FINAL REPORT

DEVELOPMENT OF A QUALITY/COMPLETENESS SCORING INSTRUMENT FOR USAID EVALUATION REPORTS

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

TRITON

TABLE OF CONTENTS

PHASE ONE SUMMARY

The first phase of this contract concentrated on identifying factors reflecting a "quality" evaluation report. These factors would subsequently be ranked through consultations with relevant persons outside AID and within the Agency, accomplished by sending selected individuals questionnaires containing statements about major quality factors, as well as subfactors within various major factors. Following this process and the determination of its results, TRITON would proceed to develop forms and numerical scoring tools (see subsequent discussions).

Quality and Completeness Factors

TRITON Corporation's initial identification of key quality and completeness indicators for AID evaluation reports was accomplished in essentially three stages. First, TRITON staff assigned to this project developed a list of factors they believed should ideally be found in

evaluation reports. These were divided into factors expressing "completeness" and "qualitative" (Exhibit A). Concurrent with this TRITON effort, the Program Evaluation Systems Division* within the Office of Evaluation/USAID prepared a set of statements about "good" evaluations (Exhibit B). Utilizing its own list and the PES list, TRITON integrated inputs into a working list of attribute statements. The presence of these attributes were intended to indicate that a report was complete and was of desirable "quality."

The next two stages of the project involved both a review and synthesis of evaluation literature from within AID and outside the Agency, and a series of interviews. These interviews, following the literature analysis, were conducted by telephone and in person with AID personnel associated with evaluations, individuals from other relevant agencies, and academicians (Exhibit C). Interviews were conducted with personnel of institutions such as the World Bank, ACTION, the Inter-American Foundation, and the American Council on Volunteer Agencies for Foreign Services.

Examples of some of the 20-30 reports, papers, etc. reviewed as part of the literature search include: Metaevaluation: Concepts, Standards and Uses, Daniel L. Stifflebean; and Standards for Program Evaluation (Exposure Draft), Evaluation Research Society, May 1980.

TRITON then prepared a report combining the results of these three stages in order to identify attributes of a "high quality" evaluation (Appendix I). This document, "Compilation of Attributes for Potential Use in Scoring Evaluation Reports," submitted October 14, 1981, describes in detail both the literature reviewed and interviews conducted. The compilation of attributes was then to serve as a basis for developing a scoring system for AID evaluation reports.

The Program Evaluation Systems Division performed a content analysis of the categories identified by TRITON. The purpose of this analysis was to

*Primarily Ms. Molly Hageboeck

EXHIBIT A

QUALITY FACTORS

Completeness Factors:

(1) Restatement of Log Frame*
(2) Evaluation Compared with Log Frame (e.g., inputs vs. outputs)
(3) Purpose (of Evaluation)
(4) Lessons Learned as a Result of Project
(5) Transferability of Experience

Qualitative Factors:

(1) Methodology - How was Evaluation Conducted
(2) Evaluation Justification
    (a) Appropriateness of Evaluation Method
    (b) Adaptability of Method
    (c) Acceptability of Method
    (d) Data Collection Procedures
(3) Logic of Methodology
(4) Analysis of the Methodology vis-a-vis Other Methods
(5) Appropriateness of Evaluation Method
(6) Timing
(7) Recommendations
(8) New Information Obtained from the Evaluation
(9) Significance
(10) Transferability
(11) Completeness of Evaluation Indicators
(12) Utility

---

*AID projects begin conceptually be developing a methodology know as the logical framework or "logframe." This logframe establishes the critical framework in a project, including project success and problems.

EXHIBIT B

WHAT MAKES AN EVALUATION A GOOD EVALUATION

1. It's legitimate--done for some comprehensible reason, even if that's just to meet a requirement.

2. It's focused--what is to be examined/learned is understood and its appropriate, given the stage of the project or program. The clearer the focus the better--since waste is associated with lack of clarity about what's needed.

3. It's methodology is appropriate--neither excessive or weak and the ways in which data was collected/analyzed are shared--no "magic."

4. It passes high on rules of evidence--unsubstantiated assertions and opinions aren't passed off as facts.

5. It uses evidence professionally--it neither hides findings nor reaches conclusions/recommendations for which there is no basis in the evidence accumulated by the evaluation.

6. It takes things to a bottom-line; i.e., follows a fact to its logical conclusion and spells our what needs to be done (or the options).

PPC/E/PES, Molly Hageboeck, 9/4/81--rough input for discussions with TRITON.

EXHIBIT C

PERSONS INTERVIEWED/CONTACTED

USAID Staff

Rich Rhoda
Nina Vreeland
Bob Berg
Bernice Goldstein
Emily Baldwin

External Interviewees

Santo Pietro, America Council on Volunteer Agencies for Foreign Service
Mary Anne Delancey, Consultants in Development
Jim Roberts, ACTION/Evaluation
Heather Clark, Inter-American Foundation
Jim Cotter

identify the major quality and completeness factors found in the TRITON data and to segregate a number of subfactors found with each major category (Exhibit D).

TRITON, in conjunction with PES, next used this master list to prepare a set of factors identified as relevant to determining the quality and completeness of an evaluation report, as distinct from the evaluation itself. Initially, the factors described on the master list were divided into two categories. All nine (9) factors which could be measured solely by reviewing the evaluation report were isolated into one category. Three (3) additional factors from the master list were characterized as factors which could not be measured exclusively from the evaluation report itself, but whose analysis would require additional information. Non-meaningful (non-measurable) factors were also deleted during this part of the research effort.

Factor Measurement and Ranking Process

Two concurrent activities were then undertaken. First, an iterative process was conducted between TRITON and PES staff to refine the factor and subfactor statements, to eliminate duplication and to coalesce all relevant attributes within the same factor. Next, all resulting statements were ranked in order of priority (relative importance). To objectively accomplish this, TRITON, in conjunction with AID/PES, initiated a limited Delphi exercise. In this exercise, TRITON identified individuals both within AID and outside of the Agency who would be provided with a questionnaire for ranking the factors (Exhibit E).

Each questionnaire consisted of four forms (Appendix II). Form 1 listed all statements identified as being characteristics of a high quality evaluation. The second form presented only those characteristics that TRITON and PES determined as assessable solely by reviewing the text of the evaluation report. Form 3 listed factors that cannot be adequately assessed by reading an evaluation report exclusively. Lastly, the fourth form identified subfactors associated with the previous key characteristics of an evaluation report.

EXHIBIT D

CATEGORIES OF STATEMENTS PRODUCED BY THE CONTENT ANALYSIS

- Characteristics of the Written Evaluation Report
- User Orientation/Focus in the Evaluation Study
- Adequate Caveats About Limitations of the Study
- Evaluation Timing and Costs
- Clear and Comprehensive Objectives Stated
- Potential Outcomes Considered Before Study Begins
- Evaluation Design/Overall Methodology
- Restrictions on the Use of Evaluation Data
- Data Collection Procedures/Processes
- Analysis Plan/Data Analysis Procedures
- Data Use/Treatment of Findings, Conclusions and Recommendations
- Actual Coverage/Scope of the Evaluation
- Value and Type of Information Produced by the Study
- Action and Other Implications of the Information

EXHIBIT E

PARTICIPANTS IN RANKING OF FACTORS

"PROVIDERS" OF REPORTS

Mr. Henry Miles
AFR/DP/PPEA
Agency for International Development

Ms. Bernice Goldstein
LAC/DP/PPE
Agency for International Development

Mr. Frank Campbell
S&T/PO
Agency for International Development

Mr. Robert Berg
AAA/PPC/E
Agency for International Development

Mr. Richard Blue
PPC/E/S
Agency for International Development

Mr. Twig Johnson
PPC/E/S
Agency for International Development

Mr. Steve Giddings
PRE/H
Agency for International Development

Ms. Barbara W. Searle
Operations Division, Educaton Section
International Bank for Reconstruction
and Development

Mr. Mike Wargo
Program Evaluation Staff
Department of Agriculture

Ms. Lois Ellen Data
Assistant Director
Education and Work Group
National Institute of Education

"USERS" OF REPORTS

Mr. Frank Kenefick
PPC/PDPR/PDI
Agency for International Development

Mr. Thomas McKee
LAC/DR
Agency for International Development

Mr. G.R. Van Raalte
ASIA/PD
Agency for International Development

Mr. Laurance Bond
AFR/DR/CCWAP
Agency for International Development
Department of State

Mr. Robert Bell
NE/PD
Agency for International Development

Mr. Richard K. Archi
PPC/PDPR
Agency for International Development

Former AID Directors

Mr. Joseph S. Toner

Mr. Gordon B. Ramsey

Mr. Stanley J. Siegel

Mr. Thomas Niblock

NOTE: Persons not currently or formerly with USAID were considered as the "external" group of respondents.

EXHIBIT E (Cont'd)

PARTICIPANTS IN RANKING OF FACTORS

"EXPERTS" IN EVALUATION DESIGN

Mr. Thomas D. Cook
Professor of Psychology
Northwestern University

Ms. Anita Weiss
Department of Sociology
University of California

Professor Robert Boruch
Department of Psychology
Northwestern University

Mr. Peter Rossi
Social and Demographic Research Institute
University of Massachusettes

Mr. Howard Freeman
Institute for Social Science Research
University of California

Mr. Herbert Turner
DIESA/PPCO
United Nations Headquarters

Mr. Michael Scriven
University of San Francisco, Calif.

Dr. Karl White
Exceptional Child Center
Utah State University

Each participant was asked to rank the order of all statements contained on each form. Once these responses were obtained, TRITON utilized this data to develop the numerical scoring process.

RESULTS OF RANKING PROCESS

A total of 34 persons were identified to participate in the ranking of quality factors and subfactors--22 currently or formerly with USAID and 12 representing external organizations/agencies. A total of 16 USAID-affiliated persons submitted completed responses for a 73% response rate, and 11 external persons responded (a 92% response rate).

In order to analyze the results, the rankings on each set of major factors and subfactors were assigned scores according to rank. For example, if there were seven factors to rank, then a factor ranked as the most important was scored as a 7, with the lowest ranked score yielding a 1. The scores for each form (set of factors or subfactors) were summed to yield the consensus of the respondents. For purposes of analysis and to identify any major inconsistencies among respondents in scores, the data was segmented by various groups of respondents:

- All respondents,
- USAID vs. external respondents,
- "Users" of USAID evaluation reports vs. "providers" of reports vs. "experts" in the field of evaluation design.

The results of this analysis is shown on the tables labeled as Exhibit F.

The key findings of this analysis revealed that:

- Those factors which cannot be adequately assessed by solely reviewing the evaluation report itself were ranked as the fourth, ninth and tenth most important factors out of the twelve (12) factors in Form 1. Thus, with the exception of the factor ranked fourth (relating to cost-effectiveness and timeliness of evaluation), it was felt by the project team that the instrument would be prividing input on quality factors that were perceived by the respondents as relatively important, while not being able to address factors that appeared to be relatively unimportant.

EXHIBIT F

SUMMARY OF ANALYSIS OF RESPONDENTS RANKINGS OF QUALITY FACTOR AND SUBFACTORS

FORM 1: RELATIVE PRIORITY OF ALL QUALITY FACTORS

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	180	1	64	3	87	1	72	2	42	5	179	2
2	44	11	52	6	58	7	63	7	48	3	112	10
3	97	5	60	4	78	3	47	10	48	4	169	4
4	122	2	67	2	82	2	75	1	51	2	235	1
5	84	7	28	12	72	5	29	11	28	11	125	9
6	73	10	43	10	53	9	57	8	21	12	83	11
7	91	6	78	1	77	4	66	4	55	1	172	3
8	76	9	43	9	30	12	51	9	36	7	143	7
9	40	12	55	5	39	11	27	12	39	6	125	8
10	82	8	49	8	58	8	65	5	32	8	78	12
11	105	4	34	11	71	6	65	6	31	9	144	6
12	115	3	50	7	52	10	72	3	30	10	154	5

FORM 2: RELATIVE PRIORITY QUALITY FACTORS THAT CAN BE ASSESSED BY CONDUCTING A REVIEW OF A WRITTEN EVALUATION REPORT

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	89	1	45	2	67	1	58	1	31	4	126	1
2	86	2	46	1	65	2	57	2	37	2	125	2
3	64	6	38	6	52	3	24	9	28	5	111	5
4	65	5	27	9	50	4	40	7	43	1	103	6
5	58	7	37	7	28	8	46	5	26	6	73	9
6	41	9	42	3	24	9	28	8	31	3	102	7
7	69	4	42	4	40	7	51	3	24	7	125	3
8	84	3	32	8	41	6	50	4	14	9	111	4
9	49	8	39	5	46	5	41	6	24	8	76	8
10												
11												
12												

FORM 3: RELATIVE PRIORITY OF QUALITY FACTORS
ABOUT WHICH AN EVALUATION REPORT MAY BE SILENT

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	30	3	25	1	26	3	24	1	22	1	55	2
2	39	1	19	2	31	1	20	4	19	2	58	1
3	25	4	19	3	15	4	23	2	11	3	44	4
4	36	2	15	4	28	2	21	3	10	4	51	3
5												
6												
7												
8												
9												
10												
11												
12												

FORM 4a: SUBFACTOR RANKINGS WITHIN CHARACTERISTICS
THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	37	1	20	1	25	1	26	1	14	1	57	1
2	26	2	17	2	21	2	17	2	11	2	43	2
3	21	3	13	3	14	3	14	3	11	3	34	3
4												
5												
6												
7												
8												
9												
10												
11												
12												

FORM 4b: SUBFACTOR RANKINGS WITHIN CHARACTERISTICS
THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT
(Cont'd)

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	26	2	16	2	20	2	18	2	13	1	31	3
2	36	1	19	1	24	1	24	1	12	2	42	1
3	18	3	13	3	19	3	12	3	11	3	35	2
4												
5												
6												
7												
8												
9												
10												
11												
12												

FORM 4c: SUBFACTOR RANKINGS WITHIN CHARACTERISTICS
THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT
(Cont'd)

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	45	4	31	4	34	4	38	3	25	4	69	5
2	85	2	43	1	62	2	59	1	41	1	132	1
3	89	1	36	3	64	1	36	4	32	2	127	2
4	39	5	43	2	33	5	48	2	26	3	76	3
5	37	7	23	6	30	6	16	7	10	7	52	7
6	46	3	25	5	35	3	28	5	20	5	75	4
7	39	6	13	7	21	7	23	6	14	6	62	6
8												
9												
10												
11												
12												

FORM 4d: SUBFACTOR RANKINGS WITHIN CHARACTERISTICS
THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT
(Cont'd)

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	85	1	46	1	60	1	51	1	41	1	127	1
2	83	2	30	5	52	3	41	3	30	2	111	3
3	65	3	44	2	53	2	48	2	29	3	115	2
4	41	5	16	7	34	5	33	5	8	7	70	5
5	60	4	31	4	47	4	36	4	20	5	91	4
6	35	6	22	6	14	7	17	7	18	6	39	7
7	23	7	35	3	30	6	24	6	28	4	63	6
8												
9												
10												
11												
12												

FORM 4e: SUBFACTOR RANKINGS WITHIN CHARACTERISTICS THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT
(Cont'd)

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	55	3	21	6	33	5	27	5	27	3	79	4
2	42	6	25	5	37	4	24	6	24	4	67	5
3	46	5	19	7	30	6	23	7	21	5	55	7
4	36	7	27	4	16	7	44	1	14	7	67	6
5	62	1	39	1	47	1	29	4	29	1	94	1
6	46	4	33	2	39	3	31	3	28	2	89	2
7	56	2	29	3	45	2	37	2	19	6	85	3
8												
9												
10												
11												
12												

FORM 4f: SUBFACTOR RANKINGS WITHIN CHARACTERISTICS THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT (Cont'd)

Rank / Statement	USAID		External		Providers		Users		Experts		Overall	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
1	78	1	47	1	42	1	57	1	47	1	125	1
2	49	4	31	5	30	5	36	5	27	2	85	4
3	39	6	32	4	36	4	29	6	25	3	70	5
4	46	5	23	6	29	6	49	2	23	5	69	6
5	59	2	36	3	42	2	40	4	19	6	91	3
6	54	3	37	2	38	3	41	3	23	4	91	2
7	25	7	14	7	17	7	19	7	14	7	39	7
8												
9												
10												
11												
12												

- 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 review by PES and TRITON staff of the few significant discrepancies indicated that such differences were, in most cases, due to the particular roles of the respondents and, hence, the perspective from which they view evaluation reports. The overall results led the project team to conclude that a second iteration of the limited Delphi technique was unnecessary.

- 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.

USE OF RANKINGS TO WEIGHT FACTORS AND SUBFACTORS

In order to translate the relative importance of factors reported by the respondents into quantitative values for the scoring instrument, each factor (out of 9 to be addressed by the instrument) and set of subfactors (one set for each of six of the factors) was assigned a weighting by:

- Clustering characteristics that received approximately the same score (sum of all respondents' rankings). In general, if statements had scores within 10 points, they were equalized.

- Summing the scores for all the characteristics on a given form; e.g., on Form 2:

(2) Factors x 125 = 250
(4) Factors x 105 = 420
(3) Factors x 75 = 225

Total = 895

- Assigning a normalized weight to each factor in proportion to its score's percentage of the total; e.g.,

Factor 3, Form 2 = 105 ÷ 895 = .11

- Rounding up or down the weighted values to insure the sum of such values equals 1.0 for each form (set of factors or subfactors).

The exact use of these weighted values (ranging from .06 to .43) in the scoring instrument is described subsequently.

DEVELOPING THE SCORING INSTRUMENT: FIRST DRAFT

Given the nine factors and their attendent subfactors, the next developmental step was to organize and structure the first draft of the scoring instrument itself. It was first determined that all but two of the subfactors for one of the factors could be scored in a similar manner--by assigning 0, 1, 2, 3, or 4 points on dimensions of completeness, clarity, and/or appropriateness. The definitions of these dimensions are shown in Exhibit G. For these characteristics, the reviewer would simply choose the appropriate score based on his/her perception of the evaluation report's standing on that characteristic and quality dimension.

Two subfactors dealing with the characteristic "the overall design of the evaluation is appropriate for answering the evaluation questions," were felt to require a more indepth approach to assessment. These dealt with: 1) the measurement procedures used by the evaluation and their validity, appropriateness, etc; and 2) the evaluation design's procedures for addressing hypothesized cause and effect linkages.

In order to assess these subfactors, worksheets and supporting materials were developed that:

- Identified planned objectives/effects, unplanned effects, assumptions/external factors, and management transformations/ hypotheses presented in the evaluation report.
- Scored each individual indicator employed in the evaluation report to address the above evaluation components according to such quality dimensions as validity, reliability, consistency, replicability and objectivity.

These materials (Appendix III, Attachments 2-7) result in normalized (0-100) scores for six (6) aspects of the evaluation report:

- For the subfactor dealing with measurement procedures:
  - Unplanned effects/results
  - Planned objectives/inputs/effects/results
  - External factors and assumptions;

EXHIBIT G

COMPLETENESS: Select the response that best reflects your perception of how completely the particular factor/topic/issue is addressed by the report:

0----------1----------2----------3----------4

Not addressed. Factor/topic/issue is totally absent.	Minimally addressed and/or addressed in a very superficial manner. Several key aspects of factor/topic/issue are not dealt with.	Most key aspects are addressed and in adequate detail.	All aspects are addressed and are adequately explored.

CLARITY: Select the response that best reflects your perception of how clearly the particular factor/topic/issue is addressed by the report:

0----------1----------2----------3----------4

Not clear. Can't understand point or concept that is being presented. Material not logically presented.	Can be understood, but reader has to "work" to determine point(s) being expressed. Not certain that understanding by reader corresponds to author's intent. Redundancy in presentation confusing. Presentation understandable but not logical.	Fully understandable. Expressed in very clear language. Reader is certain of author's points. Author fully conveys his/her thoughts.

APPROPRIATENESS: Select the response that best reflects your perception of how appropriately the particular factor/topic/issue is addressed by the report:

0----------1----------2----------3----------4

Totally inappropriate. Methods employed, analytical techniques, units of measure, statistical techniques, etc. are not appropriate for what is being analyzed, data being collected, and/or results being derived.	Generally addressed enappropriately, but selected aspects of the factor/topic/issue are appropriately analyzed, measured, etc.	Generally addressed in an appropriate manner but selected aspects (e.g., one of four units of measure) are not appropriately addressed.	Totally appropriate. The methodology, analyses, measurement tools, etc. are fully consistent with generally accepted principles and practices regarding evaluations and the particular factor/topic/issue being addressed.

- For the subfactor dealing with cause-effect linkages and hypothesis:
  - Processes/management transformations that cause an unplanned effect
  - Process/management transformations that caused a planned effect
  - Processes/management transformations that used an external factor to result in (or contribute to) the occurrence of an effect;

The 0-100 values for the former three components are combined to provide a score for the measurement subfactor, while the latter three values combine to provide a score for the hypotheses/linkage subfactor.

Computing an Overall Quality Score for an Evaluation Report

Appendix III depicts the complete first draft of the scoring instrument package. The steps involved in conducting a review are (keyed to the instrument's various components):

Step 1: Complete Attachment 1 for Characteristics I-VIII and Characteristic IX, subfactors 1 and 4-7, directly from reading the report. This form asks for scores of 0-4 on completeness, clarity and/or appropriateness (based on the scales described on the last page of Attachment 1) regarding various subfactors.

Step 2: To complete the scoring for Characteristic IX, Subfactors 2 and 3:

Step 2a. Complete one copy of Attachment 2, which enables the reviewer to "dissect," "diagram," and identify the key components of the evaluation: objectives, unplanned effects, planned effects, hypotheses, management transformations, etc. These are the O, A, and U-numeric and alpha elements discussed in this attachment.

Step 2b. Complete one copy of Attachment 3 for each O, A, U-numeric element identified by the reviewer on Attachment 2. This form scores each of these elements along various dimensions and criteria.*

The scores for these elements are sumarized (i.e., the results of all Attachment 3s completed) on Attachment 4; only one such attachment per evaluation is, therefore, filled out. This sheet enables the calculation of normalized scores for the O, A, and U-numeric element groups on a scale of 0-100. (A computation formula is included in this attachment.)

---

*Attachment 7 provides narrative material on such topics as validity, reliability, bias, objectivity, representation, adequacy and replicability, which are dimensions that must be scored by the reviewer.

Note that Item F of this attachment, Findings Analysis, is not part of the overall scoring system, but is designed to be part of the AID database on findings and their confidence levels.

Step 2c. Complete one copy of Attachment 5 for each O, A, U-alpha element identified by the reviewer on Attachment 2. This form scores each of these elements along various dimensions and criteria.*

The scores for these elements are summarized (i.e., the results of all Attachment 5s completed) on Attachment 6; only one such attachment per evaluation report is, therefore, filled out. This sheet enables the calculation of normalized scores for the O, A, and U-alpha element groups on a scale of 0-100/

Note that Item E of this attachment, Findings Analysis, is not part of the overall scoring system, but is designed to be part of the AID data base on findings and their confidence levels.

Step 3 The scores from Attachment 1 are entered on the Scoring Worksheet (Attachment 8) in the appropriate blanks (Co = Completeness, Cl = Clarity, Ap = Appropriateness) and the calculations shown on the worksheet are performed. These calculations result in:

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

TESTING THE FIRST DRAFT OF THE INSTRUMENT

In order to test the instrument, two TRITON staff consultants** were selected to each independently apply it to five USAID evaluation reports. These reports were:

I. Village Development, Bolivia, 1980;

II. Impact Evaluation of the Haiti Small Farmer Improvement Project, January 1979;

* Attachment 7 provides narrative material on such topics as validity, reliability, bias, objectivity, representation/adequacy and replicability, which are dimensions that must be scored by the reviewer.

**One has completed USAID's Evaluation Training Course.

III. Mid-Term Evaluation of the Primary Health Care Project, Kitui District, Kenya; August 1981;

IV. Assessment of the Lower Moulouya Irrigation Project, November 1981;

V. U.S. Assistance to the Family Planning and Population Program in Bangladesh, 1972-1980; April 1981 (Published).

The testing was performed to determine:

- Inter-rater reliability (i.e., how similar was the same report scored by the two reviewers);
- Absolute score levels among the reports, given general perceptions of the reports' relative quality;
- 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;
- Overall reviewer perceptions of instrument's usefulness and comprehensibility.

Summary data of the test's results is shown in Exhibit H. The key findings were:

- The scores appeared to be relatively high in general, averaging 79.5 with 4 scores above 80 and only 1 below 70.
- On three of five reports, there was a 5 point or less difference betweer the overall scores given by the two reviewers. However, there was a sharp difference on the remaining two reports: one rater scoring Report II as 93, the other as 59; conversely, Rater A gave Report V a 76, while Rater B scored it as a 97. Overall, the average scores of the two raters differed by 1 point. (See Table H-1.)
- The relative difference between rater scores for a given report ranged from a low of 0.5% to + 22.5% (Table H-1).

EXHIBIT H

SUMMARY OF INSTRUMENT TEST

H-1 Overall Scores

Report	Reviewer: A	Rank	Reviewer: B	Rank	Absolute Difference (A-B)	Relative Difference
I	70	5	75	3	5	+ 3.5%
II	93	1	59	5	34	+22.5%
III	89	2	90	2	1	+ 0.5%
IV	74	3	72	4	2	+ 1.5%
V	76	4	97	1	21	+12.0%
					12.6	+ 8.0%

H-2 Completion Times (minutes)

Report	Reviewer: A Reading	Reviewer: A Instrument	Reviewer: B Reading	Reviewer: B Instrument
I	120	120		105
II	60	70	60-150	150
III	50	120		90
IV	155	120		105
V	100	75		90
	97	101	105	108

H-3 Numerical Rankings Vs. General Rankings

Report	Reviewer: A NR	Reviewer: A Gen	Reviewer: B NR	Reviewer: B Gen
I	5	5	3	4
II	1	1	5	5
III	2	2	2	2
IV	3	3	4	3
V	4	4	1	1

EXHIBIT H

SUMMARY OF INSTRUMENT TEST

H-4 Overall Scores Vs. Sequence of Reports Ranked

	Reviewer A			Reviewer B		
Order of Review	Report Reviewed	Score		Report Reviewed	Score	
1	I	70		II	59	
2	III	89	+19	III	90	+31
3	V	76	-13	I	75	-15
4	IV	74	- 2	V	97	+22
5	II	93	+19	IV	72	-25

H-5 Summary of Inter-Rater Reliability Findings

	Average Absolute Difference	Maximum Absolute Differences	Average Relative Differences (Ranges)
Overall Scores	12.6 points	34.0 points	±8.0% (±.5% to ±22.5%)
Characteristics	2.0 points	6.0 points	±11.0% (± 8% to ±14%)
Subfactors	4.1 points	16.1 points	±16.0% (± 6% to ±31%)

EXHIBIT H (Cont'd)

SUMMARY OF INSTRUMENT TEST

H-6 Comparison of Weighted Scores For Each Characteristic

Report:	I		II		III		IV		V			
Characteristic	A	B	A	B	A	B	A	B	A	B	Average Difference	Relative Difference (%)
I	11.7	13.3	15.0	10.0	14.4	10.3	11.3	10.7	11.5	15.0	3.0	±12%
II	10.5	12.0	15.0	9.0	9.7	14.6	7.2	9.0	14.3	15.0	3.0	±13%
III	4.8	7.4	7.3	6.0	7.8	7.7	6.3	6.3	5.2	8.4	1.4	±10%
IV	6.7	6.6	10.2	7.6	9.9	10.6	9.4	7.4	8.7	11.0	1.5	± 8%
V	6.1	7.9	9.8	6.4	9.6	8.4	8.7	8.5	7.0	9.5	2.0	±12%
VI	6.6	8.3	9.6	4.1	11.0	11.0	8.3	8.3	9.6	10.0	1.7	±10%
VII	6.9	6.9	11.0	5.5	11.0	9.6	8.3	8.3	6.9	10.0	2.2	±13%
VIII	9.0	5.3	5.2	4.5	6.8	9.0	6.0	6.0	6.0	9.0	1.9	±14%
IX	7.7	7.0	9.6	6.2	9.1	8.4	8.2	7.7	6.6	9.3	1.6	±10%

EXHIBIT H (Cont'd)

SUMMARY OF INSTRUMENT TEST

H-7 Comparison of Weighted Scores For Each Subfactor

Report:	I		II		III		IV		V		Average Difference	Relative Difference (%)
	A	B	A	B	A	B	A	B	A	B		
Characteristic I:												
Subfactor 1:	43.0	37.6	43.0	26.9	43.0	43.0	21.5	21.5	26.9	43.0	7.3	±10%
Subfactor 2:	19.5	32.0	32.0	24.0	28.0	24.0	32.0	28.0	28.0	32.0	6.5	±11%
Subfactor 3:	15.6	18.8	25.0	15.6	25.0	21.9	21.9	21.9	21.9	25.0	3.8	± 9%
Characteristic II:												
Subfactor 1:	19.5	34.1	39.0	24.4	24.4	39.0	14.6	24.4	34.0	39.0	11.3	±19%
Subfactor 2:	34.1	29.3	39.0	24.4	29.3	39.0	19.5	24.4	39.0	39.0	6.8	±10%
Subfactor 3:	16.5	16.5	22.0	11.0	11.0	19.3	13.8	11.0	22.0	22.0	4.4	±13%
Characteristic III:												
Subfactor 1:	12.2	19.2	17.5	12.2	17.5	19.2	15.7	15.7	14.0	21.0	4.2	±13%
Subfactor 2:	9.5	11.9	19.0	11.9	0	14.3	11.9	14.3	9.5	14.3	6.2	±26%
Subfactor 3:	9.5	19.0	15.8	14.3	14.3	14.3	12.7	11.1	14.3	17.4	3.1	±10%
Subfactor 4:	11.3	11.3	11.3	7.5	15.0	11.3	11.3	7.5	7.5	15.0	4.2	±19%
Subfactor 5:	6.7	10.0	5.0	7.5	0	10.0	5.0	11.0	8.3	9.2	4.5	±31%
Subfactor 6:	4.5	5.5	3.0	3.5	0	6.0	3.6	6.0	4.5	6.0	2.3	±27%
Subfactor 7:	0	5.0	10.0	10.0	10.0	10.0	10.0	5.0	0	10.0	4.0	±27%

EXHIBIT H (Cont'd)

SUMMARY OF INSTRUMENT TEST

H-7 Comparison of Weighted Scores For Each Subfactor

Report:	I		II		III		IV		V			
	A	B	A	B	A	B	A	B	A	B	Average Difference	Relative Difference (%)
Characteristic IV:												
Subfactor 1:	8.0	12.0	16.0	12.0	16.0	16.0	12.0	12.0	14.0	16.0	2.8	±10%
Subfactor 2:	4.0	10.0	14.0	12.0	12.0	12.0	16.0	8.0	12.0	16.0	3.6	±15%
Subfactor 3:	10.0	5.0	10.0	7.5	10.0	10.0	10.0	7.5	8.8	10.0	2.2	±12%
Subfactor 4:	5.0	5.0	8.8	7.5	6.3	10.0	7.5	7.5	10.0	10.0	1.0	± 6%
Subfactor 5:	12.0	8.0	16.0	12.0	16.0	16.0	12.0	8.0	12.0	16.0	3.2	±12%
Subfactor 6:	8.0	12.0	16.0	10.0	14.0	16.0	12.0	12.0	10.0	16.0	3.6	±15%
Subfactor 7:	14.0	8.0	12.0	8.0	16.0	16.0	16.0	12.0	12.0	16.0	3.6	±14%
Characteristic V:												
Subfactor 1:	15.3	13.4	21.1	11.5	19.2	17.3	17.2	21.1	13.4	23.0	5.4	±15%
Subfactor 2:	6.5	6.5	9.8	6.5	13.0	9.8	9.8	9.8	6.5	13.0	2.6	±14%
Subfactor 3:	6.5	9.7	13.0	6.5	N/A	9.8	10.8	9.8	8.7	11.9	3.5	±18%
Subfactor 4:	9.8	9.8	13.0	8.1	9.8	9.8	11.4	9.8	9.8	13.0	1.9	± 9%
Subfactor 5:	10.7	14.7	14.7	9.3	N/A	12.0	12.0	10.7	10.6	16.0	4.0	±16%
Subfactor 6:	4.0	12.0	4.0	12.0	0	12.0	12.0	12.0	12.0	12.0	5.6	±30%
Subfactor 7:	3.0	5.3	3.8	4.5	6.0	6.0	6.0	4.5	3.0	6.0	1.7	±17%
Characteristic VI:	- No Subfactors -											
Characteristic VII:	- No Subfactors -											
Characteristic VIII:	- No Subfactors -											

EXHIBIT H (Cont'd)

SUMMARY OF INSTRUMENT TEST

H-7 Comparison of Weighted Scores For Each Subfactor

Report:	I		II		III		IV		V		Average Difference	Relative Difference (%)
	A	B	A	B	A	B	A	B	A	B		
Characteristic IX:												
Subfactor 1:	9.8	9.8	13.0	6.5	9.8	9.8	9.8	9.8	9.8	13.0	1.9	± 9%
Subfactor 2:	17.3	4.8	23.5	16.3	22.4	21.4	17.8	21.0	13.1	16.4	5.4	±15%
Subfactor 3:	11.0	7.1	12.1	9.0	N/A	9.3	N/A	11.3	5.5	10.3	3.9	±21%
Subfactor 4:	N/A	11.3	13.1	7.5	11.3	11.3	11.3	7.5	11.3	15.0	3.7	±17%
Subfactor 5:	N/A	10.0	N/A	5.0	N/A	8.3	N/A	6.7	N/A	9.2	N/A	N/A
Subfactor 6:	7.5	12.0	12.0	6.0	10.5	9.0	9.0	7.5	12.0	10.5	2.8	±15%
Subfactor 7:	N/A	8.3	5.0	5.8	N/A	7.5	8.3	5.8	2.5	10.0	3.6	±25%

- On average, approximately 3.33 hours were required to complete the scoring process for one report, split equally between reading the report itself and applying the instrument.(Table H-2).

- The reviewer's general assessments of the ranking of the five reports, independent of knowing the scores they had generated using the instrument, closely matched the quantitatively based rankings. Rater A's general and numerical-based ranking totally corresponded, Rater B's only transposed the third and fourth place rankings. (Table H-3)

- A general pattern appeared whereby the rater who reviewed a given report later in the sequence of his/her five reviews scored that report higher. (Table H-4).

- The weighted value scores for individual characteristics varied on the average between 1.6 and 3.0 points per characteristic which represents a ±8% to ±14% difference. For example, out of a high score of 15.0 for Characteristic I, the average score given was 12.3, with the average difference between the two rater's scorings being 3.0 points. This represents a ±12% range around the 12.3 figure. (Table H-6).

- At the subfactor level, the average absolute difference between rates scores for a given report was 4.1 points with a relative variation of of ±6% to ±31%. (Table H-5).

- The average relative difference between rater scores was smallest at the overall score level (±8.0%), somewhat larger (±11.0%) at the characteristic score level, and largest at the subfactor level (±16.0%). (Table H-5)

- There were large differences in the manner in which Subfactors 2 and 3 of Characteristic IX were scored (Appendix III, Attachments 3 and 5). The reviewers defined different numbers and types of outcomes, objectives, effects, etc. and generally had a difficult time in applying these portions of the instrument, due to the instrument's wording and conceptual definitions, and the evaluation reports not addressing these concepts explicitly or in an organized manner.

In summary, the test results indicated that the nucleus of a useful, meaningful instrument had been developed, but that further refinement was necessary to clarify concepts, reduce application time, minimize differences in interpretation and eliminate any potential learning curve bias.

REVISING THE INSTRUMENT

In order to improve the effectiveness of the instrument, several meetings were held with relevant USAID and TRITON staff to ascertain the weak points of the first draft and causes for variations in interpretation of evaluation reports, the instrument itself, and in scores assigned.

EXHIBIT I

I-1 SUMMARY OF INSTRUMENT TEST

	Reviewer A			Reviewer B		
Report	Review 1	Review 2	Change	Review 1	Review 2	Change
I	70	63	-7	--	--	--
II	93	77	-16	59	72	+13
III	89	73	-16	90	67	-23
IV	--	--	--	72	58	-14
Avg.	84	71	-13	74	66	-8

Absolute Difference Between A and B

Report	Review 1	Review 2
II	34 points	5 points
III	1 point	6 points

NOTE: Report I had been first report reviewed by A.
Report II had been fifth report reviewed by A.

Report II had been first report reviewed by B.
Report IV had been fifth report reviewed by B.

Second review of Report III was done with Characteristic IX being scored first, followed by Characteristics I-VIII.

Exhibit I (Cont'd)

SUMMARY OF INSTRUMENT TEST

I-2: Comparison of Weighted Scores For Each Subfactor

Report:	Ia	IIa		IIIa		IVa	
	A	A	B	A	B	A	B
Characteristic I:							
Subfactor 1	26.9	43.0	37.6	32.3	32.3		21.5
Subfactor 2	26.9	28.0	24.0	26.9	20.0		24.0
Subfactor 3	12.5	18.8	12.5	15.6	12.5		12.5
Characteristic II:							
Subfactor 1	29.3	19.5	24.4	19.5	24.4		24.4
Subfactor 2	29.3	24.4	24.4	24.4	19.5		19.5
Subfactor 3	16.5	13.8	13.8	16.5	11.0		13.8
Characteristic III:							
Subfactor 1	12.2	19.2	15.8	15.7	15.8		15.8
Subfactor 2	4.8	14.3	14.3	16.6	14.3		11.9
Subfactor 3	12.7	15.8	14.3	14.2	11.1		11.1
Subfactor 4	7.5	15.0	11.3	11.3	7.5		7.5
Subfactor 5	5.0	7.5	10.0	6.7	6.7		5.0
Subfactor 6	3.0	4.5	6.0	3.5	4.5		3.0
Subfactor 7	2.5	10.0	10.0	10.0	7.5		5.0

NOTE: "a" indicat second scoring of evaluation report in question.

Exhibit I (Cont'd)

SUMMARY OF INSTRUMENT TEST

I-2: Comparison of Weighted Scores For Each Subfactor

Report:	Ia		IIa		IIIa		IVa	
	A	B	A	B	A	B	A	B
Characteristic IV:								
Subfactor 1	8.0		10.0	12.0	12.0	12.0		10.0
Subfactor 2	10.0		12.0	10.0	10.0	10.0		8.0
Subfactor 3	6.3		6.3	8.8	7.5	7.5		7.5
Subfactor 4	2.5		6.3	7.5	5.0	7.5		5.0
Subfactor 5	8.0		12.0	16.0	12.0	12.0		8.0
Subfactor 6	8.0		12.0	14.0	12.0	12.0		10.0
Subfactor 7	12.0		16.0	16.0	12.0	10.0		10.0
Characteristic V:								
Subfactor 1	11.5		21.1	21.1	19.2	17.3		15.3
Subfactor 2	6.5		13.0	13.0	9.8	6.5		9.8
Subfactor 3	5.4		10.8	9.8	9.8	8.7		9.7
Subfactor 4	8.1		13.0	9.8	9.8	9.8		8.1
Subfactor 5	9.3		13.3	10.7	10.7	12.0		9.3
Subfactor 6	8.0		12.0	14.0	14.0	12.0		8.0
Subfactor 7	2.3		5.3	3.0	6.0	4.5		4.5

Exhibit I (Cont'd)

SUMMARY OF INSTRUMENT TEST

I-2: Comparison of Weighted Scores For Each Subfactor

Report:	Ia		IIa		IIIa		IVa	
	A	B	A	B	A	B	A	B
Characteristic IX:								
Subfactor 1	6.5		9.8	13.0	9.8	9.8		9.8
Subfactor 2	23.3		23.8	16.4	22.9	16.4		11.4
Subfactor 3	11.3		11.0	7.5	9.5	11.5		8.7
Subfactor 4	3.8		11.3	9.4	13.1	7.5		7.5
Subfactor 5	N/A		N/A	6.7	N/A	7.5		5.0
Subfactor 6	7.5		10.5	9.0	12.0	9.0		6.0
Subfactor 7	N/A		8.3	7.5	2.5	7.5		5.0

The first outcome of these meetings was to retest some of the evaluation reports based on feedback from the meetings. The results are shown in Exhibit I. A general reduction in absolute scores was observed. In addition, for the two evaluation reports rescored by both reviewers, the average difference in scores reduced from 17.5 points to 5.5 points. Lastly, the "learning curve bias" appeared to dissipate, with scores showing no pattern based on the sequence of review.

The second key outcome of the meetings was to determine the correlation between rater scores both on an overall report basis and on selected subfactors. These correlation factors are shown in the last column of Exhibit J. Keeping in mind that a "perfect" positive correlation between two scores would be an r value of +1.0, the report level scores indicate a very high correlation between the two raters. This, in turn, is considered to indicate a high level of interrater reliability.

At the subfactor level, there was a much wider range of correlations, as might be expected based on the absolute values involved. It was felt, however, that correlation of subfactors across several reports was not as meaningful or critical a factor as the interrater reliability measured between raters within a given evaluation report.

A third result of the meetings was to rearrange the characteristics a as set out in the instrument so that Characteristic IX, the one dealing with inputs, outputs, hypotheses, etc. and requiring the use of various worksheets, would become Characteristic I. This was done because scoring this characteristic requires the most detailed anaysis and review of the evaluation report. Hence, by completing it first, it would provide the scorer with the best framework for completing the remainder of the characteristics. Applications of this revised sequence indicated that such a strategy did give the reviewers a better feel for the reports and enabled more effective and efficient reviews of the documents.

Finally, the worksheet and supporting materials for scoring Characteristic I (previously Characteristic IX) were modified to improve the conciseness, consistency and clarity of the analysis required. This revised version is shown in Appendix IV.

EXHIBIT J

PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT FOR ASSESSING INTER-RATER RELIABILITY

	$A=X^2$	$B=X$	$C=XY$	$D=Y$	$E=Y^2$	n	r
Report I	6,546	364.3	6,707	415.3	7,919	31	+.79
Report IIa	8,583	471.6	7,886	446.9	7,481	33	+.91
Report IIIa	7,060	432.8	6,225	390.6	5,708	33	+.90
Report IV	6,318	414.4	6,072	398.1	6,218	32	+.84
Report V	7,925	426.9	9,648	563	12,322	33	+.93
Total Reports	36,432	2,110	36,538	2,214	39,648	162	+.84
Charac. II Subfactor 1	2,510	107.1	3,299	146	4,470	5	+.81
Charac. III Subfactor 5	240	34.2	318	46.9	451	5	-.34
Charac. IV Subfactor 4	246	33.8	266	37.5	294	5	+.82
Charac. V Subfactor 6	644	54	672	62	772	5	+.17
Charac. IX Subfactor 2	1,879	94.9	1,438	75	1,271	5	+.14
Charac. IX Subfactor 3	363	37	326.5	36.4	345	4	-.61

$$r = \frac{n(\text{sum } XY) - (\text{sum}X)(\text{sum}Y)}{\sqrt{[n \text{ sum}(X^2) - (\text{sum}X)^2][n \text{ sum}(Y^2) - (\text{sum}Y)^2]}}$$

TESTING THE REVISED INSTRUMENT

Based on the numerous meetings held between USAID and TRITON staff, the modifications made to the instrument, and the results of the test reviews, it was felt that the revised instrument could now be used to score a larer sample of reports. Forty (40) evaluation reports were selected by USAID staff to be scored using the revised instrument. The scorers were the same two TRITON staff who conducted the first round of tests.

The results of this second test are summarized in Exhibit K.

The scores ranged from 15 to 71 with an average score of 49 and a model score of 53. The clustering of scores was as follows:

Score	No.	% of Total
0-10	0	0
11-20	1	2.6
21-30	3	7.9
31-40	6	15.8
41-50	6	15.8
51-60	13	48.5
61-70	8	21.1
71-80	1	2.6
81-90	0	0
91-100	0	0

This represents a relatively "bell-curve" distribution without any skewed extreme clusters.

Lastly, the scoring experience of the two raters can be summarized accordingly:

	Rater A	Rater B
Range	25-71	15-62
Average	53	45
Mode	53	44

In general, this does not appear to exhibit the distinct rater tendencies originally observed.

Exhibit K

SUMMARY RESULTS OF RETEST

Project Title	USAID Project No.	Mission/AID/W Office	Score
anaging Decentralization Project	931-1053	ST/RAD	15
ural Development Planning	511-0471	Bolivia	23
ntegral Rural Development*	515-0158		25
ujarat Medium Irrigation	386-0464	India	25
orthern Sumatra Regional Planning	497-0246	Indonesia	32
itle II Food for Peace	not provided	Lesotho	34
RT: Soybean Milling	698-0407.08	Botswana	35
-Fixation Problems & Limiting Factors	931-0610	ST/AGR/RNR	37
ajasthan Medium Irrigation	386-0467	India	37
mproved Nutritional Quality of Wheat	931-0471.11	S&T/AGR	38
nhancing S&T Capabilities in LDCs**	931-1223	AID/SCI	41
ne Consequences of Small Farm Mechanization	931-1026	ST/AGR/EPP	42
ntente Food Production & Entente Livestock II	676-11-130-0203 & 676-11-130-0204	REDSO/WA Abidjah	43
nall Farm Production Systems	596-0083	ROCAP	45
jibuoti Fisheries Development Project	603-0003	Djibouti	49
raining of Paramadical Auxiliary & Community Personnel (PACs), Asia***	932-0644	ST/POP/TI	49
rakish Water Fisheries Production Project	G-497-0236	Indonesia	50
ater Resources & Soils Analysis	603-0001	Djibouti	52
echnical Health Institute	276-0019	Syria	53
oor Rural Households, Technical Change & Income Dist. in LDC's	931-0594	S&T/AGR	53

SUMMARY RESULTS OF RETEST (Continued)

Project Title	USAID Project No.	Mission/AID/W Office	Score
uras Federation of Industrial operatives (FEHCIL)	522-0179	Honduras	54
ng District Health Project	621-0138	Tanzania	56
wable Energy Technology	632-0206	Lesotho	56
Community Personnel Training oject, Cairo	263-0136	NE/TECH/HPN	57
l Rural Water Systems in Yemen	279-0044	NE/PD/NENA	57
culture Research	621-0107	Tanzania	58
-PROG Small Ruminants	931-1328	DS/AGR	58
uation of Title II: od for Peace Ghana	IQC AID/SOD/PDC-C-0262		59
rshed Management	525-0191	Panama	59
l Enterprises II	527-0176	Peru	60
al Progress Indicators - Salvador	931-0236.05	S&T/AGR/EPP	60
n Upgrading	522-0155	Honduras	62
ility Impacts of Develoment ograms in Asia	498-0270 (was 930-0068)	PPC/PDPR/HR	62
ria Control	386-0455	India	62
uctive Credit Guarantee Program	511-0486	Bolivia	62
tho Credit Union League velopment	632-0214	Lesotho	68
Industrial Export Promotion	522-0120	Honduras	69
l Farm Organizations	511-0452/511-T-055	Bolivia	71

Spanish-language evaluation with short English PES.
Evaluation Update only.
PES only.

CONCLUSION

Based upon the extensive iterative process discussed above to develop quality and completeness factors/criteria, dimensions, weightings and scoring instruments, the final result of the project appears to have achieved its original objectives. 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 is to apply the revised instrument to a large array of USAID evaluation reports and to conduct appropriate analyses of scoring trends and patterns by such variables as:

- Characteristics
- Subfactor
- Type of evaluation
- Mission/Office
- Evaluator (in-house vs. contractor).

APPENDIX I

COMPILATION OF ATTRIBUTES FOR POTENTIAL USE IN SCORING AID EVALUATION REPORTS

OVERVIEW

In order to develop the preliminary criteria by which to score/evaluate AID evaluation reports, TRITON embarked upon three approaches to compiling appropriate factors. These were:

- Developing criteria based on the project staff's own experience with evaluation reports, independent of the particular characteristics of AID evaluation reports. This array of criteria was synthesized with Ms. Hageboeck's similar delineation of attributes of a "good" evaluation, since these two perspectives turned out to be highly corroborative of one another.

- Obtaining criteria from various AID staff who are routinely involved in the preparation, review and use of evaluation reports.

- Reviewing relevant literature and contacting appropriate experts in the field (academia, World Bank, etc.) for their perspectives regarding criteria for "metaevaluation" (evaluation of evaluations).

This report summarizes the results of those three efforts at compiling the attributes of a "good" evaluation. It is intended to serve as a basis for refining the list of criteria and enhancing their specifity (in order to insure optimal objectivity). This refined list must then be prioritized/ weighted in order to proceed with a quantitative scoring system.

ATTRIBUTES OF A GOOD EVALUATION BASED ON TRITON ANALYSIS*

1. The evaluation methodology/strategy should be clearly and logically restated.

2. The evaluation methodology/strategy gives clear evidence of:

   a. Appropriateness - the methodology is appropriate given the nature and topic of the evaluation.

   b. Adaptability - the methodology has been modified ("tailored") to meet the needs of the specific project under study, and is not simply a "canned" approach from prior studies. Conversely, the methodology used can be adapted for evaluating similar projects in the future.

   c. Acceptability - the methodology "fits" the social, economic, political setting of the project; i.e., is acceptable to participants in the evaluation.

   d. Data procedures are appropriate; i.e., neither excessive or weak.

   e. Data procedures, collection and analyses are explicitly discussed, so that any ensuing conclusions and recommendations can be viewed in the context of how and what data was collected.

3. The legitimacy of the evaluation is explained; i.e., that it was done for some comprehensible reason, even if the reason is just to meet a requirement.

---

* Incorporating the input of M. Hageboeck.

a. Reasons for evaluation are given gradation; e.g., lessons for others to make decisions for future project actions (in order of importance).

4. The evaluation is focused. The report states what is being examined or is trying to be learned. The objective/purposes of the evaluation should be appropriate, given the stage of the project or program. Are the "right" questions being asked?

   a. Timing of evaluation.

5. The logic of the methodology makes sense, addresses the right unit of analysis and scale of project/program operations (e.g., individual farm, all farms in one local area, all farms in province, all farms in country.)

6. Besides the author's selected interpretation of the evaluation's results (inputs, outputs, causal links, etc.), the evaluation should discuss what alternative interpretations were considered and why they were not chosen.

7. The conclusions drawn are based on the evidence presented, which, in turn, is discussed in terms of how the evidence was collected, shortcomings, etc. Conversely, unsubstantiated assertions and opinions aren't passed off as facts.

   a. Both quantitative and qualitative evidence is presented and discussed.

   b. The evidence is used professionally. The report neither "hides" findings nor includes conclusions/recommendations for which there is no basis in the evidence accumulated by the evaluation.

8. The logic of the valuation is complete; i.e. examines causal links and assumptions between inputs, outputs, goals, purposes.

9. Recommendations are both introspective - providing insight into future action regarding the program/project under study - and outward-focused - providing insight into future action regarding other programs/projects.

10. Full use/exploitation is made of existing data.

11. The evaluation produces and presents new, meaningful information about the topic being addressed.

12. The findings are "significant" and trivial conclusions are avoided. The results of the evaluation, if followed, would appear to make a meaningful impact.

13. The evaluation methodology incorporates cost-benefit analyses of the project/program.

14. The evaluation itself proved to be cost-effective and was done on time.

15. The evaluation is useable, in terms of its utility to the intended audience (e.g., actionable recommendations).

16. The evaluation and its outcomes have transferability, external validity.

17. The evaluation takes things to a "bottom-line," i.e., follows a fact to its logical conclusion and spells out what needs to be done (or presents the options with their pros and cons).

ATTRIBUTES OF A GOOD EVALUATION
BASED ON AID INTERVIEWEES

Re: PES

1. The evaluation gives leading indicators of change (signals); i.e., is the project developmentally right or wrong.

2. Displays a clear understanding of whether implementing parties are executing properly.

3. If evaluation is conducted towards end of project, does it indicate whether project will "fly" without external parties being permanently involved.

4. Insightful "action"-oriented follow-up: what does this (evaluation recommendations/findings) mean to mission's program.

5. Lesson's learned are appropriate in scope, not "grandiose" or exhibiting unnecessary universalism.

Re: IMPACT STUDIES

1. In terms of presentation, the evaluation report relates text to appendices.

2. Gives an indication of how AID, as an institution, performed and how to improve AID's organizational performance.

Re: EOP STUDIES

1. Provides comprehensive "final tallies" of project results.

2. Denotes whether an explicit decision to "leave" project was because: 1) project could now operate on its own; or 2) ran out of money/time.

OVERALL

1. Minimized "buckslipping" (just referring to contractors report, etc.)

2. Objectivity - not slanted; doesn't just "blame" contractor; discusses mission's performance.

3. Usefulness of lessons learned - broader applications.

4. Discusses affect project is having on beneficiaries.

5. Creative use of information/data, fully exploits available information; checks sources.

6. Usefulness, "marketibility" to decision makers

7. Specific purpose of the evaluation is stated up front; objectives are well directed; appropriate timing and scope.

8. Adherence of evaluation to log frame - do linkages still make sense.

9. Identifies new directions for project itself which are realistic; what are constraints upon project's future performance.

10. Evaluation addresses specific items in project design. Did it ask all the "right" questions?

11. Evaluation is not just a "status" report; goes beyond inputs to look at impact (expected, unexpected, social, economic)

12. Talks about beneficiaries.

13. Analysis about implementation that is bringing about observed impact.

14. Compares original project design to how it has worked.

15. Self-contained document.

16. Addresses intervening variables; provides logic that supports the the contention that AID project was a facilitating factor.

17. Usability

18. Clear delineation of what was <u>important</u> in what evaluation discovered.

19. Clear statement of <u>why</u> evaluation was done
    a. Reason
    b. Need
    c. To support what? Operating mission/unit, agency as whole (programs), budgeting, etc.

20. Focused conclusions/recommendations/findings, tied to evaluation' purpose.
    a. Deal with impact issues not just technical/administrative issues.

21. Objective, credible - both on "nuts and bolts" and macro level issues

22. User-oriented focus.

23. For EOPS, did original outputs come about?

24. Candor: Did project overspend?
Analysis of time-budget performance.
Cost-benefit, internal-rate-of-return analyses
Purpose level (or goal level) progress

25. Not "too much" or "too little" data collection

26. Address what's between output and purpose; i.e., creating functioning systems: output → used by beneficiaries → purpose.

ATTRIBUTES OF A GOOD EVALUATION BASED ON AID TRAINING

CHECKLIST FOR AN EVALUATION STUDY

Objectives

1. The evaluation study (not the project) objective is stated.
2. The study provides new (and needed) information; a new method; technique; procedures; policy.
3. The final results are important or significant for the project or program. They change some policy or way of doing things. They confirm validity of earlier expectations, given the cost of the study.

Methods

1. Are the techniques, instruments, or modes of inquiry appropriate to the study design in the foreign context?
2. Have the methods been adapted to local conditions? Did this adaptation reduce the validity of the design?
3. Were there sampling problems? Are they clearly addressed?
4. If interviewing or opinion-survey techniques were used, were the questions meaningful in the local language and culture; in good taste; displayed political sensitivity; avoided religious connotation; addressed language problems?
5. Did the methods gather more or less data than required?

Data Processing

1. Are the procedures for the statistical manipulation of the data stated clearly? Is there a clearly conceived plan for the analysis that was performed in the data collected?

2. Do the analytical procedures produce meaningful statement?

Analysis and Interpretation

1. Have a wide variety of potential findings been considered?
2. Does the logic or design of the study permit clearly stated generalizations?

Costs

1. Are the total costs proportional to the scope or importance of the study? Is the study worth the cost?

General

1. Does the study answer the questions it set out to answer?
2. Does it produce explicit and usable results?
3. Does the study state what should now happen as a result of the study's findings?

ATTRIBUTES OF A GOOD EVALUATION
BASED ON SELECTED LITERATURE

"Metaevaluation: Concepts, Standards and Uses Daniel L. Stufflebeam;
"Educational Evaluation Methodology: The State of the Art (1981)

STANDARDS FOR METAEVALUATION

I. Utility: informative, timely, influential
- a - audience identification
- b - evaluator credibility
- c - information scope and selection
- d - valuational interpretation
- e - report clarity
- f - report timeliness
- g - evaluation impact

II. Feasibility: recognize natural setting of study; realistic, prudent, diplomatic, frugal
- a - Practical procedures
- b - political viability
- c - cost effectiveness

III. Propriety: legal, ethical, due regard for welfare of participants
- a - formal obligation
- b - conflict of interest
- c - full and frank disclosure
- d - public's right to know
- e - human interactions
- f - balanced reporting
- g - fiscal responsibility

IV. Accuracy: Obtained information should be technically adequate and that conclusions are linked logically to the data

a - object identification
b - context analysis
c - defensible information sources
d - described purposes
e - valid measurement
f - reliable measurement
g - systematic data control
h - analysis of quantitative information
i - analysis of qualitative information
j - justified conclusions
k - objective reporting

"Standards for Program Evaluation" (Exposure Draft); Evaluation Research Society, May 1980

FORMULATION

The evaluation report should insure that the audience for the report has a clear understanding of what was done, how it was to be done, and why, and an appreciation of constraints or impediments.

1. The purposes and characteristics of the program or activity addressed in the evaluation effort should be specified as precisely as possible.

2. The clients, relevant decisionmakers, and potential users of the evaluation results should be indentified, and their inforamtion needs and expectations made clear. Where appropriate, evaluators should also help identify areas of public interest in the program.

3. The type of evaluation effort undertaken should be identified and its objectives made clear; the range of activities undertaken should be specified.

4. An estimate of the cost of the evaluation effort should be provided.

5. The report should present evidence that the evaluation produced information of sufficient value, applicability, and potential for no utilization to justify the resources used.

6. Restrictions, if any, on access to the data and results from an evaluation should be clearly stated.

7. Conflicts of interest should be identified as well as the steps taken to avoid compromising the evaluation processes and results.

8. Respect for and protection of the rights and welfare of all parties to the evaluation should be evident from the evaluation report.

STRUCTURE AND DESIGN

The design for any evaluation cannot be conceived in a vacuum. It is necessarily influenced by logistical, ethical, political, and fiscal concerns, and therefore must take these into account as well as methodological requirements. Designs will vary in rigor and not all instruments are equally objective. However, even with these broad variations, the following standards generally apply. (For example, the approach to a case study is as subject to specification as the design of an experimental study; the reliability of judgments is as much at issue as the reliability of objective tests.)

9. A clear approach or design should be specified and justified as appropriate to the types of conclusions and inferences drawn.

10. For impact studies, the central evaluation design problem of estimating the effects of non-treatment, and the choice of a particular method for accomplishing this, should be fully described and justified.

11. If sampling was used, the details of the sampling method (choice of unit, method of selection, time frame, etc.) should be described and justified, based on explicit analysis of requirements of the evaluation, including generalization beyond the population sampled.

12. The measurement methods and instruments should be specified and described, and their reliability and validity of application to the characteristics to be measured should be estimated.

13. Justification should be provided that the best and most appropriate procedures and instruments have been utilized.

14. The report should address whether the necessary cooperation of program staff, affected institutions and members of the community, as well as those directly involved in the evaluation, was obtained.

DATA COLLECTION AND PREPARATION*

15. The data collection and preparation plan should be discussed.

16. Provisions made for the detection, reconciliation, and documentation of departures from the original design should be addressed.

17. Evidence should be presented that all data collection activities were conducted so that the rights, welfare, dignity, and worth of individuals were respected and protected.

18. The estimated validity and reliability of data collection instruments and procedures should be verified under the prevailing circumstances of their use.

19. Analysis of the source of error should be addressed as well as the provisions for quality assurance and control established to adequately meet the requirements of the overall design and anticipated data analyses.

20. The data collection and preparation procedures provided safeguards so that the findings and reports are not distorted by any biases of data collectors.

21. Data collection activities were conducted with minimum disruption to the program under study and with minimum imposition on the organizations or persons from whom data are gathered.

---

*Where secondary data are used, the evaluator should describe what is known about whether these standards have been met by the processes through which the data were originally produced.

22. Procedures that entailed adverse effects or risks were subjected to external independant review and then used only with informed consent of the parties affected.

DATA ANALYSIS AND INTERPRETATION

23. The analytic procedures matched the general purposes of the evaluation, the design, and the data collection.

24. All analytic procedures, along with their underlying assumptions and limitations, are described explicitly, and the reasons for choosing the procedures are clearly explained.

25. Analytic procedures were appropriate to the properties of the measures used and to the quality and quantity of the available data.

26. The units of analysis were appropriate to the way the data were collected and the types of conclusions to draw.

27. Justification is provided that the best and most appropriate analytic procedures have been applied.

28. Documentation is adequate to make the analyses replicable.

29. When quantitative comparisons are made (e.g., x is greater than y), tests of statistical significance are applied and interpretations stated with some indication of confidence.

30. Cause-and-effect interpretations are bolstered not only by reference to the design, but also by recognition and elimination of plausible rival explanations.

31. Findings are reported in a manner that distinguishes among objective findings, opinions, judgements, and speculation.

COMMUNICATION AND DISCLOSURE

Good communication is obviously essential to a well-formulated and executed evaluation report and to any utilization of the results. In particular, good communication is necessary to clarify the nature of the program, the expectations for the evaluation, and even the type of evaluation effort required, and to distinguish clearly objective findings and other information

32. Findings are presented clearly, completely, and fairly.

33. Findings are organized and stated in language understandable by decisionmakers and other audiences, and any recommendations are clearly related to the findings.

34. Findings and recommendations are presented in a framework that indicates their relative importance.

35. Assumptions are explicitly acknowledged.

36. Limitations caused by constraints on time, resources, data availability, etc. are stated. (Suggestions should be included on how to study those issues and questions that need further study and encouragement or assistance in doing so should be offered).

37. Complete explanation and description of how findings and results were derived should be accesible.

38. The finished data base and associated documentation should be organized in a manner consistent with accessibility policies and procedures.

UTILIZATION

The usual reason for conducting an evaluation is a functional one: to help those affected to be better informed about the feasibility of undertaking the program, the reasonableness of evaluating it, the program operation and its effects, and the results of previous evaluation efforts. Utilization cannot be guaranteed, of course, but it will be more likely if careful attention is given to the information needs of the potential users of the results throughout all phases of the evaluation.

39. Evaluation results should be timely; i.e., available to appropriate users before relevant decisions must be made.

40. The report should try to anticipate and prevent misinterpretations and misuses of evaluative information.

41. The report should bring to the attention of decisionmakers and other relevant audiences suspected side effects--positive or negative-- of the evaluation process.

42. The report should clearly distinguish between the findings of the evaluation and any policy recommendations based on them.

43. In making recommendations about corrective courses of action, the report should indicate what is known as a basis for estimating the probable effectiveness and costs of the recommended courses of action.

REFERENCE SOURCES FOR COMPILATION OF EVALUATION ATTRIBUTE LISTS

Persons Contacted/Interviewed

Molly Hageboeck, AId
Rick Rhoda, AID
Bob Berg, AID
Nina Vreeland, AID
Bernice Goldstein, AID
Santo Pietro, American Council on Volunteer Agencies for Foreign Services
Mary Ann Dulaney, Consultants in Development
Jim Roberts, ACTION/Evaluation
Jim Cotter, Inter American Foundation

Articles Reviewed

"Standards for Program Evaluation"(Exposure Draft), Evaluation Research Society, May 1980.

Metaevaluation: Concepts, Standards and Uses; "Educational Evaluation Methodology; The State of the Art; Daniel Stufflebeam; 1981.

"Metaevaluation Rsearch; Evaluation Quarterly; Thomas Cook and Charles Gruder; February 1978.

"Overview: Internal and External Validity in an Experimental Design;" Donald Campbell and J.C. Stanley; 1966.

"Purposes and General Methods of Program Evaluation;" Sourse unknown.

"Planning Useful Evaluations," Leonard Rutman.

"Draft Guide for Program Evaluation Design and Meta-evaluation;" Michae Wargo, ACTION; 1977.

"ACTION's Evaluation Role;" ACTION; 1977.

APPENDIX II

QUESTIONNAIRE FOR RANKING QUALITY FACTORS AND SUBFACTORS

-DRAFT-

Dear :

The Office of Evaluation of the Agency for International Development is seeking to develop a procedure for assessing the quality of its evaluation work. AID and TRITON Corporation, working together, have reached a point in this effort where the judgements of a wide range of individuals are needed to establish the relative priority of a series of evaluation characteristics -- all of which have been identified by AID or by evaluation literature as aspects of "quality."

You have been identified as an expert in the field of evalutation and evaluations theory, and we would appreciate your assistance in assigning a level of priority to these characteristics. Naturally, your cooperation is voluntary, but we hope you will participate in this exercise and promptly complete the attached questionnaire.

In the attachments to this letter, you will find a more detailed explanation of the effort with certain checklists. We would appreciate your review of this material and completion of the forms. A stamped envelope has been enclosed for you to return your response.

Thank you for your cooperation and participation in this important task.

Sincerely,

Robert J. Berg

Dear :

The Agency for International Development, working with TRITON Corporation, is developing procedures to be used to review evaluation reports as these reports are completed.

During the first stage of this project, TRITON conducted interviews with AID staff and reviewed literature on evaluation standards and quality. The product of that process was a list of statements which, ideally, would be true for all evaluations that were of high quality. Quality is defined as having all the characteristics deemed important by AID or identified in the evaluation literature.

The initial list developed by TRITON has been examined by AID's Program Evaluation Systems Division in the Office of Evaluation. Working together, TRITON and PPC/E/PES have organized the list into clusters of factors, some of which have a series of associated subfactors. The list has also been annotated to note which factors can be reviewed by examining an evaluation report and which cannot.

Factors which cannot be reviewed by examining an evaluation report appear to be characteristics of the evaluation process itself. They could be properly reviewed only through observation and interviews with those for whom an evaluation was carried out and those who requested it.

Using the list of important characteristics of an evaluation, TRITON and PPC/E/PES are now engaged in a second stage of this project to develop a procedure to be used in reviewing AID evaluation reports. This step has two objectives:

1. Identify priorities among factors identified as key characteristics of a high quality evaluation; and

2. Develop an evaluation report review form to be used to record information on the strengths and weaknesses of AID evaluations in terms of those factors which can be assessed by reading a report.

The first part of this task depends upon the combined effort of many individuals within and beyond AID.

Consequently, PPC/E/PES and TRITON have decided that the most appropriate way to define priorities among key characteristics of a high quality evaluation is to ask a fairly wide range of individuals which factors they consider to be of highest priority. We would appreciate your assistance in this survey.

In the material on the following pages, please record your judgements concerning the relative priority statements.

1. Form 1 lists all of the statements identified as being characteristic of a high quality evaluation. It makes no distinction between characteristics which can and cannot be assessed solely by reading an evaluation report. On Form 1 please rank the order of all statements on the page.

2. Form 2 lists only those statements from Form 1 which PPC/E/PES and TRITON have determined can be assessed by reading an evaluation report. You are asked to rank the order of all statements about evaluations that fall into this category.

3. Form 3 lists only those statements from Form 1 which PPC/E/PES and TRITON have determined cannot be adequately assessed by reading an evaluation report. You are asked to rank order all statements that fall in this category.

4. Form 4 deals with sub-factors that PPC/E/PES and TRITON have identified as being associated with key characteristics of an evaluation which can be assessed through a review of an evaluation report. You are asked to rank the order of the sub-factors listed for number of the key characteristics you have already judged in terms of their relative priority.

In completing the four forms, please approach each one independently and in order. In assigning ranks not that the number "1" should always be

assigned to the top priority item. Further, you are asked not to assign the same number to two factors -- the rules of this exercise do not allow "ties." Each factor must be given a different number in your ranking. The only thing to consider in making rankings is your own judgement. The rankings will be compared once all copies of the forms are returned to TRITON.

While we do not at present expect that a second ranking will be needed to complete the effort to assign priorities, we may require further assistance should the rankings suggest significant conflicts among those who provide us with rankings. Consistent with this attempt to use a modified Delphi approach for assigning priorities, we will provide you with information in how others have ranked factors should we require your assistance in a second round of rankings.

Please return copies of all the forms, with your name printed at the top of each form, no later than January 12, 1982, to:

Mr. Sonny Bloom
TRITON Corporation
1825 Connecticut Avenue, N.W.
Suite 408
Washington, D.C. 20009

A stamped envelope is enclosed to return the forms.

Because of the schedule in completing this project, I appreciate your response no later than January 12, 1982. As I mentioned, we may provide you with the results of this ranking for further comment, should this be required. In addition, we will share the project results with you, if you would like them.

Sincerely,

Molly Hageboeck
Chief, Program Evaluation
Systems Division
Office of Evaluation
Agency for International Development

Test Name: _____

FORM 1: RELATIVE PRIORITY OF ALL QUALITY FACTORS

Please assign ranks to all of the statements listed below. The number "1" should be assigned to the factor which you consider to have the highest priority on the list. All factors on the list have been identified as being characteristics of a "high quality" evaluation.

RANK	STATEMENT OF QUALITY CHARACTERISTIC
___	The evaluation focuses on the evaluation users and their needs/questions.
___	The evaluation clearly identifies methodological limitations and other factors that limit the study as well as restrictions on the use of study data.
___	The evaluation is carried out in a timely and cost-effective manner that is appropriate to the stage of the project or program, its size and the need for evaluative evidence.
___	The evaluation clearly identifies the objectives of the project or program which is being evaluated as well as the evaluation objectives and questions.
___	The evaluation report is a well written, self-contained document.
___	The evaluators and those for whom the evaluation is conducted considered the possible evaluation outcomes and their implications before the evaluation began.
___	The overall design of the evaluation is appropriate for answering the evaluation questions.
___	The data collection procedures and/or use of secondary data are appropriate and adequate, not excessive or inappropriate.
___	The data analysis procedures are appropriate and adequa
___	Findings, conclusions and recommendations are presented in a way that clearly separates facts from interpretati
___	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.
___	Action implications of the evaluation are clearly state and are annotated to indicate who or what unit should a

Your Name: _____

FORM 2: RELATIVE PRIORITY OF QUALITY FACTORS THAT CAN BE ASSESSED BY CONDUCTING A REVIEW OF A WRITTEN EVALUATION REPORT

Please assign ranks to all of the statements below. The number "1" should be assigned to the factor which you consider to have the highest priority on the list. In assigning ranks on this form, your judgements need to be based on the assumption that while information on these factors will be available -- no other information will be accessable. Hence, your priorities should reflect what you would consider important if these were the only factors you could examine to judge evaluation "quality".

RANK	STATEMENT OF QUALITY CHARACTERISTIC
___	The evaluation focuses on the evaluation users and their needs/questions.
___	The evaluation clearly identifies the objectives of the project or program which is being evaluated as well as the evaluation objectives and questions.
___	The evaluation report is a well written, self-contained document.
___	The overall design of the evaluation is appropriate for answering the evaluation quesitons.
___	The data collection procedures and/or secondary data are appropriate and adequate, not excessive or inappropriate.
___	The data analysis procedures are appropriate and adequate
___	Findings, conclusions and recommendations are presented in a way that clearly separates facts from interpretation
___	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.
___	Action implications of the evaluation are clearly stated and are annotated to indicate who or what unit should act

Your Name: _____

FORM 3: RELATIVE PRIORITY OF QUALITY FACTORS ABOUT WHICH AN EVALUATION REPORT MAY BE SILENT

Please assign ranks to all of the statements below. The number "1" should be assigned to the factor which you consider to have the highest rank. In assigning ranks to statements, you may expect that an evaluation report may be silent on all of these factors -- they are part of the evaluation process or they are factors which, if not discussed in an evaluation report, cannot be inferred from the report.

RANK	STATEMENT OF QUALITY CHARACTERISTIC
___	The evaluation clearly identifies methodological limitations and other limits on the study as well as restrictions on the use of study data.
___	The evaluation is carried out in a timely and cost-effective manner that is appropriate to the stage of the project or program, its size and the need for evaluative evidence.
___	The evaluators and those for whom the evaluation is conducted considered the possible evaluation outcomes and their implications before the evaluation began.

Your Name: _____

FORM 4: SUB-FACTOR RANKINGS WITHIN CHARACTERISTICS THAT CAN BE ASSESSED BY REVIEWING AN EVALUATION REPORT

A number of the characteristics that can be assessed by examining an evaluation report have several components. On this form you are asked to rank order the components of several characteristics. The procedure you are asked to use parallels that used in prior forms. In this form each boxed items is to be treated independently, i.e., within each of the boxes a ranking of "1" will be assigned to the highest priority sub-factor. Other factors in the box will be ranked in order. No "ties" are permitted.

A

CHARACTERISTIC:	The evaluation clearly identifies the objectives of the project or program which is being evaluated as well as the evaluation objectives and questions.
RANK	SUB-FACTORS TO BE ASSESSED FOR THIS CHARACTERISTIC
___	Project or program objectives are clearly stated.
___	The objectives of the evaluation are clearly stated; priorities among reasons are clear.
___	The evaluation questions are clearly stated; priorities among questions are clear.

B

CHARACTERISTIC:	The evaluation focuses on the evaluation users and their needs/questions.
RANK	SUB-FACTORS TO BE ASSESSED FOR THIS CHARACTERISTIC
___	Evaluation clients/users are clearly identified
___	User needs/expectations are clearly identified
___	Areas of "public interest"/broad concern covered by the evaluation are clearly identified.

Form 4, continued

C

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

RANK	SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC
___	The units of analysis are appropriate given the evaluation questions.
___	As needed, the design contains measures of the presence/absence of treatments and/or effects and changes in treatements and/or effects. The measures are valid measures of concepts and they consider such factors as duration, intensity, etc. as required.
___	As needed, the design contains procedures for dealing with rival causal explanations, e.g., for assessing the effects of treatment and non-treatment. The procedures are legitimate given the type of causal explanations the evaluation considers.
___	Assumptions made by the design are clearly stated.
___	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.
___	The design is acceptable to those who are to be examined or considered by the evaluation; it is ethical in the sense that it respects the rights and welfare of all parties.
___	The evaluation design is fully and clearly described by the evaluation report.
___	The design includes procedures for recording any changes in the methodology as are made during the course of the evaluation and where such changes occur the evaluation report discusses them.

D

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

RANK	SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC
___	Instruments/approaches for collecting data are valid and reliable; validity and reliability of any secondary data is checked and found acceptable.
___	Sources of error, biases, in the instruments or data collection procedures are described as fully as possible
___	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.
___	Neither too much nor too little data is secured.
___	Where cross-cultural sensitivity, language etc. are potential issues, they are properly handled, e.g. local data collectors used, female data collectors, etc.
___	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.
___	Instruments used to collect raw data, such as questionnaires, are included as exibits to evaluation reports.

E

CHARACTERISTIC: The data analysis procedures are appropriate and adequate.

RANK	SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC
___	The analysis procedures match the purposes of the evaluation and fit the evaluation questions and data collected to answer those questions.
___	The analysis procedures are appropriate, they are neither weak nor excessive.
___	Where appropriate the confidence level of findings is given, e.g., statistical significance for 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 data analysis.
___	Both quantitative and qualitative data are analyzed if both were secured.
___	Where appropriate the evaluation examines the realism of the project's original estimates of cost, economic return, etc., as well as data on project/program effectiveness and impact
___	The strenghts and weaknesses of the data analysis aspects of the evaluation are clearly stated.
___	Where appropriate, the raw data from the study are included, or their availability made known, should it be necessary/appropriate to reanalyze all or part of the study data.

Form 4, continued

F

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

RANK	SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC
___	Facts are separated from interpretations.
___	Alternative interpretations are discussed and the reason for selecting a specific interpretation or conclusion is made clear.
___	Conclusions are separated from recommendations.
___	Alternative recommendations are discussed and the reason for selecting a specific recommendation is made clear.
___	The study findings, conclusions and recommendations are well organized and presented in a fashion that is understandable to a busy reader/decisionmaker who may not be familiar with how such studies are conducted.
___	The material on findings, conclusions and recommendations is objective in the sense that it neither "hides" data nor makes assertions without adequate facts.
___	The evaluators come to 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?

APPENDIX III

FIRST DRAFT OF
SCORING INSTRUMENT
(INCLUDING SUPPORTING WORKSHEETS)

ATTACHMENT 1

OVERALL SCORING INSTRUMENT

(with scales for Completeness, Clarity and Appropriateness)

CHARACTERISTIC I: 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 II: 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 III: 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; validity and reliability of any secondary data is checked and found acceptable.

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

2. 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

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.

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. Neither too much or too little data is secured.

Appropriateness:	0	1	2	3	4

5. 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

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.

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

7. 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 IV: 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 and the reason for selecting a specific interpretation or conclusion is made clear.

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

3. Conclusions are separated from recommendations.

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

4. 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

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.

   Clarity: 0 1 2 3 4

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.

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

7. The evaluators come to 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 V: 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
Clarity:	0	1	2	3	4
Appropriateness:	0	1	2	3	4

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 appropriate, 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 VI: The evaluation report is a well-written, self contained document.

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

CHARACTERISTIC VII: 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.

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

CHARACTERISTIC VIII: 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

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

SUB-FACTORS TO BE ADDRESSED FOR THIS CHARACTERISTIC

1. The units of analysis 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 O-numeric elements: _____

   Summary Score for A-numeric elements: _____

   Summary Score for U-numeric 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 O-alpha elements: _____

Summary Score for A-alpha elements: _____

Summary Score for U-alpha 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
Clarity:	0	1	2	3	4
Appropriateness:	0	1	2	3	4

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

COMPLETENESS: Select the response that best reflects your perception of how completely the particular factor/topic/issue is addressed by the report:

0----------1----------2----------3----------4

Not addressed. Factor/topic/issue is totally absent.	Minimally addressed and/or addressed in a very superficial manner. Several key aspects of factor/topic/issue are not dealt with.	Most key aspects are addressed and in adequate detail.	All aspects are addressed and are adequately explored.

CLARITY: Select the response that best reflects your perception of how clearly the particular factor/topic/issue is addressed by the report:

0----------1----------2----------3----------4

Not clear. Can't understand point or concept that is being presented. Material not logically presented.	Can be understood, but reader has to "work" to determine point(s) being expressed. Not certain that understanding by reader corresponds to author's intent. Redundancy in presentation confusing. Presentation understandable but not logical.	Fully understandable. Expressed in very clear language. Reader is certain of author's points. Author fully conveys his/her thoughts.

APPROPRIATENESS: Select the response that best reflects your perception of how appropriately the particular factor/topic/issue is addressed by the report:

0----------1----------2----------3----------4

Totally inappropriate. Methods employed, analytical techniques , units of measure, statistical techniques, etc. are not appropriate for what is being analyzed, data being collected, and/or results being derived.	Generally addressed enappropriately, but selected aspects of the factor/topic/issue are appropriately analyzed, measured, etc.	Generally addressed in an appropriate manner but selected aspects (e.g., one of four units of measure) are not appropriately addressed.	Totally appropriate. The methodology, analyses, measurement tools, etc. are fully consistent with generally accepted principles and practices regarding evaluations and the particular factor/topic/issue being addressed.

ATTACHMENT 2

WORKSHEET FOR SCORING CHARACTERISTIC IX

SUBFACTORS 2 AND 3

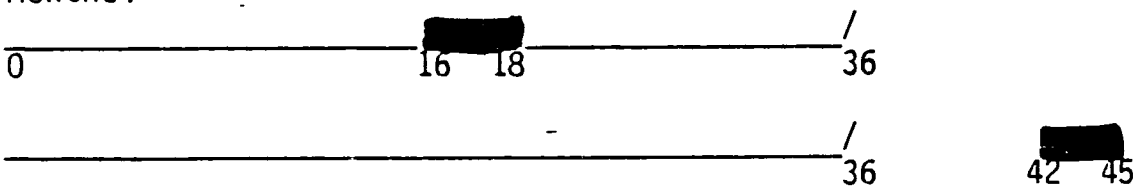
WORKSHEET FOR SCORING THE "APPROPRIATENESS" EVALUATION METHODS:

(Characteristic IX', Subfactors 2 and 3)

A. EVALUATION TIMING: Based on a review of the evaluation report, complete the following informational data:

1. Planned life of project/program/policy: _____ months
2. Period under evaluation: _____ months after start of project through _____ months after start.
3. Evaluation timing graph. (Follow examples given below):

Evaluation Timing Graph Examples:



B. EVALUATION COVERAGE

1. General purpose/character of the evaluation:
   - ___ Formative, primary focus of evaluation was replanning project/program.
   - ___ Summative, primary focus was outcome/impact assessment of project/program.
   - ___ Mixed (partially formative, partially summative)

C. IDENTIFICATION OF PLANNED OBJECTIVES/EFFECTS, UNPLANNED EFFECTS, ASSUMPTIONS/EXTERNAL FACTORS, MANAGEMENT TRANSFORMATIONS/HYPOTHESES

Using information from the evaluation report, the case diagramming example on page _, and the diagram on page , check all elements of the diagram which were covered by the evaluation:

___01	___0a	___A1	___Aa	___U1	___Ua
___02	___0b	___A2	___Ab	___U2	___Ub
___03	___0c	___A3	___Ac	___U3	___Uc
___04	___0d	___A4	___Ad	___U4	___Ud
___05	___0e	___A5	___Ae	___U5	___Ue
___06		___A6	___Af	___U6	___Uf

Note the following conventions:

U1,2... = An unplanned effect/result of the project or program.

01,2... = A planned objective or effect/result of the project. In terms of the logframe, these would be inputs, outputs, purpose(s), goal(s). These elements may also be thought of as including:

- the preparations for providing the independent variables (inputs)
- the independent variables (treatments)
- the dependent variables (outcomes).

A1,2... = An assumption that was made regarding significant external factors or conditions over which the project may have no control, but which are essential to successful project implementation.

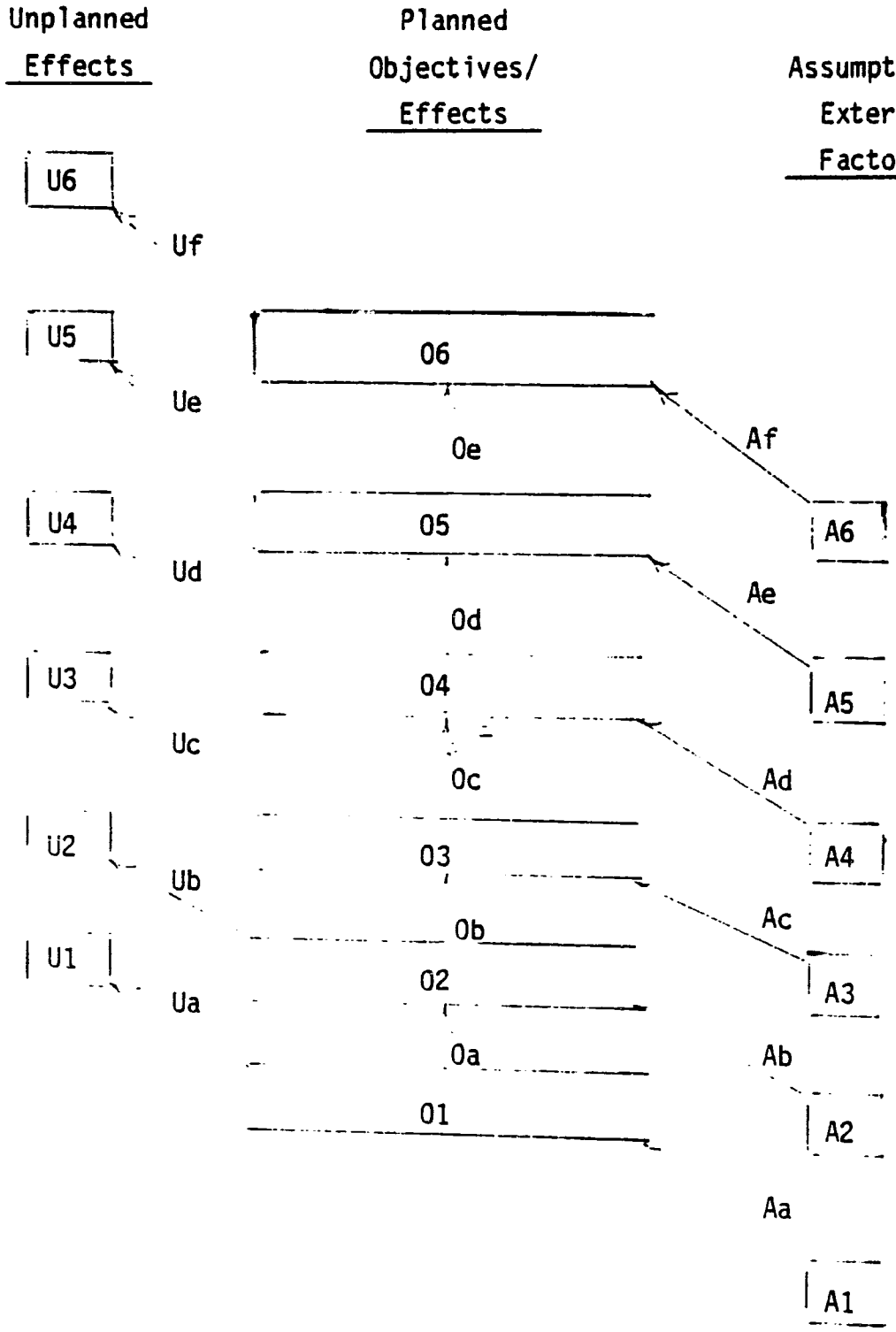
Ua,b,...= A process or management transformation that caused (or is hypothesized to cause) an unplanned effect (U1,2,...)

Oa,b,... = A process or management tranformation that caused (or is hypothesized to cause) a planned objective/effect.

Aa,b,... = A process or management transformation that caused (or is hypothesized to cause) an external factor to result in (or contribute to) the occurence of a planned effect.

For example, external factor A1, acts by some process (Aa) to influence a planned objective/inputs/ effects (O1). This planned objective/input/effect acts (or is hypothesized to act) by some process (Oa) to generate another planned objective/outcome/effect (O2). Planned objective O1 may also cause an unplanned outcome (U1) due by some process Ua.

DIAGRAM REPRESENTING SEQUENCE OF ELEMENTS IN THE CAUSAL CHAIN OF EVENTS ASSOCIATED WITH A PROJECT/PROGRAM BEING EVALUATED

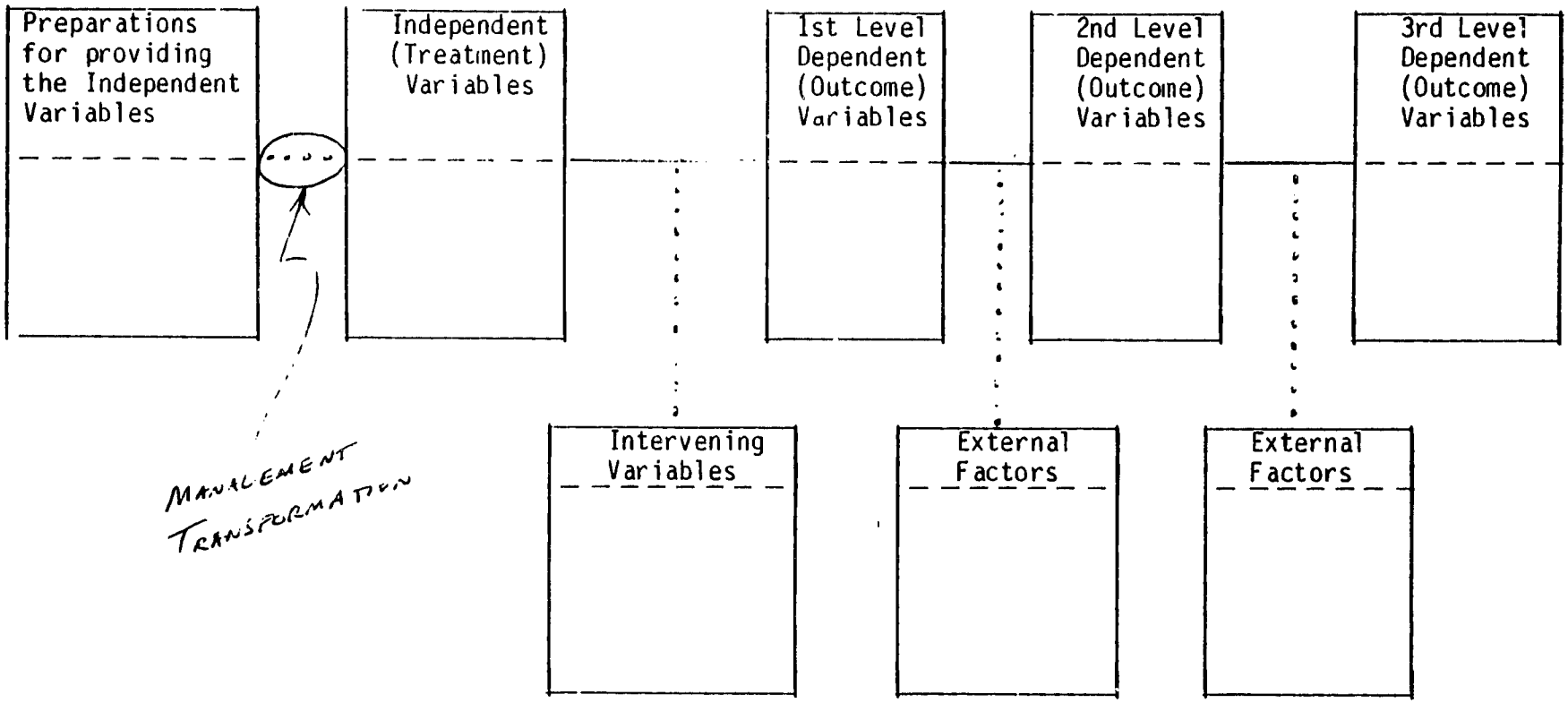


Causal process linking an external factor to a resulting planned objective/outcome/input.

Causal process that links a given planned objective/outcome/input to an unplanned effect/ result.

Causal process linking a planned input outcome/objective t a resulting subsequent outcome/objec tive/purpose.

DIAGRAM TO ASSIST IN SUMMARIZING KEY ELEMENTS OF PROJECT/PROGRAM BEING EVALUATED



The information required to fill in this chart can normally be found within the evaluation report's statement of project objectives or on the Logical Framework for the project, if that is provided. In filling out the chart: (1) First identify those factors/indicators in the development situation that the project is expcted to "change" in a causal way. List the immediate factors/variables of this type as 1st level dependent variables; e.g., farmer adoption of a new technology. If the project hypothesizes that the initial change "caused" by the project will lead to successive changes beyond the 1st level dependent variable, list these successive changes in sequence of occurence; e.g.: level 2: change in farm production; level 3: change in amount of farm production sold; level 4: increased income in farm households. The levels discussed here are not directly equated to levels in the project's Logical Framework due to differences in the way that tool is used and the fact that some levels of logic which must be stated in this chart may not be explicitly stated in project Logical Frameworks. In other words, the reviewer is expected to make such amendments to the logic as are needed to clarify it and to properly lay out the elements to be assessed by an evaluation. For definitions and more detailed descriptions of what to place in each of the on boxes the diagram, see the glossary and instructions at the end of the coding form.

ATTACHMENT 3

O, A, U-NUMERIC RATING FORM

COMPLETE ONE COPY OF THIS FORM FOR EACH O,A,U-NUMERIC ELEMENT NOTED IN ITEM C OF THE WORKSHEET.

Element being scored: _____ _____
(O,A,U) (numeric)

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 Numeric Element:

(1) Check which of these is applicable:

Ind 1	Ind 2	Ind 3	Ind 4	
__	__	__	__	Presence/absence(i.e., indicator was not present "before" activity being evaluated)
__	__	__	__	Change in status (i.e., indicator was present "before"; measure focuses on change)
__	__	__	__	Both (i.e., indicator was present "before" but not "after")

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

__	__	__	__	(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. Score 0 = No, 2 = Somewhat, 4 = Yes

__ __ __ __ (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

(4) Complete only if C (1) response = both. Score 0 = No. 2 = Somewhat, 4 = Yes

__ __ __ __ (a) Measure was valid measure of indicator which was to have changed/ existed
__ __ __ __ (b) Measures at all points were made in consistent manner
__ __ __ __ (c) Measure 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 from which a generalization was made:

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

E. Summary score on the finding/measure made:

Ind 1	Ind 2	Ind 3	Ind 4	Total for All Indicators	
__	__	__	__	__	Validity: Score from C(2)(a) or C(3)(a) or (C(4)(a)
__	__	__	__	__	Replicability/consistency: Score from C(2)(b) C(3)(b) or C(4)(b)
__	__	__	__	__	Bias: Score from C(2)(c) or C(3)(c) or C(4)(c)
__	__	__	__	__	Representativeness/Adequacy: Score form C(3)(d) or C(4)(d)
__	__	__	__	__	Objectivity: Score from C(2)(d) or C(3)(e) or C(4)(e)
__	__	__	__	__	Generalization: Score from Item D

F. Findings Analysis

(1) Status of indicators. Check for each indicator used to measure the variable being evaluated:

Ind 1	Ind 2	Ind 3	Ind 4	
__	__	__	__	Present (if only presence/absence was observed)
__	__	__	__	Positive change (if change was observed)
__	__	__	__	Negative change (if change was observed)
__	__	__	__	No change (if change was observed)

(2) Summarize the finding regarding this variable (1 or 2 sentences):

_____

_____

_____

(3) If C(1) response = presence/absence, 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						____

(4) If C(2) response = change in status, then 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						____

(5) Overall Confidence Level:
F(3) or F(4) Score + D Score = _____

ATTACHMENT 4

O, A, U-NUMERIC SUMMARY SCORING FORM

Element		Total Score From Worksheet Item E	Maximum Possible Score	Normalized Score*
01				
02				
03				
04				
05				
06				
	Total for all O-numeric elements:			
A1				
A2				
A3				
A4				
A5				
A6				
	Total for all A-numeric elements:			
U1				
U2				
U3				
U4				
U5				
U6				
	Total for all U-numeric elements:			

*See computation formula on next page.

To compute the Maximum Possible Score on Item E:

1. Determine the total number of criteria in Item E for which a score of 0, 2, or 4 was entered (e.g., if 2 indicators were scored on all 6 criteria in Item E, this would be 12).

2. Multiply the results of Step 1 by 4 (e.g., 12 x 4 = 48)

To compute the Normalized Score on Item E for all O-numeric, A-numeric, and U-numeric elements:

$$\left[\frac{100}{\text{Maximum Possible Score on Item D for all O1, 2..., or A1, 2,..., or U1, 2,.....}}\right] \times \left[\text{Total Score from Worksheet Item D for all O1,2,... or A1, 2, ... or U1, 2,...}\right]$$

ATTACHMENT 5

O, A, U-ALPHA RATING FORM

COMPLETE ONE COPY OF THIS FORM FOR EACH O,A,U-ALPHA ELEMENT NOTED IN ITEM C OF THE WORKSHEET.

Element being scored: _____ _____
(O,A,U) (Alpha)

A. 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.).
____ Other, specify nature of alpha element and describe:

_____

B. Answer if response to A = management transformation:

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

____ Outcome only (specify which outcomes, as per diagram in Item C of Worksheet: O-numeric #___ and U-numeric #___)
____ Process, from a quality standpoint
____ Process, from an efficiency standpoint (specify from which perspective(s): ___time, ___cost, ___time & cost)
____ Process, from another standpoint. Specify:

_____

_____

(2) Complete only if answer to B(1) = process in any form; 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.

C. Complete only if response to B(1) = 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 met: ___ Yes ___ No ___ Can't Tell

(3) Composite analysis of management of rival explanations:

(a) General index of validity of design: _____. See rating box below for how to score:

GENERAL INDEX OF VALIDITY

RATINGS: 4	3	2	1	0
• well executed true experimental designs • well executed double blind crossover designs with order effects balanced and sufficient time for previous drugs to become inactive	• true experimental designs with minor problems (1-3 "3" ratings) • well executed quasi experimental designs no "3" except for selection • well excuted simple subject • crossover designs with minor problems Only "3" ratings, no less than 9 points	• quasi experimental designs with minor problems (1-3 "3" ratings or 1 "3" rating • well executed pre post designs (no "4" besides selection, maturation, history • simple subject with minor problems • true experimental with moderate problems (2-4 "3" ratings or 1-3 "2" ratings) Only "1" or "2" ratings, no less than 6 points	• pre post designs with minor to moderate problems (2-4 "3" ratings or 1-2 "2" ratings) • quasi experimental with moderate problems (6 or less points, with at least 2 "2" ratings) • true experimental with major problems • simple subject with moderate problems	• any design with one or more "1" ratings • pre post designs with major problems (3 or less points with at least 2 "2" ratings) • simple subject/ case studies with major problems

(b) Narrative description of design (using "bullet" terms in above rating box):

_____

(c) Status on threats to internal validity: Use coding conventions and definitions on next page.

___	A) Maturation
___	B) History
___	C) Testing
___	D) Instrumentation
___	E) Statistical Regression
___	F) Selection Bias
___	G) Experimental Mortality
___	H) Novelty and Disruption
___	I) Experimenter Effect
___	J) Inappropriate Statistical Procedures
___	K) Selection-maturation interaction
___	L) Instability
___	TOTAL

4 = not a plausible threat to the study's internal validity

3 = potential minor problem in attributing the observed effect to the treatment; by itself, not likely to account for substantial portion of observed result

2 = plausible alternative explanation which could account for substantial amount of the observed results

1 = plausible alternative explanation which by itself could explain most or all of the observed results

(d) Status on threats to external validity: Use coding conventions and definitions on page

___	A) Interaction
___	B) Selection
___	C) Reactive Effects
___	D) Confounded Treatment Effects
___	E) Situational Effects
___	TOTAL

General Convention: Each of the "threats," listed being are coded using the following conventions. Definitions and examples of the "threats" follow the general conventions.

4 = Not plausible threat to internal validity.

3 = Potential minor problem in attributing the observed effects to treatment; by itself, not likely to account for substantial amount of the observed results.

2 = Very plausible alternative explanation which could account for substantial amount of the observed results.

1 = Very plausible alternative explanation which by itself could explain most or all of the observed results.

A) Maturation denotes natural changes in people over time which can be mistaken for program effects or the lack of intended effects. A simple before and after comparison would be inappropriate for evluating a long term health care program for instance, because as people grow older their health tends to decline; thus, there would be a systematic bias towards underestimating the program's effectiveness.

B) History includes any set of events other than program activities or treatments that are concurrent with the program and may be influencing outcomes independently of program effects. For example, the effectiveness of a program to encourage community involvement with schools would be obscured if local teachers went on strike during the program. In general, the longer the time period under consideration, the greater the danger of historical factors rivaling the program as plausible causes of change.

C) Testing refers to the effect of having taken a pretest on posttest scores. The familiarity with a particular testing format gained during a pretest may well produce an improvement on a second test, and even when different testing instruments are used the added experience of being tested in a pretest may have the same effect, which might be interpreted erroneously as a real improvement produced by the program.

D) Instrumentation (or instrument decay) refers to changes in the ways in which measures are actually taken, which by themselves can result in differences in the observed values of outcomes variables. The evaluation of a program intended to improve social adjustment, for example, might employ periodic interviews with the participants. If the psychologists conducting these interviews change their standards of judgment or interpretation in any way across the series of interviews, this could create pseudo changes in the outcomes measures.

E) Statistical regression may also be a problem when measures are repeated as in a before and after comparison. It refers to the likelihood that on any given observation, some cases take on extreme values which deviate considerably from their normal range. These cases will tend to "regress" to their normal values on subsequent observations. This threat is expecially salient when the participants in a program have been selected on the basis of extreme scores in the first place, because there will be a systematic tendency for their scores to move in a given direction on the next test, producing pseudo program effects. Thus, the effects of a remedial reading program will be overestimated if students were placed in the program on the basis of extremely low scores on a single reading test.

F) Selection Bias is a potential threat whenever an evaluation is based on the comparison of outcomes among groups of cases whose makeup has not been determined by random assignment. While the comparison groups differ in terms of the program treatments they receive, they may also differ systematically on any other variables which might influence results, and it will not be possible to sort out the program effects from these "group effects" with certainty. Although such comparison groups may be well matched on a number of important variables, the evaluator cannot be certain that non-randomly assigned groups were in fact equivalent in terms of all the factors that might have influenced final outcomes.

G) Experimental mortality refers to the attrition of cases during the program duration or evaluation period. If, for example, there is a systematic tendency for the less able participants to drop out of a program or to refuse to submit to measurement, the average score of the remaining cases will automatically go up even if the program has no other effect. If the evaluation is based on a comparison of groups exposed to different program treatments, differential rates of experimental mortality can compound the problem. It should be understood, however, that this is only a real problem if the analysis is limited to comparing outcomes in the aggregate or care is not taken to include in the analysis of program effects only those cases which remain in the program and are measured at all observation points. (Of course, separate analysis of attrition rates and comparisons of the dropouts with those completing the program can provided valuable insight as to whom the program is best suited for and the expected response to similar program initiatives in the future:)

H) Novelty and Disruption - Measurement of the behavior made in an environment that was new; plausible that the newness of the enviroment was responsible for different scores and no control group was included in the design of the study.

I) Experimenter Effect - Attitudes of experimenter regarding expected research results are known to treatment implementer, data collector, or subject.

J) Selection-maturation Interaction refers to different rates or patterns of maturation among comparison groups, such that differences in observed outcomes among the groups may be produced by systematic differences in their maturation processes but be mistaken for bona fide program effects. This threat is of particular concern whenever an evaluation is based on long term comparisons among non-randomly assigned comparison groups.

K) Instability basically reflects a lack of reliability in the operationalized measures used in an evaluation (imprecision or unsystematic inconsistency in taking the measure), random variation in sampling persons or program components, or random fluctuations in outcome indicators across time. This is the only threat which can be contained with the use of inferential statistics.

THREATS TO EXTERNAL VALIDITY

General Convention: Each of the "threats," listed being are coded using the following convention. Definitions and examples of the "threats" follow the general conventions.

4 = Not plausible threat to external validity.

3 = Potential minor problem in attributing the observed efects to treatment; by itself, not likely to account for substantial amount of the observed results.

2 = Very plausible alternative explanation which could account for substantial amount of the observed results.

1 = Very plausible alternative explanation which by itself could explain most or all of the observed results.

A) Interaction between testing and treatment includes any responses to the stimulus of being tested or observed that might interact with the treatment or be mistaken for effects of a program treatment. Pre- testing might well sensitize clients or program participants in a way that would cause them to behave differently than would clients of participants in similar program who were not tested. For example, initial interviews intended to measure homeowner's interes in burglary prevention techniques might themselves heighten that interest and make them more receptive to the program. Similarly, posttests might prompt latent reactions that would not materialize in similar situations where evaluations were not being conducted.

B) Selection can threaten external validity if the people observed in the evaluation are not representative of the larger population of clients or prosepective clients, even though these participants might have been randomly assigned to groups. If participants in a demonstration project, for example, are selected on the basis of expediency or their high potential for success, they may receive the program treatment differently from other potential recipents. If social programs intended to serve disadvantaged subpopulations are tested with relatively more advantaged subjects, the results may appear to be much more favorable than would be the case with the intended target group. Furthermore, there can be interactions between selection and measuring devices that produce misleading results. A measuring instrument that is "culture bound" with a white, middle class orientation, for instance, may fail to pick up significant effects of a program on lower income Spanish-speaking clients.

C) Reactive effects of experimental arrangements are produced by the patent artificiality of many evaluation settings. These may be guinea pig effects in which behavior is altered simply due to the fact that people know they are being observed, they may be more calculated adjustments in behavior geared to the self-interest of respondents and their perceptions of the likely consequences of alternative outcomes of the evaluation. In general, such reactive effects are likely to produce more positive or beneficial indicators, more program success, than would be obtained in more normal settings. They are often termed "Hawthorne effects" after the findings of the Hawthorne Western Electric experiments that in some instances productivity continued to increase when such conditions as illumination and rest periods were made worse as well as when they were improved. The interpretation of these findings was that the effects were due to the existence of an experiment and the additional attention paid to the workers rather than the experimental treatments, i.e., the changes in working conditions.

D) Confounded treatment effects are impacts observed in a given program evaluation that may not apply to other similar programs because they are produced by a specific mix of treatments that might not pertain to the other situations. In a sense, any program implementation is unique in its specifics. There may exist a lack of uniformity or standardization of treatments among many similar type programs which would negate the transferability of conclusions from one to another, a problem that may be particularly salient in the assessment of a nationwide program that may take on somewhat differing characteristic in each local project. The sample of projects actually observed might not be representative of the total number of such projects. Secondly, there may be a problem of multiple treatments in which the participants in the observed projects are exposed to any number of other planned and unplanned stimuli that jointly produce the observed effects. If this mixture of treatments does not parallel those impacting on the participants of other similar programs, the results of the evaluation may apply to these other programs.

E) Situational treatment effects are closely related to confounded treatment effects in that they differentiate the programs under observation from those to which it might be desirable to transfer the results. Included in this class are threats to "ecological validity" in terms of staff characteristics, the program setting, geographic coverage, or the point in time at which the evaluation is conducted which would render the results as site or time specific. Another type of a program in an evaluation setting that might generate a much more noticeable response than would occur later under more ordinary operating conditions.

(e) Summary validity score:

C(3)(a) Score _____ x 12 + C(3)(c) Score _____ x .5 + C(3)(d) Score ____

x 1.2 = _____

(D) Summary score on Alpha Element:

_____ x B(2) Score _____ + _____ x C(3)(d) Score _____ = _____

E. Findings Analysis

(1) Status of finding (check one):

_____ Measures weakly suggest hypothesis is false.

_____ Measures strongly suggest hypothesis is false.

_____ Measures weakly suggest hypothesis is valid.

_____ Measures strongly suggest hypothesis is valid.

_____ Evaluation doesn't state explicit finding on hypothesis.

(2) Narrative statement (one or two sentences) of finding (i.e., state what "a" did or did not "cause" what "b"):

_____

_____

_____

(3) Summary confidence level on the finding/measure made:

_____ High confidence in finding/measure

_____ Moderate confidence in finding measure

_____ Low confidence in finding/measure (C(1) and/or C(2) was/were answered "no" or "can't tell")

(Same scale applies for both management transformation and hypothesis)

ATTACHMENT 6

O, A, U-ALPHA SUMMARY SCORING FORM

Element		Summary Score from Item D	Normalized Score*
Oa			
Ob			
Oc			
Od			
Oe			
	Total for all O-alpha elements		
Aa			
Ab			
Ac			
Ad			
Ae			
	Total for all A-alpha elements		
Ua			
Ub			
Uc			
Ud			
Ue			
	Total for all U-alpha elements		

* Divide Summary Score Total by number of relevant (A or O or U) alpha elements for which an Item D score was obtained.

ATTACHMENT 7

DESCRIPTIVE MATERIAL ON CRITERIA/DIMENSIONS

USED TO RATE ELEMENTS

VALIDITY ISSUES

VALIDITY: Is the instrument an appropriate one for what needs to be assessed?

Is the instrument giving a "true" measure of the value of the variable, or at least something approximating the truth.

The extent to which one can rule out interpretations of an instrument's results other than the one you wish to make.

The overall validity of an evaluation study's conclusions is limited by the weakest aspect of that study. With respect to measurement, the primary threat is the use of measures that are not truly relevant; indicators that are unrelated or only tangentially related to the subject of interest. Furthermore, even when the intent of a measure is relevant, the way in which it is taken may introduce a systematic bias in the values recorded, thereby invalidating the measure.

THE MAJOR APPROACHES TO/DIMENSIONS OF DETERMINING VALIDITY:

1. Construct validity. The construct validity of a test is the extent to which one can be sure it represents the construct (skill, attitude, ability, etc.) whose name appears in its title. Construct validity is the validity with which inferences are made about constructs on the basis of particular manipulations and measures of particular sets of manipulations and measures.

2. Content Validity: Content validity refers to the representations of the behaviors being examined within the evaluation instrument. Do these behaviors reflect the construct being studied, so that by measuring the selected behaviors, an accurate reflection of the overall construct is obtained.

3. Concurrent validity: The concurrent validity of an instrument/measure is established by collecting data to see if the results obtained with the instrument/measure agree with results from other instruments/measures, administered at approximately the same time, to measure the same thing.

4. Predictive validity: Predictive validity does not focus on what test measures, but defines its value in terms of its ability to predict future behavior. Predictive validity is the substantiation of a measure that can be used to predict other measures that are considered valid; a correlation between present and future measures.

5. Face validity: Refers to the obvious relevance of a measure as determined by the evaluator himself/herself. These measures are probably the easiest to develop and can involve direct counting of objects, or some relatively objective type of measurement.

6. Consensual validity: That conferred upon a measure by experts on the basis of their special familiarity with the subject. For example, an evaluator might consult a group of professionals in a given field to collectively develop performance measures that they agree should serve as indictors of whether or not reasonable expectations are being achieved.

7. Correlational validity: Refers to a finding that a measure co-varies over a cross-section of cases with other measures that are already considered to be valid indicators of the same attribute.

8. External validity: Also deals with the question of generalizing. Here, though, the question is not one of generalizing from specific treatments and measures to more general constructs. Rather, the targets of these attempts to generalize are persons, settings, and historical times, and the validity issue is: To what extent can a causal relationship be generalized to, or across, persons, settings, or times? Generalizing both to and across is included, since some research questions demand generalizing to specified populations of persons and times.

9. Statistical conclusion validity: The validity of conclusions about the statistical association of a presumed cause and a presumed effect.

10. Internal validity: The validity of conclusions about whether the statistical association of a treatment-as-implemented and an effect-as-measured can reasonably be considered as a causal association.

REASONS WHY EFFORTS TO PRODUCE VALID MEASUREMENTS MIGHT FAIL:

1. Lack of objectivity or standardization in administration. Remember that validity depends not only upon the contents of the measure, but also on the conditions under which it is administered.

2. Response bias or evaluation apprehension. Response bias refers to a situation in which the subjects being measured develop a strategy for responding based on something other than their knowledge of the subject matter.

3. Too few items per objective. Although a test as a whole may be a valid measure of a construct in the general subject area it represents, one may be unable to make a valid judgement about the presence of the particular subfactor measured by only one item.

4. Measuring a behavior/occurrence too narrowly. Sometimes the inference made from a measure applies to a broader range of attributes than is justified by the comparatively restricted nature of the actual measurement items.

5. Mismatch between the behavior/outcomes called for by the test and the stated objective of the test.

6. Tests whose format and wording are tied to the idiosyncracies of a particular set of instructional materials.

INCREASING VALIDITY:

1. Increasing Statistical Conclusion Validity:

Increasing statistical conclusion validity is accomplished by increasing the statistical power of analyses and by using inferential statistics in ways that do not violate important assumptions or capitalize upon chance. Problems associated with statistical power are minimized by:

(1) Having "large" sample sizes (the desired size depends, of course, on the expected size of the effect relative to the expected variance);

(2) Decreasing extraneous sources of error (e.g., using homogenous populations of respondents, and standardizing the measurement setting);

(3) Accounting for extraneous sources of variance in the statistical analyses;

(4) Increasing the reliabiity of outcome measures; and

(5) Standardizing implementation of the treatment, preferably with a high level of exposure to the treatment.

2. Increasing External Validity:

External validity deals with the generalizability of a causal relationship to, and across, populations of persons, settings, or times represented in an evaluation when it ends. These may or may not be reasonable representations of the populations of initial interest. Three major models can be followed to increase external validity:

(1) Random sampling from a designated universe allows one to generalize results to that universe;

(2) Choosing heterogeneous groups of persons, settings, or times. The evaluator tries to determine across which groups the treatment was effective, irrespective of whether the members of each group were randomly chosen to be in the experiment and, therefore, irrespective of whether there is a formal correspondence between the sample and its referent population.

(3) Generalization to modal instances is another practical way of increasing external validity. It requires explicating how a treatment would most likely be implemented if it lost its "experimental" status and became formal policy, and then finding or creating at least one research setting where the treatment is implemented in a way that closely corresponds to the explication of the modal setting. Thus, this model requires one to specify targets of generalizability in advance and then to plan the selection of persons, settings and times so that there is a commonsense correspondence between the planned targets and the achieved sample.

3. Increasing Construct Validity:

Increasing construct valdity is a matter of tailoring manipulations and measures to the rigorously defined construct they are meant to represent. Since any one operationalization of a construct is imperfect, it is extremely useful to measure or manipulate a construct in several different ways if possible. For obvious reasons, it is usually easier to measure several versions of the same presumed causal construct. Accordingly, it is appropriate to consider construct validity of the cause and construct validity of the effect separately, even though the same logic applies to both.

Increasing the construct validity of the effect is usually accomplished by demonstrating, first, that different measures of the same presumed construct covary and, second, that measures of the construct do not covary with measures of related, but different, constructs.

To increase the construct validity of an effect, the evaluator examines the relationship of different measures of the presumed effect construct with each other and with measures of similar but related constructs. For obvious practical limitations, one does not usually look at multiple manipulations of a causal construct and at manipulations of similar but related constructs. Thus, construct validity of the cause is most often increased by showing that a treatment varied what it is was supposed to vary, and only that. This is often called measuring the "take" of the independent variable.

4. Increasing Internal Validity

Internal validity is concerned with the question of whether the treatment-as-manipulated caused any change in the effect-as-measured. Internal validity threats are rendered implausible by randomly assigning experimental units to treatments. In quasi-experiments, however, since there is no random assignmnt, they must be ruled out individually. This is usually accomplished (a) by the inclusion of selected design features and (b) by examining additional data which might bear on the plausibility of each threat; or (c) by assuming, because of theory or common sense, that a particular threat- while possible -is not plausible as an alternative interpretation in the particular evaluation under discussion.

RELIABILITY ISSUES

RELIABILITY: Does the instrument yield consistent results?

The instrument gives essentially the same results when re-administered.

Reliability refers to the extent to which the measurement results are free of unpredictable kinds of error (i.e., differences in measurement resutls that are not due to the skill, attribute, etc. being evaluated).

RELIABILITY IS DEMONSTRATED IN THE FOLLOWING WAYS:

1. Test-retest reliability involves readministration of the same test/measure to the same universe subset within a given time, to determine if essentially the same results are obtained.

2. Alternate-form reliability involves altering the second administration of a measurement instrument so that the effects of memory are mitigated.

3. Split-half reliability yields a measure of consistency within a single administration. It allows the evaluator to obtain the two necessary sets of measurements from the same universe subset by taking two halves of the items comprising an instrument and treating them as two administrations.

4. Inter-rater reliability must be assessed when the "measuring instrument" is actually a person; for example, an observer, interviewer or reviewer of written material. The environment, perceptions of raters, etc. may result in two people looking at the same sample of behavior and rating it differently, and the same person looking at the same sample at different times may arrive at different measurement values.

OBJECTIVITY

The extent to which the measures obtained from a given instrument are based on the application of a uniform standard, as opposed to the subjective interpretations of those taking the measurements.

REPLICABILITY

The means/instrument/process utilized to measure the indicator must not be so esoteric or complicated that a similar measurement strategy could not be repeated, if so desired or appropriate. I.e., the measurement is of a relatively lower quality if it was only a "one-time"opportunity to obtain such measurements.

REPRESENTATIVENESS/ADEQUACY

This dimension of measurement quality can best be defined by example. If the indicator being measured is personal income, one must insure that enough data points, appropriately timed, are obtained so that "peaks" and "troughs" of any inherent cyclical pattern in personal income levels are recognized. This might be due, for example, to expenditure habits by season of the year. If data points were taken only at peaks or troughs, or only from one trough to one peak, an erroneous profile of personal income trends would be postulated.

BIAS

Is the measuring mechanism or process inherently incorporating into its results a value component not due solely to the accurate measurement of the indicator, but due to evaluator-generated and/or respondent/observed populations' biases?

Bias, in effect, is the result of systematic error (as opposed to unreliability, which results from unsystematic/random error).

ATTACHMENT 8

OVERALL EVALUATION REPORT SCORING WORKSHEET

SCORING WORKSHEET

CHARACTERISTIC I:

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 II:

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 .32 = ___

Total for Characteristic = ___

x .15 = ☐

CHARACTERISTIC III:

Subfactor 1: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .21 = ___
Subfactor 2: Co ___ + Cl ___ = ___ x 12.5 = ___ x .19 = ___
Subfactor 3: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .19 = ___
Subfactor 4: Ap ___ x 25.0 = ___ x .15 = ___
Subfactor 5: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .10 = ___
Subfactor 6: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .06 = ___
Subfactor 7: Co ___ x 25.0 = ___ x .10 = ___

Total for Characteristic = ___

x .09 = ☐

CHARACTERISTIC IV:

Subfactor 1: Co ___ + Cl ___ = ___ x 12.5 = ___ x .16 = ___
Subfactor 2: Co ___ + Cl ___ = ___ x 12.5 = ___ x .16 = ___
Subfactor 3: Co ___ + Cl ___ = ___ x 12.5 = ___ x .10 = ___
Subfactor 4: Co ___ + Cl ___ = ___ x 12.5 = ___ x .10 = ___
Subfactor 5: Cl ___ x 25.0 = ___ x .16 = ___
Subfactor 6: Cl ___ + Ap ___ = ___ x 12.5 = ___ x .16 = ___
Subfactor 7: Co ___ + Cl ___ = ___ x 12.5 = ___ x .16 = ___

Total for Characteristic = ___

x .11 = [ ]

CHARACTERISTIC V:

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 .11 = [ ]

CHARACTERISTIC VI: Co ___ + Cl ___ = ___ x 12.5

Total for Characteristic = ___

x .11 = [ ]

CHARACTERISTIC VII: Co ___ + Cl ___ = ___ x 12.5

Total for Characteristic = ___

x .11 = [ ]

CHARACTERISTIC VIII: Co ___ + Cl ___ + Ap ___ = ___ x 8.33

Total for Characteristic = ___

x .09 = [ ]

CHARACTERISTIC IX:

Subfactor 1: Ap ___ x 25.0 = ___ x .13 = ___

Subfactor 2: Summary Score for O-numeric elements ___
+ Summary Score for A-numeric elements ___
+ Summary Score for U-numeric elements ___
= ___ ÷ 3.0 x .25 = ___

Subfactor 3: Summary Score for O-alpha elements ___
+ Summary Score for A-alpha elements ___
+ Summary Score for U-alpha elements ___
= ___ ÷ 3.0 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 = [ ]

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 = [ ]

APPENDIX IV

REVISED VERSION

SCORING INSTRUMENT

ATTACHMENT 1

OVERALL SCORING INSTRUMENT

Enter values from Worksheet:

Score for MT element: _____

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

5. 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

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.

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

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

Completeness:	0	1	?	3	4	N/A

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.

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

7. The evaluators come to 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

5. Where appropriate, 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

ATTACHMENT 2

WORKSHEET FOR SCORING CHARACTERISTIC I

SUBFACTORS 2 AND 3

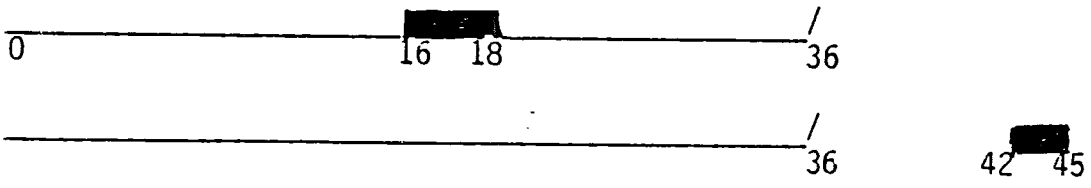
WORKSHEET FOR SCORING THE "APPROPRIATENESS" EVALUATION METHODS:

(Characteristic I, Subfactors 2 and 3)

A. EVALUATION TIMING: Based on a review of the evaluation report, complete the following informational data:

1. Planned life of project/program/policy: _____ months
2. Period under evaluation: _____ months after start of project through _____ months after start.
3. Evaluation timing graph. (Follow examples given below):

Evaluation Timing Graph Examples:



B. EVALUATION COVERAGE

1. General purpose/character of the evaluation:
   - ___ Formative, primary focus of evaluation was to provide information for replanning project/ program.
   - ___ Summative, primary focus was outcome/impact assessment of project/ program.
   - ___ Mixed (partially formative, partially summative)

C. IDENTIFICATION OF PLANNED OBJECTIVES AND THE HYPOTHESES CONCERNING THE ACHIEVEMENT OF OBJECTIVES IN A PROJECT OR PROGRAM

In order to examine the manner in which an evaluation assessed the achievement of project or program objectives, or to determine how well the evaluation examined an hypotheses, it is necessary to first identify the project/program objectives and hypotheses. In this section, the reviewer is expected to write out the objectives and hypotheses, according to a prescribed protocol and using the attached schematic. In the boxes which are placed next to the written descriptions on the schematic, the reviewer is asked to indicate:

- Whether the evaluation intended to examine particular objectives and hypotheses;
- Whether the evaluation actually examined these objectives and hypotheses.

In order to ensure that different reviewers approach this portion of the scoring effort in a similar manner, the following introductory material is provided. After reading it, please proceed to the form/schematic which is to be filled out at the end of this section.

1. Explanatory Material for the Reviewer

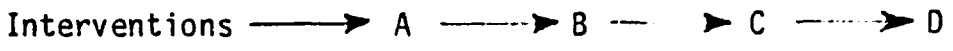
The general model of social/economic development programs/projects which are the subject of evaluations that are to be scored by this instrument by this instrument is known as a sequential effects, or linked hypotheses, model. The model, which is drawn from the physical sciences, is applied to social science situations with the full understanding that while we cannot hold all variables "constant" other than the program variables as you might in a laboratory, we can use methods of examining real world effects that approximate (in a rough way) the idea of examining the effects of one intervention at a time.

The simplest form of the general model is one which posits a dependent variable (some aspect of the real world we wish to alter; e.g., interest rates, rice production, infant mortality) and one or more independent variables (interventions we can/will make, which we hypothesize will bring about the changes we seek in the dependent variables; e.g., print more/less money, apply fertilizer, teach mothers pre-natal nutrition).

To make sense of most social/economic development programs/projects, this simple model needs to be expanded in several ways to capture the complexity of real projects and programs;

- The articulation of chains of hypotheses:

  Normally in program/projects, hypotheses about effects from an intervention are posit as being sequential; i.e., a single intervention is seen as producing a whole chain of results:



What this says is that at the first stage, a dependent variable, such as interest rate changes, can become an independent variable that produces still another effect; e.g., changes in home buying.

Stated in a little more detail, we are saying:

If A, then B

If B, then C

If C, then D

Every "if-then" combination is a separate predictive hypothesis that must be "tested" to demonstrate that the prediction was accurate. Furthermore, to "test" the predictive hypothesis, it is necessary to measure/demonstrate whether both terms of each of the hypothetical statements is valid; i.e., both A and B must first exist before we can begin to worry about whether A caused B.

In the Agency for International Development, a specialized language has been developed to characterize some element of these more general statements. In AID's specialized vocabulary, it is recognized that one must often take a number of actions to put into motion the type of intervention discussed above. The term used to describe these activities in AID is INPUTS. The INPUTS are not normally the intervention; rather, they are things we must do to set up the intervention -- much akin to putting the empty bottles on a lab table and turning the room temperature to the right level. For example, in AID terms, this may involve constructing a dam, training nurses, etc., and other preconditions for interventions that actually affect AID's target beneficiaries.

The term AID uses to describe the immediate results of these activities is OUTPUTS. In most cases, the OUTPUTS of a project/program are the actual intervention (independent variable). The process by which OUTPUTS are created is a management process/transformation, just as the filling of lab bottles is a management step. We do not say that the movement from INPUTS to OUTPUTS is an hypothesis of the same order (complexity) as those which exist between the independent variable and the chain of dependent variables our intervention is expected to effect.

As a convention, AID uses two terms to lable the first two levels of effects (dependent variables) it expects a project to yield. The term PURPOSE is applied to the dependent variable which is most proximate to the intervention (OUTPUTS); the term GOAL is applied to the next dependent variable in the chain -- the one which would be effected if PURPOSE were achieved. (In the earlier example, the production of more money would be our OUTPUT or intervention, a change in interest rates would be the PURPOSE, and a change in home buying patterns would be the GOAL). Obviously, chains of hypotheses are not limited to these few levels. In our example, additional home buying could result in a change in land values or property taxes collected or any number of other effects. In AID, we have not given specific

lables to additional levels of results -- but we know they are possible, and encourage the staff to identify them. Thus, at times, one will see additional levels described for projects/programs; e.g., "super-goal," "sub-purpose," and other such terms which stretch the chain from four to five, six or more levels.

- The Identification of "assumptions:"

  In any statement of the sort, "If A, then B," there are usually a series of factors which we are trying to hold "constant," either conceptually or in actuality. In addition, there may be factors we expect to act in a particular fashion that complements our intervention. In different contexts, different words are used to describe these factors. The word AID uses is "assumptions." To take an example,

  When we say:

  "If more money is made available, interest rates will drop."

  We are most probably saying implicitly:

  - Assuming that currency exchange rates don't change dramatically;

  - Assuming that the additional money is not hoarded;

  - Assuming that the lending system is not monopolistic and operating independent of market signals.

  All of these assumptions make the simple "If-then" statement more complex. In effect, we are forced to say:

  > If A plus a set of valid assumptions, then B.

  In AID, we ask project/program designers to identify the assumptions they are making about each level of the chain of hypotheses they set up -- including the management transformation of INPUTS and OUTPUTS. Thus, when an "If-then" hypothesis is examined for AID it looks something like this:

  > The sum of (all OUTPUTS) plus (all OUTPUT level assumptions) is predicted to yield PURPOSE.

  As was stated above, an evaluation goes about testing this predictive hypothesis by:

  --Proving the OUTPUTS were created;
  --Proving that the OUTPUT level assumptions were valid:
  --Proving that PURPOSE was achieved, and only then
  --Testing whether: OUTPUTS plus OUTPUT ASSUMPTIONS caused PURPOSE to be achieved, or

  DISCOVERING WHAT OTHER FACTOR CAUSED PURPOSE TO BE ACHIEVED IF IT WASN'T THE AID OUTPUTS.

2. Completing the Form on Project/Program Objectives and Hypotheses

At the end of Section C is a form that is to be filled in using information found in the evaluation report. In filling out the form, the reviewer must recognize that the terms INPUTS, OUTPUTS, PURPOSE and GOAL as well as other similar terms may not have been applied in the manner described here. There are a number of types of errors that crop up in AID documents with respect to the use of the AID terms and in its application of prescriptions for separating levels of effects from each other. Because of the problems of misapplication of terms and occasional failure to separate objectives, the form on the following page does not use the AID terminology. Thus, the reviewer is asked to use the evaluation document material to extract information for the form -- and correct errors -- rather than copy exactly what is found under the terms INPUTS, OUTPUTS, PURPOSE and GOAL. Most likely, correction will be needed in the following areas:

- "Double-Up" Statements about Results:

  In trying to use a four-level chain of hypothesis and results, AID staff may occasionally collapse two objectives under one term; e.g.:

  PURPOSE: To improve the availability of high-protein content food in order to improve family nutrition.

  The reviewer will be expected to correct this by using two of the boxes on the form to express this statement:

  Improve family nutrition

  Improve the availability of high-protein content food

- Incorrectly categorized assumptions:

  Due to some misunderstandings about the nature of the statement:

  "A" plus (assumptions concerning "A") yield "B"

  The reviewer may find that the assumptions are not always listed in the most appropriate place in a evaluation document. For example, assumptions about farmer motivation, which should be connected with results in terms of fertilizer use, will not appear where they should if the report suggests that both are needed to improve crop yield. Once again, the reviewer is expected to put in the appropriate place on the form/schematic provided.

- Unstated Results:

  A third problem the reviewer may face is that steps in a chain of hypotheses may not all be expressed. For example, you may find:

  If fertilizer is provided, farmer income will increase.

  A statement like this leaves out important intermediate steps:

- Fertilizer will be applied to crops
- Crop yield will increase
- The additional yield will be sold (with an assumption about prices not falling)

You may even find that the evaluation actually looks at these unstated intermediate results.

The desired approach for completing the form in this section involves filling in these unexpressed objectives -- and noting that they weren't expressed by the evaluation report.

As the above ways of correcting what is available in the evaluation document suggests, the reviewer is being asked to fully "flesh out" the program/project logic before beginning to work on what the evaluation intended to and actually did measure. The potential need for corrections is the main reason why the attached form deviates from the language AID normally uses -- and the exercise itself will help AID to better structure future projects and project evaluations.

Note that the form does not ask you to identify results of a project/program that were not anticipated in the design. These unplanned effects reported by an evaluation are addressed in a separate section.

ATTACHMENT 3

RATING FORM FOR SCORING INPUTS, OUTPUTS,
DEPENDENT VARIABLES (E1....En), ASSUMPTIONS
(A1....A-En), AND UNPLANNED RESULTS (U-O, U-E1, U-E2, ... U-En)

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
(E1, E2....En)
Complete 1 copy of Form for each set of ASSUMPTIONS
(A-I, A-O, A-E1, A-E2,....A-2n)

Element being scored: _____

(For example, Inputs, Output 1, E1, A-E1, U-E1)

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 <u>each</u> indicator measured for this Element:

(1) Check which of these is applicable:

Ind 1	Ind 2	Ind 3	Ind 4	
__	__	__	__	a. Presence/absence(i.e., indicator was <u>not</u> present "before" activity being evaluated began).
__	__	__	__	b. Change in status (i.e., indicator <u>was</u> present "before" activity being evaluated began; measure focuses on change)
__	__	__	__	c. Both (i.e., indicator was present "before" but not "after")

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

__	__	__	__	(a) Measure was <u>valid</u> measure of presence/absence for the indicator
__	__	__	__	(b) Measure was <u>replicable</u>
__	__	__	__	(c) Measure was <u>unbiased</u>
__	__	__	__	(d) Measure was <u>objective</u>

(3) Complete only if C (1) response = change in status (response b). Score 0 = No, 2 = Somewhat, 4 = Yes

Ind. 1	Ind. 2	Ind. 3	Ind. 4	
__	__	__	__	(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

(4) Complete only if C (1) response = both (response) . Score 0 = No. 2 = Somewhat, 4 = Yes

__	__	__	__	(a) Measure was valid measure of indicator which was to have changed/ existed
__	__	__	__	(b) Measures at all points were made in consistent manner
__	__	__	__	(c) Measure 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 from which a generalization was made:

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

E. Summary score on the finding/measure made:

Ind 1	Ind 2	Ind 3	Ind 4	Total for All Indicators	
—	—	—	—	—	Validity: Score from C(2)(a) or C(3)(a) or (C(4)(a)
					Replicability/consistency: Score from C(2)(b) or C(3)(b) or C(4)(b)
—	—	—	—	—	Bias: Score from C(2)(c) or C(3)(c) or C(4)(c)
—	—	—	—	—	Representativeness/Adequacy: Score form C(3)(d) or C(4)(
—	—	—	—	—	Objectivity: Score from C(2)(d) or C(3)(e) or C(4)(e)
—	—	—	—	—	Generalization: Score from Item D
—	—	—	—	—	Grand Total

F. Findings Analysis

(1) Status of indicators. Check for each indicator used to measure the variable being evaluated:

Ind 1	Ind 2	Ind 3	Ind 4	
—	—	—	—	Present (if only presence/absence was assessed; response C(1)(a) or C(1)(c)
—	—	—	—	Positive change (if change was assessed); response C(1)(b).
—	—	—	—	Negative change (if change was assessed); response C(1)(b).
—	—	—	—	No change (if change was assessed); response C(1)(b).

(2) Summarize the finding regarding this variable (1 or 2 sentences):

_____

_____

_____

(3) 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						____

(4) If C(1) response = change in status (response b), then 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							____

(5) Overall Confidence Level:

F(3) or F(4) Score + D Score = _____

ATTACHMENT 4

SUMMARY SCORING FORM

FOR ATTACHMENT 3

Element	Total Score From Worksheet Item E	Maximum Possible Score	Normalized Score*
Score for all Input elements:	____	____	____
Output 1	____	____	
Output 2	____	____	
Output 3	____	____	
Output 4	____	____	
Output 5	____	____	
Output 6	____	____	
Total for all Output elements:	____	____	____
E1	____	____	
E2	____	____	
E3	____	____	
E4	____	____	
E5	____	____	
E6	____	____	
Total for all E elements:	____	____	____
A-I	____	____	
A-O	____	____	
A-E1	____	____	
A-E2	____	____	
A-E3	____	____	
A-E4	____	____	
Total for all A elements:	____	____	____

*See computation formula which follows.

Element	Total Score From Worksheet Item E	Maximum Possible Score	Normalized Score*
U-O			
U-E1			
U-E2			
U-E3			
U-E4			
U-E5			
Total for all U elements:			

To compute the Maximum Possible Score on Item E:

1. Determine the total number of criteria in Item E for which a score of 0, 2, or 4 was entered (e.g., if 2 indicators were scored on all 6 criteria in Item E, this would be 12).

2. Multiply the results of Step 1 by 4 (e.g., 12 x 4 = 48)

To compute the Normalized Score on Item E for all Input, Output, E, A, and U elements:

$$\left[\frac{100}{\text{Maximum Possible Score on Item D for all E1, E2.., or A-I, A-O, A-E1, A-E2,...or Output 1, Output 2,..., or Inputs, or U-O, U-E1, U-E2}}\right] \times \left[\text{Total Score from Worksheet Item D for all E1, E2,... or A-I, A-O, A-E1, A-E2,.. or Output 1, Output 2,..., or Inputs or U-O, U-E1, U-E2,...}\right]$$

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 <u>each</u> HYPOTHESES (Ha, Hb, etc.)

Element being scored: _____

(For example, MT, Ha, Hb, etc.)

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.).

____ Other, specify nature of alpha element and describe:

_____

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 Item C of Worksheet: Output #___)

____ (b) Process, from a quality standpoint

____ (c) Process, from an efficiency standpoint (specify from which perspective(s): ___time, ___cost, ___time & cost)

____ (d) Process, from another standpoint. Specify:

_____

_____

(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

(3) Composite analysis of management of rival explanations:

(a) Status on threats to internal validity: Use coding conventions and definitions on next two pages.

___	A) Maturation
___	B) History
___	C) Testing
___	D) Instrumentation
___	E) Statistical Regression
___	F) Selection Bias
___	G) Experimental Mortality
___	H) Novelty and Disruption
___	I) Experimenter Effect
___	J) Inappropriate Statistical Procedures
___	K) Selection-maturation interaction
___	L) Instability
___	TOTAL

4 = not a plausible threat to the study's internal validity

3 = potential minor problem in attributing the observed effect to the treatment; by itself, not likely to account for substantial portion of observed result

2 = plausible alternative explanation which could account for substantial amount of the observed results

1 = plausible alternative explanation which by itself could explain most or all of the observed results

(b) Status on threats to external validity: Use coding conventions and definitions on previous 2 pages.

A) Interaction

B) Selection

C) Reactive Effects

D) Confounded Treatment Effects

E) Situational Effects

TOTAL

(c) General index of validity of design based on responses to C(a) and (b): _____. See rating box below for how to score:

GENERAL INDEX OF VALIDITY

RATINGS: 4	3	2	1	0
• well executed true experimental designs • well executed double blind crossover designs with order effects balanced and sufficient time for previous drugs to become inactive	• true experimental designs with minor problems (1-3 "3" ratings) • well executed quest experimental designs no "3" except for selection • well excuted single subject • crossover designs with minor problems Only "3" ratings, no less than 9 points	• quasi experimental designs with minor problems (1-3 "3" ratings or 1 "3" rating • well executed pre post designs (no "4" besides selection, maturation, history • single subject with minor problems • true experimental with moderate problems (2-4 "3" ratings or 1-3 "2" ratings) Only "1" or "2" ratings, no less than 6 points	• pre post designs with minor to moderate problems (2-4 "3" ratings or 1-2 "2" ratings) • quasi experimental with moderate problems (5 or less points, with at least 2 "2" ratings) • true experimental with major problems • single subject with moderate problems	• any design with one or more "1" ratings • pre post designs with major problems (3 or less points with at least 2 "2" ratings) • single subject/ case studies with major problems

(d) Narrative description of design's validity (using "bullet" terms in above rating box):

_____

_____

THREATS TO INTERNAL VALIDITY

General Convention: Each of the "threats," listed being are coded using the following conventions. Definitions and examples of the "threats" follow the general conventions.

4 = Not plausible threat to internal validity.

3 = Potential minor problem in attributing the observed effects to treatment; by itself, not likely to account for substantial amount of the observed results.

2 = Very plausible alternative explanation which could account for substantial amount of the observed results.

1 = Very plausible alternative explanation which by itself could explain most or all of the observed results.

A) Maturation denotes natural changes in people over time which can be mistaken for program effects or the lack of intended effects. A simple before and after comparison would be inappropriate for evluating a long term health care program for instance, because as people grow older their health tends to decline; thus, there would be a systematic bias towards underestimating the program's effectiveness.

B) History includes any set of events other than program activities or treatments that are concurrent with the program and may be influencing outcomes independently of program effects. For example, the effectiveness of a program to encourage community involvement with schools would be obscured if local teachers went on strike during the program. In general, the longer the time period under consideration, the greater the danger of historical factors rivaling the program as plausible causes of change.

C) Testing refers to the effect of having taken a pretest on posttest scores. The familiarity with a particular testing format gained during a pretest may well produce an improvement on a second test, and even when different testing instruments are used the added experience of being tested in a pretest may have the same effect, which might be interpreted erroneously as a real improvement produced by the program.

D) Instrumentation (or instrument decay) refers to changes in the ways in which measures are actually taken, which by themselves can result in differences in the observed values of outcomes variables. The evaluation of a program intended to improve social adjustment, for example, might employ periodic interviews with the participants. If the psychologists conducting these interviews change their standards of judgment or interpretation in any way across the series of interviews, this could create pseudo changes in the outcomes measures.

E) Statistical regression may also be a problem when measures are repeated as in a before and after comparison. It refers to the likelihood that on any given observation, some cases take on extreme values which deviate considerably from their normal range. These cases will tend to "regress" to their normal values on subsequent observations. This threat is expecially salient when the participants in a program have been selected on the basis of extreme scores in the first place, because there will be a systematic tendency for their scores to move in a given direction on the next test, producing pseudo program effects. Thus, the effects of a remedial reading program will be overestimated if students were placed in the program on the basis of extremely low scores on a single reading test.

F) Selection Bias is a potential threat whenever an evaluation is based on the comparison of outcomes among groups of cases whose makeup has not been determined by random assignment. While the comparison groups differ in terms of the program treatments they receive, they may also differ systematically on any other variables which might influence results, and it will not be possible to sort out the program effects from these "group effects" with certainty. Although such comparison groups may be well matched on a number of important variables, the evaluator cannot be certain that non-randomly assigned groups were in fact equivalent in terms of all the factors that might have influenced final outcomes.

G) Experimental mortality refers to the attrition of cases during the program duration or evaluation period. If, for example, there is a systematic tendency for the less able participants to drop out of a program or to refuse to submit to measurement, the average score of the remaining cases will automatically go up even if the program has no other effect. If the evaluation is based on a comparison of groups exposed to different program treatments, differential rates of experimental mortality can compound the problem. It should be understood, however, that this is only a real problem if the analysis is limited to comparing outcomes in the aggregate or care is not taken to include in the analysis of program effects only those cases which remain in the program and are measured at all observation points. (Of course, separate analysis of attrition rates and comparisons of the dropouts with those completing the program can provided valuable insight as to whom the program is best suited for and the expected response to similar program initiatives in the future:)

H) Novelty and Disruption - Measurement of the behavior made in an environment that was new; plausible that the newness of the enviroment was responsible for different scores and no control group was included in the design of the study.

I) Experimenter Effect - Attitudes of experimenter regarding expected research results are known to treatment implementer, data collector, or subject.

J) Selection-maturation Interaction refers to different rates or patterns of maturation among comparison groups, such that differences in observed outcomes among the groups may be produced by systematic differences in their maturation processes but be mistaken for bona fide program effects. This threat is of particular concern whenever an evaluation is based on long term comparisons among non-randomly assigned comparison groups.

K) Instability basically reflects a lack of reliability in the operationalized measures used in an evaluation (imprecision or unsystematic inconsistency in taking the measure), random variation in sampling persons or program components, or random fluctuations in outcome indicators across time. This is the only threat which can be contained with the use of inferential statistics.

THREATS TO EXTERNAL VALIDITY

General Convention: Each of the "threats," listed being are coded using the following convention. Definitions and examples of the "threats" follow the general conventions.

4 = Not plausible threat to external validity.

3 = Potential minor problem in attributing the observed efects to treatment; by itself, not likely to account for substantial anount of the observed results.

2 = Very plausible alternative explanation which could account for substantial amount of the observed results.

1 = Very plausible alternative explanation which by itself could explain most or all of the observed results.

A) Interaction between testing and treatment includes any responses to the stimulus of being tested or observed that might interact with the treatment or be mistaken for effects of a program treatment. Pre- testing might well sensitize clients or program participants in a way that would cause them to behave differently than would clients of participants in similar program who were not tested. For example, initial interviews intended to measure homeowner's interes in burglary prevention techniques might themselves heighten that interest and make them more receptive to the program. Similarly, posttests might prompt latent reactions that would not materialize in similar situations where evaluations were not being conducted.

B) Selection can threaten external validity if the people observed in the evaluation are not representative of the larger population of clients or prosepective clients, even though these participants might have been randomly assigned to groups. If participants in a demonstration project, for example, are selected on the basis of expediency or their high potential for success, they may receive the program treatment differently from other potential recipents. If social programs intended to serve disadvantaged subpopulations are tested with relatively more advantaged subjects, the results may appear to be much more favorable than would be the case with the intended target group. Furthermore, there can be interactions between selection and measuring devices that produce misleading results. A measuring instrument that is "culture bound" with a white, middle class orientation, for instance, may fail to pick up significant effects of a program on lower income Spanish-speaking clients.

C) Reactive effects of experimental arrangements are produced by the patent artificiality of many evaluation settings. These may be guinea pig effects in which behavior is altered simply due to the fact that people know they are being observed, they may be more calculated adjustments in behavior geared to the self-interest of respondents and their perceptions of the likely consequences of alternative outcomes of the evaluation. In general, such reactive effects are likely to produce more positive or beneficial indicators, more program success, than would be obtained in more normal settings. They are often termed "Hawthorne effects" after the findings of the Hawthorne Western Electric experiments that in some instances productivity continued to increase when such conditions as illumination and rest periods were made worse as well as when they were improved. The interpretation of these findings was that the effects were due to the existence of an experiment and the additional attention paid to the workers rather than the experimental treatments, i.e., the changes in working conditions.

D) Confounded treatment effects are impacts observed in a given program evaluation that may not apply to other similar programs because they are produced by a specific mix of treatments that might not pertain to the other situations. In a sense, any program implementation is unique in its specifics. There may exist a lack of uniformity or standardization of treatments among many similar type programs which would negate the transferability of conclusions from one to another, a problem that may be particularly salient in the assessment of a nationwide program that may take on somewhat differing characteristic in each local project. The sample of projects actually observed might not be representative of the total number of such projects. Secondly, there may be a problem of multiple treatments in which the participants in the observed projects are exposed to any number of other planned and unplanned stimuli that jointly produce the observed effects. If this mixture of treatments does not parallel those impacting on the participants of other similar programs, the results of the evaluation may apply to these other programs.

E) Situational treatment effects are closely related to confounded treatment effects in that they differentiate the programs under observation from those to which it might be desirable to transfer the results. Included in this class are threats to "ecological validity" in terms of staff characteristics, the program setting, geographic coverage, or the point in time at which the evaluation is conducted which would render the results as site or time specific. Another type of a program in an evaluation setting that might generate a much more noticeable response than would occur later under more ordinary operating conditions.

(e) Summary validity score:

B(3)(c) Score _____ x 12 + B(3)(a) Score _____ x .5 + B(3)(b) Score ____

x 1.2 = _____

(C) Summary score on element:

6.25 x A(2) Score _____ or 1.06 x B(3)(e) Score _____ = _____

D. Findings Analysis

(1) Status of finding (check one):

_____ Measures weakly suggest hypothesis is false.

_____ Measures strongly suggest hypothesis is false.

_____ Measures weakly suggest hypothesis is valid.

_____ Measures strongly suggest hypothesis is valid.

_____ Evaluation doesn't state explicit finding on hypothesis.

(2) Narrative statement (one or two sentences) of finding (i.e., state what "a" did or did not "cause" what "b"):

_____

_____

_____

(3) Summary confidence level on the finding/measure made:

_____ High confidence in finding/measure

_____ Moderate confidence in finding measure

_____ Low confidence in finding/measure (B(1) and/or B(2) was/were answered "no" or "can't tell")

(Same scale applies for both management transformation and hypothesis)

ATTACHMENT 6

SUMMARY SCORING FORM FOR ATTACHMENT 5

Element	Summary Score from Item C	Normalized Score*
Score for Management Transformation element		
Ha		
Hb		
Hc		
Hd		
He		
Total for all H elements		

* Divide Summary Score Total by number of relevant (MT or H) elements for which an Item C score was obtained.

ATTACHMENT 7

(Unchanged from First Draft)

ATTACHMENT 8

OVERALL EVALUATION REPORT SCORING WORKSHEET

SCORING WORKSHEET

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 ___
+ Summary Score for H elements ___
= ___ ÷ 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 = ☐

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 .82 = ___

Total for Characteristic = ___

x .15 = ☐

CHARACTERISTIC IV:

Subfactor 1: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .21 = ___
Subfactor 2: Co ___ + Cl ___ = ___ x 12.5 = ___ x .19 = ___
Subfactor 3: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .19 = ___
Subfactor 4: Ap ___ x 25.0 = ___ x .15 = ___
Subfactor 5: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .10 = ___
Subfactor 6: Co ___ + Cl ___ + Ap ___ = ___ x 8.33 = ___ x .06 = ___
Subfactor 7: Co ___ x 25.0 = ___ x .10 = ___

Total for Characteristic = ___

x .09 = [ ]

CHARACTERISTIC V:

Subfactor 1: Co ___ + Cl ___ = ___ x 12.5 = ___ x .16 = ___
Subfactor 2: Co ___ + Cl ___ = ___ x 12.5 = ___ x .16 = ___
Subfactor 3: Co ___ + Cl ___ = ___ x 12.5 = ___ x .10 = ___
Subfactor 4: Co ___ + Cl ___ = ___ x 12.5 = ___ x .10 = ___
Subfactor 5: Cl ___ x 25.0 = ___ x .16 = ___
Subfactor 6: Cl ___ + Ap ___ = ___ x 12.5 = ___ x .16 = ___
Subfactor 7: Co ___ + Cl ___ = ___ x 12.5 = ___ x .16 = ___

Total for Characteristic = ___

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 = [ ]

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 = [ ]