score the evaluation for quality and completeness, while the external variables (Facesheet and part of Attachment 5) are used to analyze scoring trends and patterns. Internal variables assess completeness, clarity, appropriateness, validity, replicability, reliability, adequacy, and bias. External variables taken into account are: geographic bureau, type of evaluation, timing of evaluation, AID management unit, technical code, length of time for evaluation, host country participation, levels of logframe evaluated, evaluation cost, project cost and contractor/evaluation entity. In essence, the internal variables assess the evaluation report as a self-contained entity, while external variables serve to situate the evaluation report in particular environmental, cultural and developmental contexts.

The 1984 TRITON Scoring Instrument can be found in the Appendix.

### C. APPLICATION OF INSTRUMENT

The sequence for completing the various parts to the scoring instrument is outlined in the flowchart below.



### D. SCORING PROCESS

The scoring process itself is the second stage in scoring the evaluation reports. This procedure is generally done by someone other than the reviewer who scored the report in order to avoid bias. A scoring sheet is used to calculate the overall score of an evaluation according to the forms which have been filled out by the reviewer. In this process values from the attachments are weighted and totalled. The overall score is a reflection of an evaluation's performance on a 1-100 scale, as measured by TRITON's nine internal characteristics. The scoring sheet is contained in Appendix B.

### E. DATA PROCEDURES

Once all the data were collected, there is only one intervening step required before analysis is begun. This intervening step

requires putting the data into a computer-readable form and "cleaning" those data sets. A variety of coding formats was devised and is included as part of Appendix C. These required the transformation of names into numbers.

The computer application of this analysis was performed using the Statistical Analysis System (SAS) on AID's mainframe computer.

Computer analysis provides a rapid and accurate statistical application and the data are put into a quickly legible format.

#### F. STATISTICAL PROCEDURES

The statistics used for the analysis scores were frequency distributions and two by two tables analyzed using chi-square. These are parametric statistics: that is, they are used when the data can be assumed to have a specific type of distribution.

Frequency is the number of times a variable may occur. For example, the frequency of the Agriculture techcode is 78: there are 78 agricultural projects in the FY83 metaevaluation.

A frequency distribution is the arrangement of those frequencies, usually by another variable. Thus, the number of agricultural projects read by each coder would be a good example of such a distribution.

Two by two tables are statistical measures to determine the effect one variable may exert on another. One variable (e.g., agriculture projects) is grouped against all other techcodes on one axis, while the other axis might have design against all other major headings. In this was the proportion of design findings in agricultural projects can be assessed to determine if that frequency is statistically significant.

## G. ANALYSIS PLAN

The data were analyzed to provide answers to the following series of questions:

- o Does one (or more) bureau score significantly higher or lower than the others? If so, why?
- o What are the particular strengths/weaknesses (judged by scores of internal characteristics) of each bureau?
- o Does a project that budgets more for evaluation actually produce "better" evaluations?
- o Is there a differentiation among the types of evaluators, especially with regards to quality, completeness and data management?
- o What is the relationship between host country participation and the unit organizing the evaluation?
- o What is the relationship between host country participation and the types of evaluators?
- o How much of the difference in scores is due to the number of evaluation entities?
- o Is there a difference in overall score and the scores of individual internal characteristics with regards to technical activities? (The hypothesis is that a more scientific evaluation would tend to score better than a more narrative one.)
- o What is the relationship between the time taken to do the evaluation and the overall score? (The hypothesis is that the longer the evaluators spent on the study, the better the score.)
- o What is the relationship, if any, between the numbers of logframe levels examined and the "evaltime"? (The hypothesis is that a final or ex post evaluation would examine more levels than an interim one.)
- o Does one or more bureau, technical activity or coder/reviewer typically produce a greater number of evaluation levels than the others? Why?

- o Did one or more bureau or one technical focus use more innovative techniques? What internal characteristics does the majority of those techniques address?<sup>3/</sup>
- o Is there a relationship between reallocation of resources and overall scores? Does such a relationship exist in the bureaus or technical activities? Is a high or low-scoring evaluation more likely to examine reallocated resources?

The answers to these questions are presented in Part I of this report, together with any evidence of statistical significance.

---

<sup>3/</sup> See: Final Report: Innovative Techniques Observed During the FY 1983 Metaevaluation Project. Contract No. OTR-0000-C-00-3482-00, November 1984.

APPENDIX

FACE SHEET DATA

1. Project Title \_\_\_\_\_
2. Project Number \_\_\_\_\_
3. Mission/AID/W/Office \_\_\_\_\_
4. Year of Evaluation Review \_\_\_\_\_
5. Evaltype \_\_\_\_\_
6. Evaltime \_\_\_\_\_
7. Mangunit \_\_\_\_\_
8. Host Country Participation on the Evaluation Team?  
Yes \_\_\_\_\_ No \_\_\_\_\_ Can't Tell \_\_\_\_\_
9. Contractor(s): List principal one first  
\_\_\_\_\_
10. Author(s):
11. Time taken to do Evaluation \_\_\_\_\_
12. Time taken to Score Evaluation \_\_\_\_\_
13. No. levels examined \_\_\_\_\_
14. General Indicator of evaluation completeness/innovative techniques.
15. Mission comments:
16. Scope of Work Included in the Documents?  
Yes \_\_\_\_\_ No \_\_\_\_\_
17. Did the evaluators discover that resources needed to be reallocated among all the inputs to achieve outputs?  
Yes \_\_\_\_\_ No \_\_\_\_\_  
If yes, describe how:

FACE SHEET DATA  
(Continued)

18. Did the evaluators discover that resources needed to be reallocated among the outputs to achieve project purpose?

Yes \_\_\_\_\_ No \_\_\_\_\_

If yes, describe how:

**ATTACHMENT. 1**

**OVERALL SCORING INSTRUMENT**

**(with scales for Completeness, Clarity and Appropriateness)**

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

**SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC**

1. The indicators are appropriate given the evaluation questions.

Appropriateness            0            1            2            3            4

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.

Enter values from Worksheet:

Summary Score for U elements: \_\_\_\_\_

Summary Score for E elements: \_\_\_\_\_

Summary Score for A elements: \_\_\_\_\_

Summary Score for Output elements: \_\_\_\_\_

Summary Score for Input elements: \_\_\_\_\_

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.

Enter values from Worksheet:

Summary Score for MT elements: \_\_\_\_\_

Summary Score for H elements: \_\_\_\_\_

4. Assumptions made by the design are clearly and completely stated.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

6. The evaluation design is fully and clearly described by the evaluation report.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

**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.

**SUBFACTORS TO BE ASSESSED FOR THIS CHARACTERISTIC**

1. Project or program objectives are clearly and completely stated.

Completeness: 0            1            2            3            4

Clarity:            0            1            2            3            4

2. The objectives of the evaluation are clearly and completely stated; priorities among objectives and reasons for some are clear.

Completeness: 0            1            2            3            4

Clarity:            0            1            2            3            4

3. The evaluation questions are clearly and completely stated; priorities among questions are clear.

Completeness: 0            1            2            3            4

Clarity:            0            1            2            3            4

55

**CHARACTERISTIC III: The evaluation focuses on the evaluation users and their needs/questions.**

**SUB-FACTORS TO BE ASSESSED FOR THIS CHARACTERISTIC**

1. Evaluation clients/users are clearly and completely identified.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

2. User needs/expectations are clearly and completely identified.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

3. Areas of "public interest"/broad concern covered by the evaluation are clearly identified.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

**CHARACTERISTIC IV: The data collection procedures/secondary data are appropriate and adequate, not excessive or inadequate.**

**SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC**

1. Instruments/approaches for collecting data are valid and reliable;

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4
Appropriateness:	0	1	2	3	4

2. Validity and reliability of any secondary data is checked and found acceptable.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

3. Sources of error/biases in the instruments or data collection procedures are described as fully as possible.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

4. 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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

5. Neither too much or too little data is secured.

Appropriateness: 0 1 2 3 4

6. Where cross-cultural sensitivity, language, etc. are potential issues, they are properly handled (e.g. local data collectors used, female data collectors, etc.)

Completeness: 0 1 2 3 4 N/A

Clarity: 0 1 2 3 4 N/A

Appropriateness: 0 1 2 3 4 N/A

7. 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.

Completeness: 0 1 2 3 4 N/A

Clarity: 0 1 2 3 4 N/A

Appropriateness: 0 1 2 3 4 N/A

8. Instruments used to collect raw data, such as questionnaires, are included as exhibits to evaluation reports.

Completeness: 0 1 2 3 4 N/A

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

**SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTICS**

1. Facts are separated from interpretations.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

2. Alternative interpretations are discussed.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

3. The reason for selecting a specific interpretation or conclusion is made clear.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

4. Conclusions are separated from recommendations.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

5. Alternative recommendations are discussed and the reason for selecting a specific recommendation is made clear.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

6. The reasons for selecting a specific recommendation are made clear

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

7. 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.

Clarity:                    0            1            2            3            4

8. 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.

Clarity:                    0            1            2            3            4

Appropriateness: 0            1            2            3            4

9. 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?

Completeness:            0            1            2            3            4

Clarity:                    0            1            2            3            4

**CHARACTERISTIC VI: The data analysis procedures are appropriate and adequate.**

**SUB-FACTORS 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.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4
Appropriateness:	0	1	2	3	4

2. The analysis procedures are appropriate; they are neither weak nor excessive.

Appropriateness:	0	1	2	3	4
------------------	---	---	---	---	---

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

4. Both quantitative and qualitative data are analyzed if both were secured.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

6. The strength and weaknesses of the data analysis aspects of the evaluation are clearly and completely stated.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A

**CHARACTERISTIC VII:** The evaluation report is a well-written, self contained document.

**Completeness:** 0      1      2      3      4

**Clarity:** 0      1      2      3      4

**CHARACTERISTIC VIII:** The evaluation produces the types of information it was expected to produce; i.e., insofar as possible, the full set of evaluation questions are answered.

**Completeness:** 0      1      2      3      4

**Clarity:** 0      1      2      3      4

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

**Completeness:** 0      1      2      3      4

**Clarity:** 0      1      2      3      4

**Appropriateness:** 0      1      2      3      4

**ATTACHMENT 1**

**OVERALL SCORING INSTRUMENT**

**(with scales for Completeness, Clarity and Appropriateness)**

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

**SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC**

1. The indicators are appropriate given the evaluation questions.

Appropriateness            0            1            2            3            4

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.

Enter values from Worksheet:

Summary Score for U elements: \_\_\_\_\_

Summary Score for E elements: \_\_\_\_\_

Summary Score for A elements: \_\_\_\_\_

Summary Score for Output elements: \_\_\_\_\_

Summary Score for Input elements: \_\_\_\_\_

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.

Enter values from Worksheet:

Summary Score for MT elements: \_\_\_\_\_

Summary Score for H elements: \_\_\_\_\_

4. Assumptions made by the design are clearly and completely stated.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

6. The evaluation design is fully and clearly described by the evaluation report.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

**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.**

**SUBFACTORS TO BE ASSESSED FOR THIS CHARACTERISTIC**

1. Project or program objectives are clearly and completely stated.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

2. The objectives of the evaluation are clearly and completely stated; priorities among objectives and reasons for some are clear.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

3. The evaluation questions are clearly and completely stated; priorities among questions are clear.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

**CHARACTERISTIC III: The evaluation focuses on the evaluation users and their needs/questions.**

**SUB-FACTORS TO BE ASSESSED FOR THIS CHARACTERISTIC**

1. Evaluation clients/users are clearly and completely identified.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

2. User needs/expectations are clearly and completely identified.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

3. Areas of "public interest"/broad concern covered by the evaluation are clearly identified.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

**CHARACTERISTIC IV: The data collection procedures/secondary data are appropriate and adequate, not excessive or inadequate.**

**SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC**

1. Instruments/approaches for collecting data are valid and reliable;

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4
Appropriateness:	0	1	2	3	4

2. Validity and reliability of any secondary data is checked and found acceptable.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

3. Sources of error/biases in the instruments or data collection procedures are described as fully as possible.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

4. 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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

5. Neither too much or too little data is secured.

Appropriateness: 0 1 2 3 4

6. Where cross-cultural sensitivity, language, etc. are potential issues, they are properly handled (e.g. local data collectors used, female data collectors, etc.)

Completeness: 0 1 2 3 4 N/A

Clarity: 0 1 2 3 4 N/A

Appropriateness: 0 1 2 3 4 N/A

7. 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.

Completeness: 0 1 2 3 4 N/A

Clarity: 0 1 2 3 4 N/A

Appropriateness: 0 1 2 3 4 N/A

8. Instruments used to collect raw data, such as questionnaires, are included as exhibits to evaluation reports.

Completeness: 0 1 2 3 4 N/A

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

**SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTICS**

1. Facts are separated from interpretations.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

2. Alternative interpretations are discussed.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

3. The reason for selecting a specific interpretation or conclusion is made clear.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

4. Conclusions are separated from recommendations.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

5. Alternative recommendations are discussed and the reason for selecting a specific recommendation is made clear.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

6. The reasons for selecting a specific recommendation are made clear

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

7. 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.

Clarity:                    0            1            2            3            4

8. 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.

Clarity:                    0            1            2            3            4  
Appropriateness: 0            1            2            3            4

9. 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?

Completeness:            0            1            2            3            4  
Clarity:                    0            1            2            3            4

**CHARACTERISTIC VI: The data analysis procedures are appropriate and adequate.**

**SUB-FACTORS 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.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4
Appropriateness:	0	1	2	3	4

2. The analysis procedures are appropriate; they are neither weak nor excessive.

Appropriateness:	0	1	2	3	4
------------------	---	---	---	---	---

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

4. Both quantitative and qualitative data are analyzed if both were secured.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A
Appropriateness:	0	1	2	3	4	N/A

6. The strength and weaknesses of the data analysis aspects of the evaluation are clearly and completely stated.

Completeness:	0	1	2	3	4
Clarity:	0	1	2	3	4

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.

Completeness:	0	1	2	3	4	N/A
Clarity:	0	1	2	3	4	N/A

**CHARACTERISTIC VII:** The evaluation report is a well-written, self contained document.

<b>Completeness:</b>	0	1	2	3	4
<b>Clarity:</b>	0	1	2	3	4

**CHARACTERISTIC VIII:** The evaluation produces the types of information it was expected to produce; i.e., insofar as possible, the full set of evaluation questions are answered.

<b>Completeness:</b>	0	1	2	3	4
<b>Clarity:</b>	0	1	2	3	4

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

<b>Completeness:</b>	0	1	2	3	4
<b>Clarity:</b>	0	1	2	3	4
<b>Appropriateness:</b>	0	1	2	3	4

**ATTACHMENT 2  
LOGICAL FRAMEWORK**

E-4

**Hd**

E-3

**Hc**

E-2

**Hb**

E-1

**Ha**

**OUTPUTS**

**MANAGEMENT  
TRANSFORMATION**

**INPUTS**

A-E3

A-E2

A-E1

o

A-0

**ATTACHMENT 3**

**RATING FORM FOR SCORING INPUTS, OUTPUTS,  
DEPENDENT VARIABLES ASSUMPTIONS,  
AND UNPLANNED RESULTS**

Note: Complete 1 copy of Form to address all INPUTS together;  
Complete 1 copy of Form for each OUTPUT.  
Complete 1 copy of Form for each DEPENDENT VARIABLE  
Complete 1 copy of Form for each set of ASSUMPTIONS

A. Type of variable addressed by this project element being evaluated:

- \_\_\_\_\_ Independent variable (for this project/program/policy)
- \_\_\_\_\_ Dependent variable (for this project/program/policy)
- \_\_\_\_\_ Other. Specify type of variable/element and describe:

B. Number of indicators used in evaluation report to measure status of variable. \_\_\_\_\_

C. Answer for each indicator measured for this element:

(1) Check which of these is applicable:

Ind Ind Ind Ind Ind Ind  
1 2 3 4 5 6

- \_\_\_\_\_ a. Presence/absence (i.e., indicator was no present "before" activity being evaluate began).
- \_\_\_\_\_ b. Change in status (i.e., indicator was present "before" activity being evaluate began; measure focuses on change)

(2) Complete only if C (1) response = presence/absence (response a). Score 0 = No, 2 = Somewhat, 4 = Yes:

Ind Ind Ind Ind Ind Ind  
1 2 3 4 5 6

- \_\_\_\_\_ (a) Measure was valid measure of presence/absence for the indicator
- \_\_\_\_\_ (b) Measure was replicable
- \_\_\_\_\_ (c) Measure was unbiased
- \_\_\_\_\_ (d) Measure was objective

(3) Complete only if C (1) response = change in status (response

b). Score 0 = No, 2 = Somewhat, 4 = Yes

Ind Ind Ind Ind Ind Ind  
1 2 3 4 5 6

- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ (a) Measure was valid measure of indicator which was to have changed
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ (b) Measures at all points were made in consistent manner
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ (c) Measures of indicator was unbiased
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ (d) Measure was adequate, given inherent variability in indicator
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ (e) Measures at all points were objective

D. Generalization: Complete only if evaluation sought/attempted to generalize for a universe based on measures made of indicator for a subset of that relevant universe. Enter one value for each indicator form which a generalization was made:

Ind Ind Ind Ind Ind Ind  
1 2 3 4 5 6

- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Statistically sound/representative sample = 4
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Random selection procedure/universe size unkn = 3
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Criteria or other purposive sample = 2
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Convenience or volunteer sample = 1
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Single case (of larger universe) = 1
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Only case (automatic census)/all cases = 4
- \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ \_\_\_ Can't tell from evaluation report = 0

20

Ind 1 Ind 2 Ind 3 Ind 4 Ind 5 Ind 6 Total for All Indicators

_____	_____	_____	_____	_____	_____	_____	Validity: Score from C(2)(a) C(3)(a)
_____	_____	_____	_____	_____	_____	_____	Replicability/consistency: Score from C(2)(b) or C(3)(b)
_____	_____	_____	_____	_____	_____	_____	Bias: Score from C(2)(c) or C(3)(c)
_____	_____	_____	_____	_____	_____	_____	Representativeness/Adequacy: Score from C(3)(d)
_____	_____	_____	_____	_____	_____	_____	Objectivity: Score from C(2)(e) C(3)(e)
_____	_____	_____	_____	_____	_____	_____	Generalization: Score from Item E
_____	_____	_____	_____	_____	_____	_____	Grand Total

F. Summary score on indicators

(1) If C(1) response = presence/absence (response a), then complete the following computation:

	Score from Item E	Max. Poss. Score	Norm. Score	
Validity Score	_____	_____	_____ x .40	= _____
Reliability Score	_____	_____	_____ x .30	_____
Objectivity Score	_____	_____	_____ x .15	_____
Unbiasedness Score	_____	_____	_____ x .15	_____
Total				_____

complete the following computation:

	Score from Item E	Max. Poss. Score	Norm. Score	
Validity Score	_____	_____	_____	x .30 = _____
Reliability Score	_____	_____	_____	x .30 = _____
Objectivity Score	_____	_____	_____	x .20 = _____
Unbiasedness Score	_____	_____	_____	x .20 = _____
Total	_____	_____	_____	_____

(3) Overall Confidence Level:  
F(1) or F(2) Score + D Score = \_\_\_\_\_

WP-436

gk

**ATTACHMENT 5**

**RATING FORM FOR SCORING THE MANAGEMENT  
TRANSFORMATION AND HYPOTHESES (Ha, Hb, Hc....)**

**Note: Complete 1 copy of Form for the MANAGEMENT TRANSFORMATION  
Complete 1 copy of Form for all HYPOTHESES (Ha, Hb, etc.)**

Element being scored: \_\_\_\_\_

(MT or H)

---

Type of alpha element (check one):

\_\_\_\_\_ Management transformation (no hypothesis presented; i.e., "effective management" is the primary process needed to generate desired effects).

\_\_\_\_\_ Hypothesis (from independent to dependent variable, planned or unplanned, etc.)

---

A. Answer if element = Management Transformation:

(1) What was examined to determine ~~whether~~ transformation occurred:

\_\_\_\_\_ (a) Outcome only (specify which outcomes, as per diagram in Attachment 2: Output # \_\_\_\_\_)

\_\_\_\_\_ (b) Process, from a quality standpoint

\_\_\_\_\_ (c) Process, from an efficiency standpoint (specify from from which perspective(s): \_\_\_\_\_ time, \_\_\_\_\_ cost, \_\_\_\_\_ time and cost)

\_\_\_\_\_ (d) Process, from another standpoint. Specify:

---

---

ed

(2) Complete only if answer to A(1) = process in any form  
(response b, c or d); Score 0 = No, 2 = Somewhat, 4 = Yes:

\_\_\_\_\_ Process measure was valid for situation.

\_\_\_\_\_ Process measure was reliable.

\_\_\_\_\_ Process measure was unbiased.

\_\_\_\_\_ Process measure was objective.

B. Complete only if element = hypothesis:

(1) Was the logic requirement that the hypothesized cause  
preceded the effect met: \_\_\_\_\_ Yes \_\_\_\_\_ No \_\_\_\_\_ Can't Tell

(2) Was the logic requirement that the hypothesized cause  
and effect covaried (both changed in status) met: \_\_\_\_\_  
Yes \_\_\_\_\_ No \_\_\_\_\_ Can't Tell

C. Attribution

1. Did the evaluation attribute some result to some aspect of the  
project?

Yes \_\_\_\_\_ No \_\_\_\_\_

2. If the evaluation made such a statement, was the proof:

Adequate:	0	1	2	3	4
Unbiased:	0	1	2	3	4
Valid:	0	1	2	3	4

3. To what extent were exogenous variables (price, self-selection, initial economic order) examined?

0            1            2            3            4

4. To what extent were exogenous variables responsible for project achievements/failures?

0            1            2            3            4

5. Were exogenous variables examined in the evaluation?

Yes \_\_\_\_\_ No \_\_\_\_\_

If yes, list: 1. \_\_\_\_\_  
2. \_\_\_\_\_  
3. \_\_\_\_\_

6. Did the evaluators come to a conclusion about the project's sustainability?

Yes \_\_\_\_\_ No \_\_\_\_\_

7. If the evaluators came to a conclusion, was the project considered sustainable?

0            1            2            3            4

Summary score on element:

6.25 x A(2) Score \_\_\_\_\_ or 2.27 x (B(1) + B(2) + C Score)

SCORING INSTRUMENT

Attachment 1

CHARACTERISTIC I:

Subfactor 1: Ap \_\_\_\_\_ x 25.0 = \_\_\_\_\_ x .13 = \_\_\_\_\_

Subfactor 2: Summary Score for U Elements \_\_\_\_\_  
+ Summary Score for E Elements \_\_\_\_\_  
+ Summary Score for A Elements \_\_\_\_\_  
+ Summary Score for Output Elements \_\_\_\_\_  
+ Score for Input Elements \_\_\_\_\_  
= \_\_\_\_\_ + 5.0\* x .25 = \_\_\_\_\_

Subfactor 3:  
Score for MT element \_\_\_\_\_  
+ Score for H element \_\_\_\_\_  
= \_\_\_\_\_ + 2.0\* x .15 = \_\_\_\_\_

Subfactor 4:  
Co \_\_\_\_\_ + Cl \_\_\_\_\_ = \_\_\_\_\_ x 12.5 = \_\_\_\_\_ x .15 = \_\_\_\_\_

Subfactor 5:  
Co \_\_\_\_\_ + Cl \_\_\_\_\_ + Ap \_\_\_\_\_ = \_\_\_\_\_ x 8.33 = \_\_\_\_\_  
x .10 = \_\_\_\_\_

Subfactor 6:  
Co \_\_\_\_\_ + Cl \_\_\_\_\_ = \_\_\_\_\_ x 12.5 = \_\_\_\_\_ x .12 = \_\_\_\_\_

Subfactor 7:  
Co \_\_\_\_\_ + Cl \_\_\_\_\_ + Ap \_\_\_\_\_ = \_\_\_\_\_ x 8.33 = \_\_\_\_\_  
x .10 = \_\_\_\_\_

Total for Characteristic = \_\_\_\_\_

x .11 =

\* Precisely, by the number of elements present, which varies.

81

**CHARACTERISTIC II:**

Subfactor 1: Co      + Cl      =      x 12.5 =      x .43 =       
Subfactor 2: Co      + Cl      =      x 12.5 =      x .32 =       
Subfactor 3: Co      + Cl      =      x 12.5 =      x .25 =     

Total for Characteristic =                     

x .15 =

**CHARACTERISTIC III:**

Subfactor 1: Co      + Cl      =      x 12.5 =      x .39 =       
Subfactor 2: Co      + Cl      =      x 12.5 =      x .39 =       
Subfactor 3: Co      + Cl      =      x 12.5 =      x .22 =     

Total for Characteristic =                     

x .15 =

**CHARACTERISTIC IV:**

Subfactor 1: Co      + Cl      + Ap      =      x 8.33 =       
x .105 =       
Subfactor 2: Co      + Cl      + Ap      =      x 8.33 =       
x .105 =       
Subfactor 3: Co      + Cl      =      x 12.5 =      x .19 =       
Subfactor 4: Co      + Cl      + Ap      =      x 8.33 =       
x .19 =       
Subfactor 5: Ap      x 25.0 =      x .15 =       
Subfactor 6: Co      + Cl      + Ap      =      x 8.33 =       
x .10 =       
Subfactor 7: Co      + Cl      + Ap      =      x 8.33 =       
x .06 =       
Subfactor 8: Co      x 25.0 =      x .10 =     

Total for Characteristic =                     

x .09 =

28

**CHARACTERISTIC V:**

Subfactor 1: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .16 = \_\_\_  
 Subfactor 2: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .08 = \_\_\_  
 Subfactor 3: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .08 = \_\_\_  
 Subfactor 4: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .10 = \_\_\_  
 Subfactor 5: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .05 = \_\_\_  
 Subfactor 6: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .05 = \_\_\_  
 Subfactor 7: Cl \_\_\_ x 25.0 = \_\_\_ x .16 = \_\_\_  
 Subfactor 8: Cl \_\_\_ + Ap \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .16 = \_\_\_  
 Subfactor 9: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .16 = \_\_\_

Total for Characteristics = \_\_\_\_\_

x .11 =

**CHARACTERISTIC VI:**

Subfactor 1: Co \_\_\_ + Cl \_\_\_ + Ap \_\_\_ = \_\_\_ x 8.33 = \_\_\_  
 x .23 = \_\_\_\_\_  
 Subfactor 2: Ap \_\_\_ x 25.0 = \_\_\_ x .13 = \_\_\_\_\_  
 Subfactor 3: Co \_\_\_ + Cl \_\_\_ + Ap \_\_\_ = \_\_\_ x 8.33 = \_\_\_\_\_  
 x .13 = \_\_\_\_\_  
 Subfactor 4: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .13 = \_\_\_\_\_  
 Subfactor 5: Co \_\_\_ + Cl \_\_\_ + Ap \_\_\_ = \_\_\_ x 8.33 = \_\_\_\_\_  
 x .16 = \_\_\_\_\_  
 Subfactor 6: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .16 = \_\_\_\_\_  
 Subfactor 7: Co \_\_\_ + Cl \_\_\_ = \_\_\_ x 12.5 = \_\_\_ x .06 = \_\_\_\_\_

Total for Characteristic = \_\_\_\_\_

x .10 =

CHARACTERISTIC VII: Co \_\_\_\_\_ + Cl \_\_\_\_\_ = \_\_\_\_\_ x 12.5

Total for Characteristic = \_\_\_\_\_

x .10 =

51

CHARACTERISTIC VIII: Co \_\_\_\_\_ + Cl \_\_\_\_\_ = \_\_\_\_\_ x 12.5

Total for Characteristic = \_\_\_\_\_

x .10 =

CHARACTERISTIC IX: Co \_\_\_\_\_ + Cl \_\_\_\_\_ + Ap \_\_\_\_\_ = \_\_\_\_\_ x 8.33

Total for Characteristic = \_\_\_\_\_

x .09 =

SUMMARY (OVERVIEW) SCORE FOR REPORT

Weighted Score

- Characteristic I \_\_\_\_\_
- Characteristic II \_\_\_\_\_
- Characteristic III \_\_\_\_\_
- Characteristic IV \_\_\_\_\_
- Characteristic V \_\_\_\_\_
- Characteristic VI \_\_\_\_\_
- Characteristic VII \_\_\_\_\_
- Characteristic VIII \_\_\_\_\_
- Characteristic IX \_\_\_\_\_

Total Score =

#535 . . . =

20