Sample

September 7, 1991

[redacted] [redacted]

From: Chuck Walters

Subj: Useability Questionnaire Analysis

Ref: My Memo of 30 Aug 81 w/Questionnaire Attached

1. The reference provided the questionnaire that is used by the Marine Corps and Air Force operational test agencies (OTA) to evaluate the adequacy of the software to meet user requirements. The contents of this memo provides the methodology for converting the subjective questionnaire responses into a quantitative evaluation tool.

2. The subject questionnaire consists of 95 questions broken down as follows with associated category weights to stratify the questionnaire results over all categories:

Category	Number of Questions	Weight
Confirmability	16	.1818
Controllability	14	.1591
Workload	19	.2159
Descriptiveness	15	.1705
Consistency	10	.1136
Simplicity	14	.1591
General by Category	7	NA

The weights for each category are determined (based on the 88 questions associated to the categories) by dividing the number of questions in a category by the total number of questions e.g. Confirmability: 16/88 = .1818. In the event that questions are to be deleted or added, reevaluate the weights for each category e.g. Deleting two questions from Workload results in a total of 86 questions divided into 17 equals .1977. Each of the remaining category weights would be recalculated based on 86 questions vice 88.

3. The questionnaire provides for the following responses with associated scores:

A.	Completely Agree	- 6.0	
B.	Strongly Agree	- 5.0	
C.	Generally Agree	- 4.0	> 3.5 = (4.0 + 3.0)/2
D.	Generally Disagree	- 3.0	
E.	Strongly Disagree	- 2.0	
F.	Completely Disagree	- 1.0	

Note: The midpoint of Generally Agree and Generally Disagree generates a value of 3.5. This is the threshold value established by the Marine Corps and Air Force OTAs to evaluate the adequacy of software useability from the Users viewpoint. From zero to 3.49, the software is inadequate. From 3.5 to 6.0, the software is adequate. The software is evaluated within each category to isolate deficient areas for correction (below 3.5) and an overall score based on weighted values across the six categories. The values for the General questions (last 7 of 95) are calculated and compared to the category questions to obtain the Users overall viewpoint.

4. To calculate the mean scores per category ($\bar{x}$) and overall, the following applies:

By Question	By Category	Overall
$\sum_{i=1}^{n} \frac{QV_i}{n}$	$\sum_{i=1}^{n} \sum_{j=1}^{NQ} \frac{QV_{ij}}{n(NQ)}$	$\sum_{i=1}^{n} \sum_{j=1}^{NQ} \frac{QV_{ij}}{n(NQ)}$

where: n - Number Questionnaire Respondents
QV - Question Value (1.0 to 6.0)
NQ - Number Questions in Category

To stratify the overall questionnaire values across the categories (excluding the seven general questions):

$$\bar{X} = \sum_{i=1}^{6} \bar{X}_i * wt_i$$

where: $\bar{x}_i$ - The mean values by category
wt_i - The category weights

5. The following provides an example (data was extracted from a 1986 operational test). The data set is the last page of this memo. The number of questions and category weights are same as page 1 of this memorandum. There were 25 respondents.

a. Calculation of Question # 1 value. From the data set: the sum of scores for column Ø1 is 109; therefore:

$$\bar{X}_{01}=\frac{\sum_{i=1}^{25} QV_i}{n}=\frac{109}{25}=4.360$$

b. For the calculation of Confirmability category value, using the attached data set: The sum of question scores within the category is 1782 where n=25 and NQ=16 derived from:

$$\bar{X}_{Confirm}=\frac{\sum_{i=1}^{25}\sum_{j=1}^{16} QV_{ij}}{n*NQ}=\frac{1782}{25*16}=4.455$$

c. Calculation of overall across 95 questions. From the data set: the overall sum of scores is 9286 where n=25 and NQ=95 derived from:

$$\bar{X}_{overall}=\frac{\sum_{i=1}^{25}\sum_{j=1}^{95} QV_{ij}}{n*NQ}=\frac{9286}{25*95}=3.910$$

d. Following data summary applies:

(1) Controllability

QUESTION NUMBER	SUM SCORE	$\bar{X}$
17	99	4.0
18	99	4.0
19	120	4.8
20	113	4.5
21	121	4.8
22	136	5.4
23	127	5.1
24	96	3.8
25	130	5.2

QUESTION NUMBER	SUM SCORE	$\bar{x}$
26	63	2.5 *
27	98	3.9
28	121	4.8
29	101	4.0
30	100	4.0

Note: Question # 26 has a score below the 3.5 threshold: "Strings of commands can be created and executed as a single command"

(2) <u>Category Means and Weighted Means:</u>

where: $\bar{x}_{wt} = \bar{x}_{cat} * wt$

e.g. Confirmability: $\bar{x}_{wtconfirm} = 4.455 * .1818 = .8099$

Category	$\bar{x}$	$\bar{x}_{wt}$
Confirmability	4.455	.8099
Controllability	4.354	.6928
Workload	3.918	.8456
Descriptiveness	3.195	.5447 *
Consistency	3.774	.4253
Simplicity	3.617	.5755
		3.894

* The category of Descriptiveness has a score below the 3.5 threshold and is considered a problem area. [The low score was attributed to deficiencies in the operators manual.]

(3) <u>Overall Assessment</u>

(a) Calculated mean across all categories: 3.910

(b) Weighted Mean Value: 3.894

The software is marginally mature from a useability viewpoint. Deficiencies in Descriptiveness has the most severe impact on the User.

Chuck W

[redacted] - SOFTWARE EVALUATION DATABASE DUMP

QUESTION VALUES & RESPONSES

10/03/86 16:05:13

A.

Page — Confirmability | Controllability | Workload | Descriptiveness | Consistency | Simplicity | Ge[neral]

Label	Page	000000000111111111122222222222333333333334444444444555555555556666666666777777777788888888889999 / 123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123
NCO – E	3104	DBBEBCBBBABCDEBCBFBBDBBABCCBCCCBCABBADCDBABCCCCBACEDDDBBCBBBBAADBBBBBBCCBBBBBBBBBABBCBCBCCBBB
SNCO E	6790	CBDCCCBDCCBCCCCCDDCCCCCDCDCCDCCCDCCCCEDCCDCCCCCCCDCEDDDCCDDDCCCCCCCCCCCCCCCCCCCCCCCCDCDCDCCCCC
NCO E	6312	CBDCDCBCBACECCCBDDCCCCDDCDDCCCCDDCBBCDDCDCDEEDCDDEDDFEEDEECFEEEEBBDDCDDDDECDCCCDBCCECFEDCCDDD
NCO E	4980	CACDDDEDAACECCCCDECCDCDDCEDCCCCDDCBBDCDCDCDECDCECEEDFFEDEFCFEEEECCDDCCDDDDCECCCDDCDEEEEDDDCEDC
E5 E	7425	BCEEDCBBCBBDCBCCDDBCCCCDBDBBDBDDDBBBBBDDBDBCCEEDEEDDEDCCDCEEEDDBBDCCCCDEDCECECBBCCEDCECDBDDBB
O	1004	BBCDCCCDBBCDCCCCDCBBBBBDBDCBCCCDDCCACCCCCCDCCCAEBECCFDCCDCCDDCCCBCCCCDCDCDCCCCDCCCCDDCDCCCCDCC
O	9838	BCDCCBBCDABCECBEDCCCDBCEBECCCCCDFDBCCBBBCCBCCBCCBDCCFDCCDEBEBCCCBBCBDCCEDCABCBCDCCCDDECDCCBCCC
O	2973	DCDDCCBBCACDCCCECCBDCBBDBECBCCBCEBBDECCBECBFEFCCCECFFFFEEEBCDEFCCCDDFFCEFCDCCDCFDDCFADDCEBDECC
O	2280	BBBEAAAAAAADBDBDCCBBBAADCDCAEBBDDBBBECCABBBEBDCCCDCDEECCCCBCCDDCDBCCCCDDCCBDCDEBDCEEEEECCDDCE
O	2512	CCCDCCBCCCCCDCCBDCBBBAACACCCDCCCCCBCDCCDCCDDDDCDDFCDDFCDDCAADDFCBBCDCCCEDDDCCCDCBCCDEFFDCCCCDC
NCO	1619	BAAACBCBBAABBCBACCACBABCABCCBCBAACCADBCACCCDEDFBCBBCCBDDBBBDCBBBDCDCCCCCBADCCCCBEBCBECCCBBCCCC
NCO	7498	ACBACCCABADCEAAACBBAAAAAAFAAEFBBCBCAABDCCCCFFDCBCFDDFFFEAAAAFFFFADDFBEDFFDCDCDCDAACFEFFFBCCBBC
	4930	BBBBCCBCAACBBCBBCCACBABCCDCCBCABBBABABCBCBCEDEECCCCDFDCDCACBCDDCCBCBCCBBCBBBBCCBBCEEEDDCBCABB
	2134	BBCACBBCBBBBDCBBCCCBBAABAFCBBCABBBBBBDDBDBEFECCCCFBCFDEFCCBBFDFCBBECBBCDDCBBCCBCBBCEFFFECBCCCB
	5086	CAACCCCCCAAACCACBBACAAABCCCABBAACBACCCCCCCBCCCCABCACCCCAACCCCCCCAACACCBCCCCCCCCCABAAACCCAAAAAA
NCO	845	BBBEAAAAAAAAEBAABACBAABEAFBACDABCABEEBBADBFFFFDCCDADFFEFEDDBCDDCCCEDBCBDDECBCBCDABCEAFEEBCDFCD
E3	6984	CBCBACACAABCDBAACBBCBACCAFCCBCACCCAABBBCCCCCCCFFCCCCCBBCDBBBBDBBBCBCCCCBCAABCCAABBCADDCABBBBB
E2	7181	BBCCBCBBCBCCCCBBCBBCAAAAAFBBBCBBCBBCBCCCCBCFDCBBCEDDFFFDECBBEEEDBBCDBBBEEECBCCBCBBBEEFFECBCDCC
O	6231	CBBACCCACACCDDBDCCDCBAACBCCCBCBCECBCCCCDCCCDFCDCCCDDEEDCCDCCCDDDCCCCCCCDCCBDCBCDCCCCDDDDCCCCCD
SNCO	9656	CCBACCACAAABBBACCBABAAABAFCBBABBCCCABCCBCBFFFCCBFDCDDDDCCCBBCFFBBBCCCCCCCFACCCCCABCCCFFFBBCCCC
NCO	454	CBCCBCCBCBBCDCBBCCBBBABBADBABCABBBABCDDBCAFEDBCCCEDCFDEEECBCEEFEBBEDCBCEFEBCCCBCBBCECFFECBCDCC
SNCO	7186	CBCCAAACBABCFCABBBBBAAACACCACCCABBABBDDCCAFDCDBCBCBBFFCCCCABBDCCBBBFCCBCBBBBBBCBBBAFCFDCDBBBBB
O	180	DBFECCBEAABDEECBBEBAAAAEBFFBFFFBEBEBEFECEBFFFFEEFFEFDFFEEFECFEFCDCEEFEFFFFFFDEFFCFEFFFFFDCCFEE
SNCO	3436	BBABBBABBABAACBBABBBBAACAFFCAABCBBBBCBABBAFFECCBBDEEEEEFCCBBCFFCBBBDBBCEFFBCBBCDABBEFFEFCBBCBB
SNCO	6033	CCDCCDCBFABDCCCCDCCCCCCDBDCBCCDCDCCCDCCBCCECCCCDCDDCEEECDDBCECCCCCCCCCCCECCDCCCDCCCDDCCDCCCCCC

O - ofer

W - ofer

E5/4 NCO

E6-9 SNCO

VALUES

B.

Confirmability (1–16) | Controllability (17–30) | Workload (31–49) | Descriptiveness (50–64) | Consistency (65–74) | Simplicity (75–88) | General (89–95)

	00000000011111111112222222222233333333333444444444455555555555666666666677777777778888888888999999 / 12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345
3104	3552545556543254615535565445444546556343565444456423335545555663555555544555555556554545445555 5
6790	4534445344544444334444434344344434442344344444443423334433344444444444444444444434343444444 4
6312	4534345456424445334444334334444334554334343223433233122322412222553343333243444354424123443334 3
4980	4643332366424444324434334234444334553434343243424223112321412222443344333342444334322223334234 3
7425	5422345545534544335444435355533335555555333535442232233234434222335534444323424245544234243533554
1004	5543444355434444345555535345444334464444443444625244134434433444454444343434444344443334344434444
9838	5434455436542452544435425244444313544555445445445344134432525444554534423465454344433243445444 4
2973	3433445546434442445345535245445425532445245121444241111222543214443311421434434133416334253244 3
2280	5552666666635353445556634346255335552446555253444343224444544334354444433445343253422222443342 3
2512	4443445444443445345556646444344444543443443333433143314334663314554344423334443454432113444434 4
1619	5666454556655456446456546544545664463546444323154554453335553455534344444563444452545244465444 5
7498	6456444656342666455666666166215545466534444113454133111266661111633152311343434366412111544554 4
4930	5555445466455455446456544344546555656545454232244443134346454334454544554555554455422233454655 5
2134	5546455455553455444556656145546555555333535212444154132144551314552455433455445455421112454445 4
5086	4664444446664464556466654446556645644444445444465464444664444446646445444444444656664446666666
845	5552666666662566564566526156436546522556351111344363112123354334442354533245454365426122543143 6
6984	4545646466543566455456444614454644466555444444411444445543555553554544445466544665546334555555 4
7181	5544545545444455455466666155545545545444454134554233111324552223554355522245445455522112454344 4
6231	4556444646443353443456645444545424544443444314344433223443444333444444434453454344443333444443 4
9656	4456446466655564456566656145565544465445451114451343333444554115544444441644444654441115544444
454	4544544545543455445556556356546555654335461235444234132224542212552345421254444554241124543444
7186	4544666455541465555566646446446556553344613435454551144446553445551445455555545556141345555555
180	3512445266532245525666625115111525252124251111221121311221241214342212111111321141211111544122 4
3436	5565556556566645565555664611466545555456556112445532222144554114655355421154554365521121455455 4
6033	4434434516534444344444435345443434443445442444434334222433542444444444424434443444334434444444
Category Total Value:	1782 / 1524 / 1861 / 1198 / 936 / 1266 / 719

OVERALL TOTAL: 9286

TOTAL FOR QUESTION # 01: 109

CATEGORY	QUESTION NO.	J.R.	N.E.	C.W.	SUM OF ROWS	MEAN SCORE
WORKLOAD SUITABILITY +	31	3	2	4	9	3.0
	32	6	5	6	17	5.67
+	33	2	2	2	6	2.0
	34	6	6	6	18	6.0
	35	5	5	6	16	5.33
	36	5	5	6	16	5.33
	37	6	5	6	17	5.67
	38	5	6	6	17	5.67
	39	5	6	6	17	5.67
+	40	3	2	2	7	2.33
	41	5	6	5	16	5.33
	42	5	6	6	17	5.67
	43	6	6	6	18	6.0
	44	6	6	6	18	6.0
	45	4	5	6	15	5.0
	46	5	5	6	16	5.33
	47	5	5	6	16	5.33
	48	5	5	5	15	5.0
	49	6	5	5	16	5.33
SUM OF COLUMNS		93	93	101	287	5.04
DESCRIPTIVENESS +	50	2	1	1	4	1.33
	51	5	5	6	16	5.33
	52	6	5	6	17	5.67
	53	6	6	6	18	6.0
	54	6	6	6	18	6.0
+	55	4	1	4	9	3.0
	56	6	6	5	17	5.67
+	57	1	2	1	4	1.33
	58	6	6	6	18	6.0
+	59	2	1	1	4	1.33
	60	5	5	6	16	5.33
	61	6	5	6	17	5.67
	62	6	6	6	18	6.0
	63	6	6	6	18	6.0
	64	6	5	6	17	5.67
SUM OF COLUMNS		73	66	72	211	4.69

Check #72

NTCS-A Useability Questionnaire Analysis

CATEGORY	QUESTION NO.	J.R.	N.E.	C.W.	SUM OF ROWS	MEAN SCORE
CONFIRMABILITY	1	6	6	5	17	5.67
	2	6	6	6	18	6.0
	3	6	6	6	18	6.0
	4	6	6	6	18	6.0
	5	6	6	6	18	6.0
	6	5	5	6	16	5.33
	7	6	5	6	17	5.67
	8	6	6	6	18	6.0
	9	6	6	6	18	6.0
	10	6	6	6	18	6.0
	11	6	5	6	17	5.67
	12	5	5	4	14	4.67
	13	0	0	0	0	0
*	14	0	0	0	0	0
	15	6	4	6	16	5.33
+	16	1	1	1	3	1.0
SUM OF COLUMNS		77	73	76	226	5.38
CONTROLLABILITY	17	5	5	5	15	5.0
	18	5	5	6	16	5.33
+	19	2	1	1	4	1.33
	20	5	5	6	16	5.33
+	21	3	1	1	5	1.67
	22	6	5	6	17	5.67
*	23	0	0	0	0	0
	24	6	5	6	17	5.67
	25	6	5	6	17	5.67
	26	5	5	6	16	5.33
+	27	1	1	1	3	1.0
	28	4	4	5	13	4.33
	29	5	5	5	15	5.0
+	30	3	2	2	7	2.33
SUM OF COLUMNS		56	49	56	161	4.13

Sample Data set (3 respondents)

CATEGORY	QUESTION NO.	J.R.	N.E.	C.W.	SUM OF ROWS	MEAN SCORE
CONSISTENCY	65	6	6	6	18	6.0
	66	6	6	6	18	6.0
	67	4	4	4	12	4.0
	68	5	5	5	15	5.0
	69	6	5	6	17	5.67
	70	6	6	6	18	6.0
	71	5	5	5	15	5.0
*	72	~~0~~ 5	~~0~~ 5	~~0~~ 5	~~0~~ 15	5.00
	73	6	6	6	18	6.0
	74	6	5	6	17	5.67
SUM OF COLUMNS		~~50~~ 55	~~48~~ 53	~~50~~ 55	~~148~~ 163	~~5.48~~ 5.43
SIMPLICITY	75	6	6	6	18	6.0
	76	6	6	6	18	6.0
	77	5	6	6	17	5.67
	78	6	6	6	18	6.0
	79	6	6	6	18	6.0
	80	6	5	5	16	5.67
	81	6	5	6	17	5.67
	82	5	5	6	16	5.33
	83	5	5	5	15	5.0
	84	6	5	6	17	5.67
	85	6	5	6	17	5.67
	86	6	6	6	18	6.0
	87	5	6	6	17	5.67
	88	5	5	5	15	5.0
SUM OF COLUMNS		79	77	81	237	5.64
GENERAL	89	6	6	6	18	6.0
	90	6	5	5	16	5.33
	91	6	6	6	18	6.0
	92	5	6	6	17	5.67
	93	6	6	6	18	6.0
	94	5	6	6	17	5.67
	95	5	6	5	16	5.33
SUM OF COLUMNS		39	41	40	120	5.71

+ = Scored less than 3.5
* = *Mean Scores* adjusted for "Not Enough Info" Responses

J.R. = Mr. Jon Roth
N.E. = Mr. Norvin Eyrich
C.W. = Mr. Charles Walters

NTCS-A USEABILITY QUESTIONNAIRE ANALYSIS

Calculations:

$$\text{Overall (all Categories)} = \frac{226+161+287+211+148+237+120}{(3x91)+(2x0)+(1x0)+(0x4)} = \frac{1390}{273} = \mathbf{5.09}$$

Applying questions weights:

CATEGORY	MEAN SCORE	CATEGORY WT.	MEAN WT.
Confirmability	5.38	.1666	.8967
Controllability	4.13	.1548	.6392
Workload Suit.	5.04	.2262	1.1400
Description	4.69	.1786	.8375
Consistency	5.48	.1071	.5871
Simplicity	5.64	.1667	.9400
		TOTAL MEAN =	5.04

Overall score of 5.04 indicates that the software and operator/user manuals are acceptable. The General category (overall useability) indicates a "very good" to "excellent" level of acceptability.

NTCS-A Useability Problem Areas. Question scores 3.5 or less are as follows:

CATEGORY	QUESTION NO.	SCORE	QUESTION ISSUE
1. Workload Suitability	44	2.67	Not easy to locate information within user/operator manual
2. Descriptive.	63	2.0	User/Operator Manual does not contain a useful index
	64	3.3	User/Operator Manual does not contain a useful glossary
3. Simplicity	84	2.5	Appropriate checklists are not available
	85	2.5	Number checklists required not manageable

LINES AND CABLES

1. Cables should be routed so as to be accessible for maintenance.

2. Test cables terminating on control panels should not interfere with controls and displays.

3. Cables routed through holes in metal should be protected from mechanical damage by grommets or other protective devices.

4. Cables used for checking units should be long enough for the purpose.

5. When cable clamps are used, they should be spaced approximately every 12".

6. Gas, fluid, and electrical conduit lines should be properly identified.

7. Cables should be labeled to indicate the equipment they are used with and the connectors with which they are mated.

8. Cables should be routed so as not to be pinched by doors, walked on, used for hand holds, or bent.

9. Cables containing individual insulated conductors in a common sheath should be coded.

FASTENERS AND CONNECTORS

1. There should be a 1" minimum space between connectors for grasping.

2. Connecting plugs and receptacles should be color-coded.

*3. Plugs of one voltage should be incapable of being inserted into receptacles of another voltage.

4. Noninterchangeable connectors should be used for different uses.

*5. Fasteners should require only one complete clockwise turn to tighten and one complete counterclockwise turn to loosen.

6. Use identical screw/bolt heads where possible.

7. Aligning pins should extend beyond electrical pins to ensure alignment before pins engage.

8. Use stripes, arrows, etc. to show position of aligning pins.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

CASES AND COVERS

1. Cases should be sufficiently larger than the units they cover to prevent damage when the case is removed and replaced.

2. It should be obvious when a cover is not secured even though it is in place.

*3. Cases should be capable of being lifted from units rather than units lifted from cases.

4. Edges and corners on cases and covers should be rounded or otherwise finished to prevent injury to personnel.

5. Guides, tracks, and stops should be provided as necessary to facilitate handling and prevent damage or injury.

6. If the method of opening a cover is not obvious, instructions should be prominently displayed on cover.

7. Proper orientation of a unit in a case should be obvious through design or labels.

8. Stowage locations should be labeled.

9. Bulkhead, brackets, and other units should not interfere with removal or opening of covers.

10. Mounting screw holes in covers that attach to the chassis should be large enough so that perfect case alignment is not necessary.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

ACCESS OPENINGS

1. When possible, an access should be available whenever frequent maintenance operations would otherwise require removing a case, opening a fitting, or dismantling a component.

2. Size and shape of openings for physical access should agree approximately with dimensions in Figure 4-3.

3. Access covers that are not completely removable should be self-supporting when opened.

4. Accesses and covers should avoid sharp edges to preclude injury.

5. Accesses should be labeled to indicate items to be accessed, operations to be accomplished, and any hazards beyond access.

6. Access warnings should be large enough to be read at a reasonable distance.

*7. Space for gloved hand or clothed body should be provided in access.

8. Access covers should be equipped with grasp areas for openings.

*9. Accesses should be large enough to permit required operations.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

MINIMAL TWO-HAND ACCESS OPENINGS WITHOUT VISUAL ACCESS

Reaching with both hands to depth of 6 to 19.25 inches:

Light clothing:	Width:	8" or the depth of reach*
	Height:	5"
Arctic clothing:	Width:	6" plus 3/4 the depth of reach
	Height:	7"

Reaching full arm's length (to shoulders) with both arms:

Width:	19.5"
Height:	5"

Inserting box grasped by handles on the front:

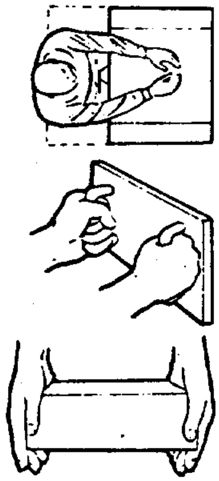
1/2" clearance around box, assuming adequate clearance around handles

Inserting box with hands on the sides:

Light clothing:	Width:	Box plus 4.5"
	‡Height:	5" or 0.5" around box*
Arctic clothing:	Width:	Box plus 7"
	‡Height:	8.5" or 0.5" around box*

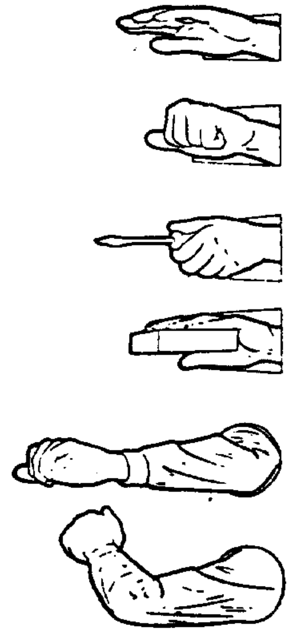
* Whichever is larger.

‡ If hands curl around bottom, allow an extra 1.5" for light clothing, 3" for arctic clothing.



MINIMAL ONE-HAND ACCESS OPENINGS WITHOUT VISUAL ACCESS

	Height		Width		
Empty hand, to wrist:					
Bare hand, rolled:	3.75"	sq	or dia		
Bare hand, flat:	2.25"	x	4.0"	or	4.0" dia
Glove or mitten:	4.0"	x	6.0"	or	6.0" dia
Arctic mitten:	5.0"	x	6.5"	or	6.5" dia
Clenched hand, to wrist:					
Bare hand:	3.5"	x	5.0"	or	5.0" dia
Glove or mitten:	4.5"	x	6.0"	or	6.0" dia
Arctic mitten:	7.0"	x	8.5"	or	8.5" dia
Hand plus 1" dia object, to wrist:					
Bare hand:	3.75"	sq	or dia		
Gloved hand:	6.0"	sq	or dia		
Arctic mitten:	7.0"	sq	or dia		
Hand plus object over 1" in dia, to wrist:					
Bare hand:	1.75" clearance around object				
Glove or mitten:	2.5" clearance around object				
Arctic mitten:	3.5" clearance around object				
Arm to elbow:					
Light clothing:	4.0"	x	4.5"	or	4.5" dia
Arctic clothing:	7.0"	sq	or dia		
With object:	Clearances as above				
Arm to shoulder:					
Light clothing:	5.0"	sq	or dia		
Arctic clothing:	8.5"	sq	or dia		
With object:	Clearances as above				



MINIMAL FINGER ACCESS TO FIRST JOINT

Push button access:	Bare hand:	1.25" dia
	Gloved hand:	1.5" dia
Two finger twist access:	Bare hand:	object plus 2.0" dia
	Gloved hand:	object plus 2.5" dia

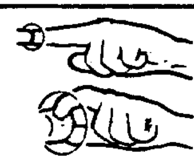


Figure 4-3. Arm and hand access dimensions.

MAINTAINABILITY (TOOLS, TEST EQUIPMENT, TEST POINTS)

1. Special tools should be securely mounted in equipment and accessible to the technician.

2. Test points should be accessible, clearly marked, and close to the units with which they are used.

3. Space should be provided within portable test equipment to store leads, probes, manuals, and tools.

4. Calibration and adjustment controls with limited motion should have mechanical stops to prevent damage.

5. Test points, cables, and connectors should be accessible and visible during maintenance.

6. If nonvisual screwdriver adjustments are required, they should have shaft guides.

7. Displays to indicate failure of equipment units should be provided.

*8. Lamps and light bulbs should be removable from front of display panel without special tools.

9. Sensitive adjustment points should be guarded against accidental disturbance.

10. Larger units should not be mounted to deny access to small ones.

11. Positive and negative battery terminals should be of different size and marked "+" and "-."

12. Items frequently removed for test should, where possible, be mounted on rollout racks, slides, or hinges.

13. Lamp replacement should be possible with power on and without danger.

14. Critical units requiring fast maintenance should be more accessible than other units except that, where criticality is not a factor, units requiring most frequent access should be most accessible.

15. Field removable units should be replaceable with common hand tools.

16. Where applicable, interlocks should be provided to disconnect equipment that would otherwise be damaged by withdrawal of racks or drawers.

Notes. Maintainability deficiencies can also be ascertained by asking personnel to perform required maintenance actions and then interviewing them and/or by asking them to rate the adequacy of maintainability features. For rating scale development, see Section Three of this report.

Criteria marked with an asterisk can be evaluated by physically performing required actions.

COMMUNICATIONS

1. Communication devices should be located within easy reach of the work station.

2. Foot control of the communications device should be available for the seated operator who needs both hands.

3. Headsets should be provided for high noise workspaces.

*4. The speaker should hear his own voice in the headset in phase with his speech.

5. Noise cancelling or bone conduction microphones should be utilized in high noise environments.

6. Microphones, headphones, and telphone headsets should permit normal hands-free operation.

*7. If, in actual use, the operator finds that the volume permitted by his device is too low for him to communicate easily or if he notes any distortion of his speech or of the message received, he should communicate with a Human Engineering specialist.

<u>Notes</u>: Communications adequacy can also be ascertained by interviewing personnel and/or by asking them to rate the adequacy of communications practice. For rating scale development, see Section Two of this report.

Criteria marked with an asterisk can be evaluated by physically performing required actions.

SECTION FIVE--SELF REPORT RATING SCALES

This section describes rating scales which can be used to elicit quantitative judgments from test personnel concerning the adequacy of various aspects of the equipment/system being tested.

Introduction

Among the techniques available to the Marine Corps evaluator for gathering personnel performance data are those involving subjective personnel reactions to characteristics of the equipment/system being tested.

Such subjective data can supplement and amplify objective data. Where objective data cannot be secured, as, for example, to determine an operator's reaction to vehicle driving or riding qualities, subjective methods may be the only evaluation methods available.

This section describes a number of ways of gathering subjective data in quantitative form, principally by means of rating scales to be completed by test participants at the conclusion of a test operation.

Rating Scales

The scales described report the operator's reactions to the following system characteristics:

1. Environmental conditions (noise, temperature/humidity, vibration).
2. Illumination.
3. Handling (driving) qualities of a vehicle.
4. Riding qualities of a vehicle.
5. Control accessibility.
6. Display readability.
7. Control-display arrangement.
8. Information presented.
9. Workspace.
10. External visibility.
11. Vehicle entrance/exit.
12. Accessibility of internal components.
13. Ease of troubleshooting malfunctioning equipment.

14. Test points.

15. General equipment maintainability.

16. Safety.

17. Operating procedures and/or technical manuals.

18. Workload.

19. Communications.

In addition, a Critical Incident Report form and a Satisfaction Checklist are described.

The scales take advantage of the equipment operator/maintainer's experience with the system under test and his general background on comparable systems to tap his evaluation of the system and of personnel performance in utilizing the system. Any or all of these scales can be included in any questionnaire administered to OST personnel or they can be utilized as part of an interview with these personnel. They require minimal explanation to respondents and can be completed very quickly. The evaluator can select his scales to investigate those system aspects about which he wishes information. It is not necessary that all the scales described in this report be used in the same test evaluation.

The evaluator is interested primarily in having the forms available, but he may also wish to know the theoretical foundation of these scales. Information on this point is provided in a Rationale which follows each scale. In a number of scales, the underlying dimension is the amount of effort the operator has to expend in doing his job as related to the particular equipment/system characteristic under investigation. In others, the scale dimension may be the difficulty the operator has experienced in relation to the factor (e.g., workspace) being evaluated. A few scales have more than one dimension, each one contributing to the scale value.

The scales are designed to supply a number representing the operator's evaluation. They therefore differ from the checklist evaluation performed by someone other than the test participant. These scales can be used to supplement the Human Engineering Checklist described in Section Four of this report which was designed to be used by the evaluator himself.

The scales in this report are oriented vertically, rather than horizontally across the paper. This is because a horizontal orientation (which saves paper) tends to crowd the written ratings and descriptors unduly. If space in the questionnaire is a desideratum, the scales can be reoriented horizontally.

Since few operators can differentiate more than five major intervals on a continuum, the scales have five points representing a continuum ranging from Excellent (1.0) through Good (2.5), Fair (4.0), and Poor (5.5) to Unacceptable (7.0). Each point is identified by a behavioral descriptor which "anchors" the point. Intermediate intervals between anchor points (1.5, 2.0, 3.0, 3.5, 4.5, 5.0, 6.0, 6.5) are indicated without being numbered. Test personnel are asked to check anywhere along the scale, paying particular attention to the behavioral descriptors; the resultant checks can easily be transformed into numerical equivalents. Values between anchor and intermediate points are interpolated visually.

The advantage of having a numerical rating of the operator's response to various system characteristics is that the evaluator can treat these ratings statistically (averaging several operator ratings, determining their variability by means of a standard deviation, and comparing the mean ratings of one equipment or one test condition with another).

1. REACTIONS TO ENVIRONMENTAL CONDITIONS (Noise, temperature/humidity, vibration)

RATING ITEM

(Noise, temperature/humidity, vibration)* affected my performance during my operation of the vehicle/equipment, such that

Rating		Descriptor
Not at all	1.0	No discomfort noted; no increase in effort required; no performance impairment.
Slightly	2.5	Minimal discomfort; slight increased effort to perform tasks; minimal performance impairment.
Moderately	4.0	Moderate increase in effort required and/or some discomfort noted; some performance impairment.
Seriously	5.5	Considerable increase in effort required to perform tasks; great discomfort; considerable performance impairment.
Excessively	7.0	Maximum effort required to perform tasks and/or extraordinary discomfort; serious performance impairment.

If rating is seriously or excessively, please comment further.

*Select one as appropriate.

Rationale: This item would be used to determine the impact of any undesirable environmental condition on the vehicle/equipment operator. Instrumentation can determine whether any of these conditions would be painful or even damaging to the operator; however, the effect on the operator's performance (short of these extreme conditions) can be most easily determined by the operator's self report. This scale has three dimensions: the amount of (1) effort required to perform the tasks, (2) discomfort experienced, and (3) performance impairment noted. It is assumed that undesirable environmental conditions will increase all three dimensions.

2. ILLUMINATION

RATING ITEM

Because of the illumination within the vehicle or operating compartment, tasks requiring fine visual discrimination

Rating		Descriptor
Excellent	1.0	Can be performed effortlessly.
Good	2.5	Can be performed with only slight effort.
Fair	4.0	Can be performed with moderate effort.
Poor	5.5	Can be performed only with considerable effort.
Unacceptable	7.0	Almost impossible to perform because of inadequate lighting.

If rating is poor or unacceptable, please comment further.

Rationale: This item has essentially the same rationale as item 1. The absolute level of illumination in the vehicle or operating compartment can be measured, but the effect of that illumination on the operator himself can be most readily determined by his self report. (One could of course set up an experiment to measure his performance under different levels of illumination, but this is usually not possible under OST conditions.) The emphasis in this scale is on tasks requiring fine visual discrimination, such as reading dials. The dimension employed in this scale is the amount of extra effort required of the operator by lack of proper illumination.

3. HANDLING (DRIVING) QUALITIES OF VEHICLE

RATING ITEM

The handling (driving) qualities* of my (aircraft, jeep, APC, etc.) ** were such that

Rating		Descriptor
Excellent	1.0	Little or no effort is required.
Good	2.5	Slight amount of effort is required.
Fair	4.0	Moderate amount of effort is required.
Poor	5.5	Considerable effort is required.
Unacceptable	7.0	Very strenuous effort is required.

If rating is poor or unacceptable, please comment further.

---

*Defined as ease of turning vehicle, starting/stopping, shifting gears, etc.
**Insert appropriate term.

Rationale: The dimension in the scale is again the amount of effort or additional effort required to drive the vehicle because of handling deficiencies. It is assumed that, when more than moderate effort is required, the vehicle design is poor or unacceptable. The less effort required, the better designed the vehicle.

4. RIDING QUALITIES OF VEHICLE

RATING ITEM

The riding qualities* of my (aircraft, jeep, APC, etc.)** were such that the ride was

Rating		Descriptor
Excellent	1.0	Always very comfortable.
Good	2.5	Generally comfortable; occasional slight bumpiness.
Fair	4.0	Occasionally very bumpy.
Poor	5.5	Always very bumpy; seat belts required.
Unacceptable	7.0	Heavy pitching/rolling; almost impossible to remain seated during ride despite seat belts.

If rating is poor or unacceptable, please comment further.

---

*Defined as smoothness of ride, sway, vibration, rattles, etc.
**Insert appropriate term.

Rationale: Riding must be differentiated from handling the vehicle. This scale describes what the driver or the passenger feels as a result of vehicle motion apart from efforts made to control the vehicle (item 3). Comfort (and its reverse, bumpiness) is the scale dimension and is assumed to affect operational performance. Since comfort is a subjective response, it cannot be measured objectively.

5. CONTROL ACCESSIBILITY

RATING ITEM

Were controls reachable when you were normally seated?

Rating		Descriptor
Excellent	1.0	All controls are reachable without effort.
Good	2.5	A few controls require slight effort to reach.
Fair	4.0	All controls are reachable but with some additional effort.
Poor	5.5	Considerable straining required to reach a few controls.
Unacceptable	7.0	All controls require considerable straining to reach.

If rating is poor or unacceptable, please comment further.

Rationale: This scale is based on two dimensions: the additional effort required to reach for controls, and the number of controls for which the effort is demanded. The more effort required to reach more controls, the less acceptable the control accessibility. It is possible to determine accessibility objectively by seating the operator, asking him to touch each control in turn, and noting the degree of muscular strain. However, this situation does not impose the same demand on the operator as does normal operations. The scale is designed to measure accessibility in the latter situation.

6. DISPLAY READABILITY

RATING ITEM

Were all displays readable from the normal operating position?

Rating		Descriptor
Excellent	1.0	All displays readable without effort.
Good	2.5	One or two displays require slight additional effort to read.
Fair	4.0	All displays readable but some eye straining required.
Poor	5.5	Intensive eye straining required to read some displays.
Unacceptable	7.0	All displays difficult to read even with intensive straining.

If rating is poor or unacceptable, please comment further.

Rationale: Readability is the ability to discriminate individual characters/numerals on the display. This scale includes two dimensions: amount of effort and number of displays. The greater the amount of eye strain and the more displays for which this eye strain is required, the less acceptable the display readability. Again, this factor can be measured objectively, but only with great difficulty in an OST context.

7. CONTROL-DISPLAY ARRANGEMENT

RATING ITEM

The way controls and displays are arranged on the equipment console is such that operating them was

Rating		Descriptor
Excellent	1.0	Extremely easy and effortless.
Good	2.5	Easy but required a little effort.
Fair	4.0	Slightly difficult, required moderate effort.
Poor	5.5	Very difficult, required considerable effort.
Unacceptable	7.0	Excessively difficult, required strenuous effort.

If rating is poor or unacceptable, please comment further.

Rationale: This scale assumes that the manner in which controls and displays are arranged affects equipment operation, in particular the ease or difficulty of that operation. To use the scale, the respondent must consider control-display arrangement in terms of two dimensions: the amount of effort required and the difficulty he experiences in manipulating those controls/displays. Objective determination of control-display arrangement is very difficult in an OST context.

8. INFORMATION PRESENTED

RATING ITEM

Information presented by displays was

Rating		Descriptor
Excellent	1.0	Extremely easy to understand, required almost no effort.
Good	2.5	Easy to understand but required a little effort.
Fair	4.0	Somewhat difficult to understand, required moderate effort.
Poor	5.5	Very difficult to understand, required very great effort.
Unacceptable	7.0	Either extremely difficult to understand or not enough information is available.

If rating is poor or unacceptable, please comment further.

Rationale: This scale contains two dimensions: ease of understanding the material communicated and the effort involved in doing so. The easier the understanding, the less effort, the better. It is assumed that all necessary information is being presented; when this is not true, the rating becomes Unacceptable.

The following checklist may be used with the preceding scale, or separately.

Check one or more of the following if they pertain to the information displayed on your equipment.

_____ Too much information presented at one time.

_____ Too much information must be combined from different displays.

_____ Information appears too quickly.

_____ Information changes too quickly.

_____ Some information is irrelevant to task.

_____ Not enough information.

9. WORKSPACE

RATING ITEM

Workspace within the vehicle or ground facility was such that there was

Rating		Descriptor
Excellent	1.0	No interference with others and no difficulty in performing own tasks because of space constraints.
Good	2.5	Very infrequent interference with others and/or slight difficulty in performing own tasks because of space constraints.
Fair	4.0	Occasional interference with others and/or moderate difficulty in performing own tasks because of space constraints.
Poor	5.5	Repeated interference with others and/or great difficulty in performing own tasks because of space constraints.
Unacceptable	7.0	Constant interference with others and excessive difficulty in performing own tasks because of space constraints.

If rating is poor or unacceptable, please comment further.

Rationale: Workspace is space available for performing jobs. It is assumed that, if workspace is restricted, the operator will interfere with or be interfered with by others in the same vehicle or facility and he will have difficulty in performing his tasks. Thus, there are two dimensions in this scale: frequency of interference and task performance difficulty. Again, this factor can be measured objectively, but only with great difficulty in the OST context.

10. EXTERNAL VISIBILITY

RATING ITEM

Visibility external to the vehicle was such that I could see out in every required direction

Rating		Descriptor
Excellent	1.0	Without effort.
Good	2.5	With only slight effort.
Fair	4.0	With moderate effort.
Poor	5.5	With great effort.
Unacceptable	7.0	With exceptional effort.

If rating is poor or unacceptable, please comment further.

Rationale: External visibility is defined as how much the operator can see out of windows or viewing ports. It is assumed that, if external visibility is limited but the operator must see outside to do his job, the more restricted the visibility, the more effort he will have to expend on viewing. Again, this factor can be measured objectively, but with great difficulty.

11. VEHICLE ENTRANCE/EXIT

RATING ITEM

Entrance to/exit from the vehicle in full combat gear is

Rating		Descriptor
Excellent	1.0	Extremely easy, extremely rapid.
Good	2.5	Easy, fast.
Fair	4.0	Neither particularly easy or difficult; speed satisfactory.
Poor	5.5	Somewhat slow, somewhat difficult.
Unacceptable	7.0	Very slow, very difficult.

If rating is poor or unacceptable, please comment further.

Rationale: From a performance standpoint, adequacy of entrance to and exit from a vehicle is determined by the speed and difficulty of performing this function. Hence, these two dimensions are included in this scale. It is possible to observe personnel entering/exiting the vehicle and to measure the time required to perform this function. The scale above provides an alternative to this procedure.

12. ACCESSIBILITY OF INTERNAL COMPONENTS

RATING ITEM

Internal components can be reached

Rating		Descriptor
Excellent	1.0	Without effort and without having to remove other components first.
Good	2.5	With only slight effort and after removing only a few other components first.
Fair	4.0	With moderate effort but minimal difficulty; a moderate number of other components must be removed first.
Poor	5.5	With some difficulty; many components must be removed first.
Unacceptable	7.0	Only with great effort/difficulty and after removing an excessive number of other components first.

If rating is poor or unacceptable, please comment further.

Rationale: Accessibility of internal components is defined by the number of other components one must remove first and (as a consequence of this) by the effort involved in reaching the desired component. It is unlikely that this type of accessibility can be objectively measured without great difficulty.

13. TEST POINT AVAILABILITY

RATING ITEM

Test points are available to check

Rating		Descriptor
Excellent	1.0	Every important component.
Good	2.5	Most important components.
Fair	4.0	Some important components.
Poor	5.5	A few important components.
Unacceptable	7.0	Almost no components.

If rating is poor or unacceptable, please comment further.

Rationale: Test point availability is considered to be a most significant factor affecting the capability to troubleshoot an equipment. There are other factors, such as the accessibility of these test points, that bear on troubleshooting capability, but none is as important as availability. The dimension represented on the scale is the correspondence between the number of test points and the number of major components that require testing. One could, of course, check this factor out objectively by examination of the equipment design, but we have taken the tack that the test participant, in working with the equipment, is in the best position to know how this correspondence works in actual practice.

14. EASE OF TROUBLESHOOTING MALFUNCTIONING EQUIPMENT

RATING ITEM

The malfunctioning component can usually be discovered with

Rating		Descriptor
Excellent	1.0	Almost no effort, difficulty or time.
Good	2.5	Slight effort, difficulty or time.
Fair	4.0	Moderate effort, difficulty or time.
Poor	5.5	Great effort, difficulty and time.
Unacceptable	7.0	Exceptional effort, difficulty and time.

If rating is poor or unacceptable, please comment further.

Rationale: This scale deals with troubleshooting as a total function. Equipment characteristics (such as accessibility of internal components and test point availability, scales for which were described previously) affect troubleshooting but are not maintenance functions per se.

It is possible to measure the operator's troubleshooting proficiency on the job, but to do so requires that either an observer must measure repair time or the operator himself must report this time. Often this is not feasible in the context of a test operation. Moreover, objective troubleshooting measures do not get at the effort/difficulty dimension represented in this scale (along with time, which these objective measures do deal with). Hence, use of such a scale can provide useful information describing the ease or difficulty of keeping an equipment running.

15. GENERAL EQUIPMENT MAINTAINABILITY

RATING ITEM

Preventive and corrective maintenance can be accomplished

Rating		Descriptor
Excellent	1.0	With almost no effort; very rapidly.
Good	2.5	With slight effort; quickly.
Fair	4.0	With moderate effort; acceptable time.
Poor	5.5	With much difficulty; long time.
Unacceptable	7.0	With excessively strenous effort, difficulty and time.

If rating is poor or unacceptable, please comment further.

Rationale: This scale describes general equipment maintainability, including both preventive and troubleshooting aspects. Consequently, it subsumes the preceding maintainability scales. Because of its generality, however, it is assumed that one would wish to use this scale only in conjunction with one or more of the previous ones. This scale provides a summary quantitative evaluation of maintainability from the technician's standpoint. It is unlikely that such a summary statement could be made objectively except as a conclusion based on a number of empirical tests, which might be difficult to perform in an OST context. The scale dimensions are those most pertinent to maintainability: effort and speed.

16. SAFETY

RATING ITEM

Required safety equipment

Rating		Descriptor
Excellent	1.0	All required safety equipment is available and accessible.
Good	2.5	Almost all required safety equipment is available and accessible.
Fair	4.0	Most required safety equipment is available and accessible.
Poor	5.5	Only certain items of required safety equipment are available and accessible.
Unacceptable	7.0	Very few required items of safety equipment are available and accessible.

If rating is poor or unacceptable, please comment further.

Rationale: This scale assumes that all required safety equipment must be available and accessible in the vehicle, ground facility, or weapon system. To the extent that less than all such equipment is available, the system is deficient.

17. OPERATING PROCEDURES AND/OR TECHNICAL MANUALS

RATING ITEM

(Operating Procedures and/or technical manuals)* can be understood and followed

Rating		Descriptor
Excellent	1.0	With no effort or difficulty at all.
Good	2.5	With little effort and difficulty.
Fair	4.0	With moderate effort and difficulty.
Poor	5.5	With great effort and difficulty.
Unacceptable	7.0	With extreme effort and difficulty.

If rating is poor or unacceptable, please comment further.

* Select one

Rationale: A major factor affecting how well personnel perform their jobs is their ability to understand and follow the procedures and technical documentation they must employ. This scale is designed to measure the operator's evaluation of this factor. Errors in performing procedures can of course be measured objectively, but do not describe the effort involved in using procedures and technical manuals. It is therefore almost impossible to evaluate procedures and technical manuals objectively; i.e., without securing the operator's opinion on the matter. As usual, the effort factor is the scale dimension.

18. WORKLOAD

RATING ITEM

My job can be performed effectively

Rating		Descriptor
Excellent	1.0	With no difficulty or effort at all.
Good	2.5	With little effort and difficulty.
Fair	4.0	With moderate effort and difficulty.
Poor	5.5	With great effort and difficulty.
Unacceptable	7.0	With extreme effort and difficulty.

If rating is poor or unacceptable, please comment further.

Rationale: It is often of interest to determine just how easy or difficult it is for test personnel to operate their vehicle or equipment. This scale is designed to measure the effort/difficulty associated with that operation. Since the evaluation of this factor is largely subjective, no objective measurement is feasible, particularly within the context of OST. Note that the scale deals with effective performance and the difficulty associated with effective performance. It is assumed that the respondent knows what comprises effective performance.

An alternative way of getting at workload is in terms of the following scale, which is particular to a specific equipment:

Operation (or maintenance)* of the __________ (fill in) equipment is

Very Easy	Easy	Neither Easy nor Difficult	Difficult	Extremely Difficult

*Select one

19. COMMUNICATIONS

RATING ITEM

Intercom/radio messages between personnel are

Rating		Descriptor
Excellent	1.0	Highly intelligible; no distortion at all.
Good	2.5	Fairly intelligible; little distortion.
Fair	4.0	Acceptably intelligible; moderate distortion.
Poor	5.5	Barely intelligible; great distortion.
Unacceptable	7.0	Almost unintelligible; extreme distortion.

If rating is poor or unacceptable, please comment further.

Rationale: The dimension in this scale is the physical distortion of the communicated message, as perceived by the recipient of the message. Physical distortion of the signal can be ascertained by objective methods, but this may be difficult in the OST context.

CRITICAL INCIDENT REPORT

Please report any incident during the operation of the vehicle, equipment, or weapon system that resulted or could have resulted in an abnormal or potentially dangerous situation (this includes any equipment malfunction).

Check the stage of the operation in which the critical condition occurred and then describe it in as much detail as you think necessary.

Stage of Operation	Subsystem Involved

(This listing must be provided by the evaluator; it is specific to the equipment being evaluated.)

DESCRIPTION

Symptoms (How did you first notice this problem?)

Diagnosis (How did you determine what the problem was?)

Causes (What produced the problem?)

Remedy (What did you do to solve the problem?)

Rationale: Any untoward incident, event, or phenomenon occurring during the test operation may indicate a deficiency that needs improvement before the system is released to operational use. This report is completed at the conclusion of a test run or operation--but only if an incident worthy of note has occurred. Note that this report can also be used to report equipment malfunctions (thus combining data on both personnel and equipment factors).

SATISFACTION CHECKLIST

On a scale from 0 to 100, where 0 represents complete dissatisfaction and 100 represents complete satisfaction, please check your degree of satisfaction with the following factors:

Factor	0	25	50	75	100
1. Noise, temperature/humidity, vibration	0	25	50	75	100
2. Illumination	0	25	50	75	100
3. Vehicle handling/driving qualities	0	25	50	75	100
4. Display readability	0	25	50	75	100
5. Control accessibility	0	25	50	75	100
6. Control-display arrangement	0	25	50	75	100
7. Information understandability	0	25	50	75	100
8. Workspace	0	25	50	75	100
9. External visibility	0	25	50	75	100
10. Vehicle entrance/exit	0	25	50	75	100
11. Internal component accessibility	0	25	50	75	100
12. Test point availability	0	25	50	75	100
13. Ease of troubleshooting	0	25	50	75	100
14. Equipment maintainability, general	0	25	50	75	100
15. Communications	0	25	50	75	100
16. Safety	0	25	50	75	100
17. Operating procedures	0	25	50	75	100
18. Technical manuals	0	25	50	75	100
19. Operator workload	0	25	50	75	100

Rationale: This scale can be used as a very abbreviated summary of the preceding individual scales but obviously does not provide as much information as the latter. It does not indicate the reason for the respondent's satisfaction/dissatisfaction.

SECTION SIX--INTERVIEW QUESTIONS

This section presents questions that evaluators can use in interviewing OST participants to secure data on the adequacy (from a personnel standpoint) of the equipment/system under test. The questions are of two types: general and specific. The interview begins with general questions to determine whether any personnel performance problems have been noted by the operator/maintainer. More specific questions follow to cover the range of equipment/job variables that could influence performance.

The questions listed in this section are to be asked in an interview situation. It is assumed that if the Operational System Test consists of several test operations or cycles, the test operator (or maintainer) will be interviewed following each such operation/cycle.

These questions cover the most important topics that describe how test personnel operate and maintain the system under test. The following topics are covered:

1. General questions.
2. Equipment characteristics.
3. Environment.
4. Job aids.
5. Safety.
6. Manning.
7. Training.
8. Information.
9. Communications.
10. Maintenance.

These topics are roughly the same as those covered in previous sections, but have been created specifically to be used as part of an interview.

In all cases the interviewer should begin his interview with the general questions, because these permit the interviewee (respondent) to produce information that he considers most relevant and important.

If the respondent produces significant material as a result of the general questions, he should be allowed to direct the interview into channels he wishes to follow--as long as these channels are relevant and, in the opinion of the interviewer, productive.

The more specific questions should be asked when the respondent has exhausted the material he produces as a result of the general questions or if he has not touched on any of these topics previously. They should be asked if the respondent appears uncertain as to how to proceed, is not producing any information of value, or is generally noncommunicative.

If several test operations are performed and the operator has been interviewed following each test operation, it is permissible to ignore those questions that have been asked previously and for which the test situation has not changed so that previous answers are still applicable (but not if the operator has been learning on successive operations). Under these circumstances, only the general questions need be asked.

Although it is unnecessary for the interviewer to follow the precise wording of these questions or the order in which they are presented here, the general tenor of the questions should not be changed.

The emphasis in these questions is on problems or difficulties experienced by the operator/maintainer, rather than on how well these personnel have performed. These difficulties indicate inadequacies in the equipment, procedures, etc., that should be cleared up before the system is released to general use.

GENERAL QUESTIONS

> 1. Can you recall any difficulty or problem (no matter how small) you experienced during the previous test operation(s)? If so, what was that difficulty? What do you think was responsible for it? What actions did you take (i.e., what did you do about it)?

Comment

It is always best to begin an interview with a general all-encompassing question such as the above. This permits the respondent to focus on what appears to him to be the most important of the difficulties he has experienced.

> 2. Did you observe anyone else in your team having any difficulty? If so, what was the difficulty? Do you know why he had this difficulty? What action did this individual take to relieve the difficulty?

Comment

This question is asked only if the operation is a team affair. All personnel on the team should be interviewed to cross-check individual answers.

> 3. Were there any characteristics of the job, such as the equipment, procedures, technical manuals, tools, weapons, etc., that made it difficult for you to do your job? How do you know that you had more than your usual difficulty?

Comment

Again, a general queston that allows the respondent to select what he considers the most important topic on which to zero in. It is possible that the respondent in answering question 1 will also discuss equipment/job characteristics that presented difficulty; but this question should be asked in any event, since it directs the interview to the equipment/job as a whole. It is desirable also to return to this question later in the interview in the following manner: We have discussed certain characteristics of the equipment/job that were not optimal. Can you think of any other characteristic that bothered you?

4. Assume that someone less skilled than you had to do your job. Would there be anything about the equipment, procedures, or the job as a whole that would cause a less skilled man difficulties?

Comment

It is assumed that the operator being interviewed is reasonably skilled. (Information about the type and level of training the interviewee has received should be ascertained prior to this interview.) However, it is entirely possible (even likely) when the equipment enters the operational inventory that someone less skilled will perform the job. An equipment/job feature that presents no difficulty to a skilled man may, however, have entirely different effects on a less skilled man. Test personnel can often estimate what these effects would be, and it is important to know these effects because if they are severe enough, some equipment or procedural change might be necessary. In asking this question, it may be necessary to provide an explanation of the reason for asking it, as described in the Comment.

5. Did the equipment fail in any way to perform as it was supposed to do (in any respect, no matter how small)? If so, do you know why? What did you do in reaction?

Comment

Any deviation ought to be routinely reported and might well be, but it is desirable to remind the interviewee about this possibility.

6. Did the test operation as a whole, or your job during the test, take much longer than you had expected it to take? How much longer? Do you know why? Did this extra time affect your performance in any way?

Comment

If the test operation took substantially longer than was planned, this indicates that something is wrong and the operating procedure may have to be changed. In any event, the evaluator must find out if the extra time affected the operator's performance.

7. Can you think of any changes in equipment, procedures, or the job as a whole that must be made to accomplish the mission? That should be made, if at all possible? Or that could improve the equipment, procedures, or job as a whole?

Comment

Test personnel often have excellent suggestions for improving their job (e.g., simplifying it, making it easier, or more accurate) and these should be elicited. Asking this question, moreover, cues the interviewee to think of deficiencies in equipment operation that might not previously have been reported.

8. Could you operate the equipment in accordance with the procedures you had been taught (that are in the manual)? If not, did you add any steps? Delete any steps? Perform one or more steps differently? What were these steps and why did you make these changes?

Comment

The evaluator should routinely ask about any procedure changes that were required because these changes should be included in revised technical manuals and might indicate a problem in the equipment/job.

9. In your opinion, would personnel in the field have any difficulty in operating/maintaining the equipment?

Comment

Assuming that the test personnel have worked in the Fleet Marine Force prior to becoming test personnel, they will remember their past experience.

10. Is every part of the equipment necessary?

Comment

Test personnel may have found no use for a designed feature that was considered essential during design. There may be valid justification for elimination of costly "extras."

11. Did you experience any difficulty as a consequence of operating the equipment for a prolonged period of time?

Comment

Equipment should be evaluated for a full shift and longer. Field or combat operations often require extended operations.

EQUIPMENT CHARACTERISTICS

In general, questions concerning equipment characteristics should follow general questions, although, if difficulties described in relation to general questions deal with some job aspect other than controls and displays, the interview should logically proceed on those other aspects, returning to controls and displays later.

Note. If one receives a negative answer to any of the following items (i.e., there is no difficulty), the interviewer would not proceed with the follow-up questions included in that item.

> 1. Were the controls difficult to operate? Any particular controls? Do you know why the controls are difficult to operate? How does the difficulty show itself (e.g., mushiness, sluggishness)? What was the effect of the difficulty on your performance?

Comment

If controls are difficult to operate, the effect on personnel performance is likely to be severe, particularly if the controls are in a vehicle. The interviewer should ask whether the operator knows why the difficulty exists, even though the operator often is unable to answer. The symptoms of the difficulty and the effect of the difficulty on the operator's performance may serve as clues to diagnosing the problem and determining its importance.

> 2. Were any of the controls difficult to reach? Which ones? How important are these controls? What effect does this difficulty have on your performance?

Comment

This question deals with control accessibility. It is necessary to determine the importance of the inaccessible controls and the effect of their inaccessibility on performance to properly evaluate the significance of the problem raised.

> 3. Were any of the displays (e.g., meters, indicators) difficult to read? Which ones? How important are these displays? Why were they difficult to read? What was the effect on your performance?

Comment

As in question 2, the interviewer seeks to determine which displays are difficult to read, how important these are, and the effect on performance, because these indicate just how important the problem is.

4. Did the displays provide all the information needed to do the job? What information was missing? Was there too much information? Was any of the information unnecessary or irrelevant to the job? How important was this factor in affecting your performance?

Comment

Information presented via displays may be too much, too little, unnecessary, or irrelevant. This question seeks to distinguish among these possibilities. The specific content of the information presenting the problem should be ascertained.

5. Were any of the displays difficult to understand? What precisely about the displays was difficult to understand? Which displays? How did this difficulty affect your performance?

Comment

Interviewees may have problems describing precisely what is meant by understanding. What we refer to here is the interpretation of the meaning of the information presented.

6. Did you have any difficulty reading the lettering or indicator lights when they were unlit? Lit? Which ones?

Comment

Many indicators are identified by labels which must be read even though the indicator is unlit.

7. Did any of the controls or displays seem unnecessary to perform the job? Which ones?

Comment

Ordinarily one would expect every control/display on the equipment or in the vehicle to be needed to perform the task. Occasionally, however, the nature of the job changes between the original design and the OST. This question enables the interviewer to check on this factor.

ENVIRONMENT

1. Was the lighting in your area (or vehicle) inadequate at any time for you to operate with maximum efficiency? Too little lighting? Too much lighting (glare)? What was the effect of this on your job performance?

2. Was any area (or any part of the vehicle) in which you worked excessively noisy, improperly ventilated, too cool, or too warm? How did this affect your performance?

3. Was there too much vibration in the vehicle when it was driven? Did this affect your work? In what way? How much?

4. Was there insufficient room around the equipment you operated so that it was difficult to move about? Was this true of the equipment in general or of a specific equipment only? What were the effects of this on your job? How great?

5. How difficult is it to get in and out of your vehicle? Does this affect your job performance? In what way?

Comment

These questions are obvious and consequently specific comments are not made about individual items. The operator's working area (i.e., his environment) may affect the efficiency with which he performs his job and so the evaluator will wish to examine the various aspects of that environment. Particular emphasis should be placed on the effect of these environmental factors on performance. If a factor has little effect on job performance, it will be unnecessary to probe deeper.

1. Are the tools and equipment you would need for maintenance available? Appropriate? Satisfactory? Are there any special tools you might need that are not available to you? If so, which ones?

2. Are all authorized spare parts available? Were any spare parts required and not available?

Comment

Tools and spare parts for maintenance fall into the category of job aids. If a problem in relation to these arises, it is probably because some of these tools/spares are either inappropriate or missing. The evaluator will also wish to know whether any tools, equipment, or spare parts are required that were not anticipated during design.

SAFETY

1. Is there any safety equipment you need that has not been provided?

2. Are there any desirable safety features (e.g., interlocks), that have not been included in the design of the equipment you operate? What are these features? How important are they?

3. Are there any safety hazards in the vehicle or area that you noticed? If so, what are they? Is all safety information conspicuously posted? Can anything be done to make it easier for the operator to heed these warnings?

4. Are all required safety equipment available and accessible in your area (vehicle)? If not, what is missing? Is there any safety equipment you need that has not been provided?

MANNING

1. Could you have used more men to do the job than were assigned to your team? If so, how many and of what type and what skill level should they have been? Could you have used fewer men to do the job? If so, which ones would you eliminate?

2. Was anyone on your team overloaded? Excessively fatigued by the end of the test operation? Why? What effect did this have on overall performance?

Comment

The reason for asking these questions is to verify that the appropriate number of personnel have been assigned to perform the job of operating or maintaining the test vehicle, weapon, etc. The question on workload seeks to determine indirectly whether more personnel are needed (if anyone is overloaded, presumably he needs help to carry the load).

TRAINING

1. Do you feel that the training you were given for this job was appropriate? Inappropriate? In what ways? What would you recommend to improve the training?

2. Are the men in your team properly qualified in terms of training?

3. What items were missing from the training you received that should be added? Was there anything about the training you received which you considered unnecessary or which you did not understand? Did you receive enough training to do the job?

4. What parts of the training were most important for safe, efficient operation? What parts were least important?

<u>Comment</u>

The OST is the first opportunity the Marine Corps has to check on the adequacy of the projected training that will be given personnel to operate the system. Test personnel will have been given factory training, but until OST the opportunity to check the adequacy of the training against performance has been lacking. The above items seek to gather information on the adequacy of training plans.

1. Do you feel that the procedure for operating the equipment (system) is completely adequate? Does it reflect what you have to do? Does it cover all contingencies? What is missing? What is included that is unnecessary? How could it be improved?

2. Are all required TO's, handbooks, etc., available to you? Are they complete; i.e., do they cover everything you need to know about the equipment? What was missing from these? Was any unnecessary material included? Is the material understandable? How could these publications be improved?

3. Have you had occasion to refer to technical manuals since you began the OST? On what occasion? To find out what?

Comment

Actual operating procedures may differ somewhat from those that were developed during design. These questions are asked to elicit any required procedural changes. Technical manuals should reflect the needs of the operator and, like operating procedures, may have to be brought up to date.

COMMUNICATIONS

1. Did you have any difficulty in receiving or supplying information to other personnel over internal communications equipment? What were the causes of this difficulty? How can these be changed?

2. Did the necessity for communicating interfere in any way with your job of operating the equipment? To what extent?

3. Did you have any difficulty in providing required information? Why?

MAINTENANCE

General

The following questions are asked only once (at the conclusion of the OST) and refer to the interviewee's total experience in performing maintenance:

1. Have you had any difficulties or problems in performing preventive maintenance (cleaning, oiling, adjusting, etc.) on the equipment? Did anyone else on your team have these difficulties? What were these difficulties? What caused them? How significant were these difficulties? How could these problems be eliminated?

2. Have you had any difficulties or problems in performing corrective maintenance during the test operation? What were these difficulties? Did anyone else on your team have the same difficulties? What caused these difficulties? How significant were they? How could they be eliminated?

3. How often has it been necessary for you to perform corrective maintenance during the OST? What piece of equipment failed most frequently? What impact did this have on test operations?

4. Was there anything about the equipment, procedures, or the tools you used that might make it difficult for persons less skilled than yourself to troubleshoot your equipment? Which equipment? What procedure? Which tools? What caused the difficulty?

Specific

The following questions are asked following each test operation:

1. Did you have an equipment malfunction during the test? (If the answer is no, the following questions need not be asked.)

2. How did you first become aware (by what displays or other symptoms) that a malfunction had occurred? What were the symptoms?

3. When the malfunction occurred did you have enough information to know what caused it?

4. Did you try to troubleshoot the equipment (bring it back on line) during the test? If not, why?

(The following questions need be asked only if the interviewee has attempted to troubleshoot the malfunctioning equipment.)

5. How easy or difficult was the malfunction cause to diagnose?

6. Did you refer to your technical manual? Was it of any value? How useful do you find your TM generally in troubleshooting?

7. Did you have any difficulty in securing access to the inside of the equipment? In unfastening panels? Removing the equipment chassis?

8. Approximately (to the nearest minute) how long did it take you to determine the cause of the failure? To remove and replace a component? To check that the equipment was working again? Do you consider this time excessively long? Average? Short?

9. Did you have enough room to move around the outside of the equipment while you were troubleshooting it? Within the equipment?

10. Did you have all the proper tools to perform the maintenance? Any missing? Which ones were missing? Which ones were inappropriate?

11. Were there enough test points to check out the equipment? Did you have any difficulty finding them? Are they located close to the units they check? Are they accessible?

12. Did you have a spare component to replace the failed one? Did you have to go elsewhere to find a spare?

13. Were there any safety hazards in troubleshooting the equipment? What were these? What caused them? What could be done to eliminate them?

14. Were there any difficulties in removing the failed component because of weight, shape, location?

15. Was there any difficulty installing the replacement unit?

16. Did you ask anyone else's advice while troubleshooting? Did you work as part of a team in repairing the failure?

17. Was the failure successfully cleared up? If not, what happened then?

Comment

The intent of these questions is to secure as much information as possible about maintenance, and particularly about troubleshooting equipment that has failed during the test. The questions in some respects parallel those asked about other aspects of the system (e.g., information, job aids, safety) but they must be asked again in relation to troubleshooting because their significance is different in maintenance.

SECTION SEVEN--TEST PROCEDURES

The preceding sections have dealt largely with test planning. This section describes procedures to be followed during the actual conduct of the personnel performance test.

- -

In order to secure meaningful results from the Operational System Test, the following are required:

1. Scenario Deviations

Test personnel must follow the scenario (operations plan) laid out for them. If the test objective is to determine the time for a tank to drive 1 mile, ford a 2-foot deep creek, and then drive 5 miles, this scenario must be followed in one complete and continuous action rather than as separate parts over different periods or on different days.

2. Briefing Test Participants

To ensure that the above is performed, ascertain before the test begins that the test participants have all the equipment, operating procedures, performance aids, etc. that they need to do their job properly. Test participants should be briefed before they begin the test operation to ensure that they know what they are supposed to do. This includes not only the route to be followed (if a vehicle is to be driven), but also any information they are supposed to supply to data collectors. Ask them if they have any questions and answer these.

3. Noninterference

Once a test operation has begun, no one should interfere with the performance of that operation by (a) aborting the test or (b) providing information to test participants. The reason for this is that the evaluator is trying to replicate operational conditions and outside interference or aid will not be available to personnel in the operational situation. Any such interference and aid merely cause the test results to be nonrepresentative and nongeneralizable to the operational situation.

The only exceptions to this rule are if (a) a dangerous situation arises that could hazard personnel or the system under test or (b) equipment malfunctions and makes the continuation of the test impossible. Data collectors/observers should have the authority to call off a test, but only for the most pressing reasons. Under all other circumstances, data collectors/observers should not interfere once the test has begun. In fact, observers in the physical proximity of test participants should be as unobtrusive as possible and should provide no assistance, even when asked for it, except under the hazardous conditions referred to previously.

4. Unforeseen Occurrences

However, measurement in a field environment always involves the possibility that the unforeseen will occur. For example, equipment or instrumentation may fail or the weather may not be appropriate for a particular test. The Test Conductor should have contingency plans in the event that a change in the test plan is required. Such a contingency plan may involve rescheduling a test operation, or performing part of rather than the entire scheduled event. Observers should be made aware of these contingency plans.

5. Reasons for Deviations

If test personnel fail to follow the scenario exactly, it is the observer's responsibility to determine why this has occurred. However, he should not interfere with the deviation while it is occurring, unless it involves a hazard situation, nor should he call the test participants' attention to the fact that they are deviating from the scenario. Following the test operation, he should question test participants to determine why they deviated, because the reason may have some bearing on the adequacy of the system and the procedures developed for it.

6. Specific Test Objectives

Each test operation should be defined in terms of the specific test objective it is designed to satisfy. The Test Conductor should use a check-off sheet to record that each test has in fact been performed fully, partially, or not at all. This is particularly important when evaluating a system of any size or complexity.

7. Manual Backup to Instrumentation

If data will be collected by means of instrumentation, the Test Conductor should have a manual backup method of collecting data in the event that the instrumentation fails.

8. Practice Runs

Prior to the start of formal data collection, at least one or two practice runs following all procedures exactly as intended should be conducted to try out data collection methods. Data collectors should be debriefed following these runs to determine whether any last minute changes in test procedures and/or data collection forms are necessary. Debriefing should focus on whether the desired data can be collected efficiently and whether serious data difficulties are being encountered.

9. Observer Stations

Observers of the test operation should be stationed in such a position that they can see what is occurring without their intruding unduly upon the privacy of test participants. It must be emphasized to data collectors that they are <u>not</u> test participants.

10. Equipment Failures

Data collectors should record all instances of unscheduled events occurring during the test operation. The most important of these will probably be equipment and logistics failures and any repair activities performed. This information may be needed to explain performance results.

11. Data Collection Forms

All data collection forms should be controlled in terms of their issuance to data collectors from a central office; they should be returned to the same office. All data collection forms should, as a minimum, contain the following information:

a. Identification (e.g., name) of the test participant.

b. Name of the data collector.

c. Identification of the test operation for which data are being collected.

d. Identification of the equipment being tested (in case this is not implicit in the name of the test operation).

e. Date the test was performed.

f. Scenario number.

12. Data Quality

It may be useful for data collectors to record their judgment of the quality of the data being collected in a particular test operation when, for whatever reason, they have little confidence in those data. Such information would be useful to the Test Conductor in drawing conclusions from the data.

13. Interviews

Test participants should be interviewed following each major test operation. They may be able to supply information which would amplify and explain observers' data.

14. Data Collection Monitoring

The Test Conductor should monitor all data collection activities on a sampling basis. This is to ensure that his personnel are performing as desired and that he will secure the data he desires.

15. Start/Stop Time

The start and stop time of the test activity being monitored should be ascertained by the data collector.

16. Reliability Data

All equipment failures observed to occur by test participants should be recorded by data collectors. The following information should be collected:

a. Time the failure was observed by test participants.

b. Symptoms of the failure.

c. Any diagnostic, troubleshooting, or repair activities performed by test participants.

d. Time the equipment was restored to operating status.

e. Whether or not the test was aborted as a result of failure.

f. How serious the failure was in terms of its impact on the accuracy and precision of the test operation.

One of the major parameters in terms of whether or not the system under test will be judged effective is system reliability, as measured by the occurrence of equipment failures. It is therefore essential that all such failures be reported, no matter how trivial they may appear to be on the surface. Since most of these failures will occur or first be noted during a test operation, both test participants and data collectors should be admonished to report them.

17. Availability

Availability is another important system parameter. Essentially, availability is a measure of the extent to which the system is ready to perform when it is needed.

$$\text{Availability} = \frac{\text{Total uptime (system actually operating)}}{\text{Total time (system in usable condition)}}$$

Obviously, any failure may cause the equipment or system to "go down"; until it is restored, that equipment or system is not available for use. Not every equipment malfunction will necessarily cause the equipment or the total system to fail (e.g., a light on a console failing); but every failure must be reported, if not for the determination of availability, then for the determination of reliability.

18. Maintainability

The length of time it takes to restore an equipment to operational status (otherwise known as the mean time to repair) is one index of the maintainability of the equipment or system. Other indices of maintainability relate to the equipment characteristics that make it easy or difficult to troubleshoot the equipment (e.g., accessibility of components, availability of test points, etc.). Both should be reported.

It should be obvious that the key to these measures--reliability, availability, and maintainability--is the failure report. Hence data collectors should make every effort to report the details of such failures, preferably on special report forms designed for this purpose.

SECTION EIGHT--INTRODUCTION TO STATISTICAL METHODOLOGY

Introduction

Researchers, investigators, and policy makers are often faced with the problem of obtaining or evaluating data relevant to the solution of a specific problem. Data obtained and analyzed using proper statistical techniques are likely to yield knowledge vital to the understanding of complex problems while data improperly obtained or analyzed will frequently result in poor understanding or erroneous conclusions.

In essence, statistical methodology is concerned with planning and carrying out the collection, tabulation, and analysis of data. Statistical methodology may be subdivided into two broad areas--descriptive statistics and statistical inference. Descriptive statistics is concerned with the development and utilization of appropriate arithmetic, tabular, and graphical techniques for describing data in an orderly and meaningful way. Statistical inference describes the methodology for making statements that go beyond the data that have been observed or analyzed. This chapter provides guidelines for the use of appropriate statistical techniques.

Statement of the Problem

The application of statistical methodology should not be undertaken without a clear statement of the problem being investigated. A well-defined problem statement should include clear definitions of:

1. The population(s) or universe(s) under study. The totality of individuals or units about whom knowledge is desired must be clearly specified.

2. The aspect(s) or the population(s) of interest. The characteristic(s) of the population(s) that are being studied must be rigorously defined.

3. The purpose or goal of the research. The primary goal and its associated objectives must be clearly stated. Objective(s) may include the estimation of unknown values, the answer to a specific question about a population, a comparison between populations, or the investigation of a relationship between various aspects of a population.

Some examples of simple problem statements are as follows:

1. Example 1. A new tire has been developed but it is not clear how it will perform over rough terrain. Since performance is characterized by the tire's tread life (in miles), it is necessary to estimate the average number of miles over rough terrain that the tire will travel until it needs to be replaced.

   a. Population: All new tires of this type.
   b. Aspect: Tread life (in miles).
   c. Purpose: Estimation of average tread life (in miles) over rough terrain in order to gauge tire quality.

2. Example 2. A training manual has been developed that should enable new marine recruits to utilize a specific piece of machinery with no further instruction. It is necessary to determine whether this manual is, in fact, effective so that classroom time can be diverted to other essentials.

a. Population: All new marine recruits.
b. Aspect: Ability to use a piece of machinery after reading manual.
c. Purpose: To establish whether manual is effective so that classroom time can be used for other purposes.[1]

3. Example 3. A new reading course has been established and it is not clear whether its effect will be the same in two different areas of the country. The intent of the course is to increase the reading comprehension of poor students.

a. Population: I--All poor students in area I; II--All poor students in area II.
b. Aspect: Improvement in reading score after exposure to a specific course.
c. Purpose: To determine whether the new reading course will improve reading comprehension by the same amount in two distinct areas.

4. Example 4. A final exam is given to all individuals who enroll in a given computer programming course. It is desired to use a qualifying test to determine whether an individual should be admitted to the course, but the relationship between the qualifying test and course performance is not clear. Therefore, an investigation of the relationship between final exam grade and qualifying exam grade is initiated.

a. Population: All individuals who might enroll in computer course.
b. Aspect(s): Qualifying exam score and final exam score.
c. Purpose: To determine whether the screening test is useful; that is, to assess the relationship between qualifying exam score and final exam score.

Variables

Once a statistical problem has been clearly defined, it is necessary to obtain and utilize data pertinent to its resolution. Data analyzed for statistical analysis are usually obtained from either multiple physical or mental measurements or responses to a questionnaire. It is essential that the measurements analyzed be obtained under conditions relevant to the problem being addressed. For example, if tread life of tires over rough terrain is of interest, tires should be tested under conditions that are analogous to the type of terrain about which inferences are to be drawn.

---

[1]The word "effective" was not defined precisely. In order to utilize statistical methodology, the criteria determining "effective" must be clearly stated. As an example, an "effective" manual might be one such that at least 85 percent of all recruits would properly utilize the machinery after reading the manual.

A later section of this chapter will discuss sampling techniques for obtaining data. It should be obvious that data should not be analyzed unless they are obtained in a manner such that they represent the population(s) of interest.

Recalling that statistical techniques are utilized to analyze various aspects of populations, it is necessary at this point to introduce the concept of a statistical variable.

A variable may be defined as a characteristic of the population that may differentiate individuals or units within that population. For example, the variable of interest in Example 1 above was the tread life (in miles) of a given tire. Example 2 was concerned with the variable "ability to utilize a specific piece of machinery"; and Example 3, the variable "reading score after exposure to a course." Example 4 investigated two variables: "final exam grade" and "qualifying exam grade." The variables under investigation must be clearly defined prior to the utilization of statistical techniques.

A number of different kinds of variables arise in practice. The selection of the appropriate statistical methodology is dependent upon the type of variable being analyzed. Broadly speaking, there exist two types of variables, quantitative and qualitative.

A quantitative variable is one that is recorded as a numerical value. The variables "tread life in miles," "reading score," "numerical exam grade," "blood pressure," "family income," "number of heads occurring in eight tosses of a coin," etc. are all quantitative variables since they are measured as numbers.

A qualitative variable is one that is not measured in quantitative units. Qualitative variables are defined by specifying a set of two or more categories into which individual population elements may be assigned. We "measure" or "observe" individuals with regard to qualitative variables by assigning each one to a category. Categories should be defined in such a way so that every individual or unit in the population of interest can be classified as a member of one, and only one, of these categories. This is frequently referred to as establishing a set of categories that are exhaustive and mutually exclusive. In Example 2, the variable "ability to utilize a specific piece of machinery" may be considered as a qualitative variable if individuals are rated as either "can" or "cannot." Examples of other qualitative variables include "state of origin," "eye color," "opinion towards a candidate" (will vote for, won't vote for, undecided), etc.

In actuality, there are gradations between quantitative and qualitative variables. One commonly occurring "gray area" is one in which an observation consists of response to an ordered or ranked scale (say, extremely dislike, dislike somewhat, indifferent, like somewhat, extremely like). For a discussion of the theory of measurement, see Ellis (1966) and Churchman and Ratoosh (1959).

Descriptive Statistics

Once data have been obtained, it is frequently necessary to organize and summarize them in a manner so that their meaning can be clearly understood. Descriptive statistics is concerned with the description of data without attempting to draw inferences beyond the individuals or elements from whom the measurements were taken or observed.

Data are usually summarized using tables, graphs, and summary statistics. The necessity for summarizing data is clear, since a mere presentation of observations or measurements (e.g., CAN, CAN, CAN, CANNOT, CAN, CANNOT, CANNOT, etc.) is often confusing and virtually useless. Such presentations provide data but little information about the problem.

Description of Qualitative Variables

Frequency Distributions and Graphs. Data pertaining to qualitative measures of a population are often presented in tables known as frequency distributions. A frequency distribution may be defined as a listing of all possible categories in which the variable values may occur and the number (or percentage) of individuals or units so designated. Referring to Example 2, a frequency distribution of the ability of a group of 200 recruits to use a piece of machinery may appear as shown in Table 8-1.

Table 8-1

Ability to Use New Machinery

Ability to Use Machine	Number	Percent
Could Use Machine	120	60.0
Could Not Use Machine	80	40.0
Total	200	100.0

It is essential that the total number of individuals or units tabulated be specified so as to facilitate proper evaluation of the data. Table 8-1 clearly indicates that the total number of individuals presented is 200.

Qualitative variables are frequently illustrated by means of circle and/or bar graphs. Circle graphs are especially useful when the relative proportion of individuals falling into each category is of interest; and bar graphs, when the absolute number of individuals falling into each category is of interest. Proper procedure for constructing graphical representations of data may be found in Hamburg (1970, Chapter 3).

When more than one qualitative variable is observed on elements in a population, a contingency table is often a convenient method of simultaneously summarizing such data. Consider the following example. Suppose 400 individuals were exposed to one of two teaching methods--200 were assigned to Method I and 200 to Method II. Assume that, at the end of the course, each individual was given an exam that was graded on a pass-fail basis. In this case, observations consist of (1) two measurements per individual, (2) method of instruction, and (3) exam grade. Individual observations might thus consist of pairs, such as Method I, PASS; Method II, PASS; Method I, PASS, etc. A contingency table summarizing these results might appear as illustrated in Table 8-2. Such tables can be constructed with more than two variables.

Table 8-2

Exam Scores for 400 Recruits Exposed to Two Different Teaching Methods

Teaching Method	Exam Grade		
	PASS	FAIL	Total
I	160	40	200
II	80	120	200
Total	240	160	400

Note that all categories of one variable comprise the rows of the table; and all categories of the second variable, the columns. The number in a particular cell of the table, therefore, represents the number of individuals having both a specific teaching method and exam grade. Data presented in this fashion are especially useful when one wishes to analyze the relationship between two variables. A discussion of contingency tables may be found in Neter and Wasserman (1973, Chapter 26).

Summary Statistics. When considering qualitative variables, commonly used summary statistics include the mode and category rankings.

The mode of a frequency distribution of a qualitative variable is defined as the category in which the maximum number of individuals or units have fallen. The mode (or modal value) of the frequency distribution appearing in Table 8-1 is "Could Use Machine." The mode is useful if one wishes to present the specific category that best represents the data being described. The mode should not be used as a summary statistic if two or more categories contain approximately the same number or percentage of individuals. To illustrate, the data in Table 8-3 are bimodal in nature since there are two distinct maximum categories--"extremely favorable" and "extremely unfavorable." Presenting one of the two categories as representative of this data set would be misleading.

Table 8-3

Attitude Towards Tax-Relief Bill
(Based upon a survey of 500 individuals)

Attitude	Percentage
Extremely favorable	30.0
Moderately favorable	10.0
Neutral or Undecided	15.0
Moderately Unfavorable	15.2
Extremely Unfavorable	29.8
Total	100.0

For the individuals represented in Table 8-3, the rank-ordering of the responses (e.g., 1-extremely favorable, 2-extremely unfavorable, 3-moderately unfavorable, etc.) may provide a valuable summary of the data for many applications. A good discussion of descriptions of qualitative variables is found in McCarthy (1957, Chapter 3).

Additionally, when one observes two or more qualitative variables, it is often of interest to measure the association, or relationship, between them. Considering the data of Table 8-2, we might be interested in measuring the relationship between teaching method and grade on exam. (If there is no relationship between two variables, they are known as "independent.") For a discussion of measures of association for qualitative variables, see McCarthy (1957, Chapter 11).

Description of Quantitative Variables

As in the case of qualitative variables, quantitative variables may be summarized in terms of tables, graphs, and summary statistics.

A frequency distribution of a quantitative variable may be defined as a listing of all possible values of the variable and the number (or percentage) of individuals or units within each value. In many practical situations, however, it is not feasible to list all possible values of a variable simply because that number is too large or infinite. For example, tread life (in miles) is a variable whose possible values are limited only by the accuracy of the measuring instrument (e.g., to the nearest mile, tenth of a mile, hundredth of mile, etc.). Similarly, all possible scores on a test (Example 2) may consist of the values 0, 1, 2, . . ., 100. (Note that in the case of a variable such as the number of heads appearing on eight flips of a coin, the values 0, 1, 2, 3, 4, 5, 6, 7, 8 constitute all possible values and may easily be listed.) As such, it is usually necessary to group values of quantitative variables (usually as intervals) when one summarizes values of a quantitative variable. The "best" methods of grouping values will not be discussed here other than to note that the number of intervals and specific interval values should contain all possible values of the variable and should clarify, rather than obscure, the underlying data. For a discussion of how to construct meaningful intervals, see Yamane (1964, Chapter 2). Tables 8-4 and 8-5 present frequency distributions for variables discussed in Examples 1 and 3.

Table 8-4

Number of Miles Driven Before Tire Failure
(Based upon tests of 200 tires)

Number of Miles of Tread Life	Number of Tires
More than 40,000 miles	10
35,000-39,999 miles	20
30,000-34,999 miles	30
25,000-29,999 miles	40
20,000-24,999 miles	40
15,000-19,999 miles	30
10,000-14,999 miles	20
Less than 10,000 miles	10
Total	200

Table 8-5

Reading Scores of 100 Individuals in Area I

Reading Score	Percentage of Individuals
90-100	10
89-99	20
70-79	50
60-69	5
0-59	15
Total	100

Such data are frequently illustrated by graphs known as histograms and frequency polygons. For a discussion of these graphical techniques see Yamane (1964, Chapter 2).

Note that, by further subdividing the intervals in Table 8-4, we obtain the data appearing in Table 8-6. Finer subdivisions might result in the limiting frequency distribution graphed in Figure 8-1. A possible limiting distribution for the data in Table 8-5 is presented in Figure 8-2. For a treatment of distributions of continuous variables, see McCarthy (1957, Chapter 3).

Table 8-6

Number of Miles Driven Before Tire Failure
(Based upon tests of 200 Tires)

Number of Miles of Tread Life	Number of Tires
More than 50,000 miles	3
40,000-49,999 miles	7
37,500-39,999 miles	10
35,000-37,499 miles	10
32,500-34,999 miles	15
30,000-32,499 miles	15
27,050-29,999 miles	20
25,000-27,499 miles	20
22,500-24,999 miles	20
20,000-22,499 miles	20
17,500-19,999 miles	15
15,000-17,499 miles	15
12,500-14,999 miles	10
10,000-12,499 miles	10
7,500-9,999 miles	7
Less than 7,500 miles	3
Total	200

Quite often, quantitative variables are summarized using various descriptive summary statistics. When one considers summarizing or representing a set of quantitative data, it is natural to search for measures of the "center" of the data and of the "dispersion" or "variability" of the data.

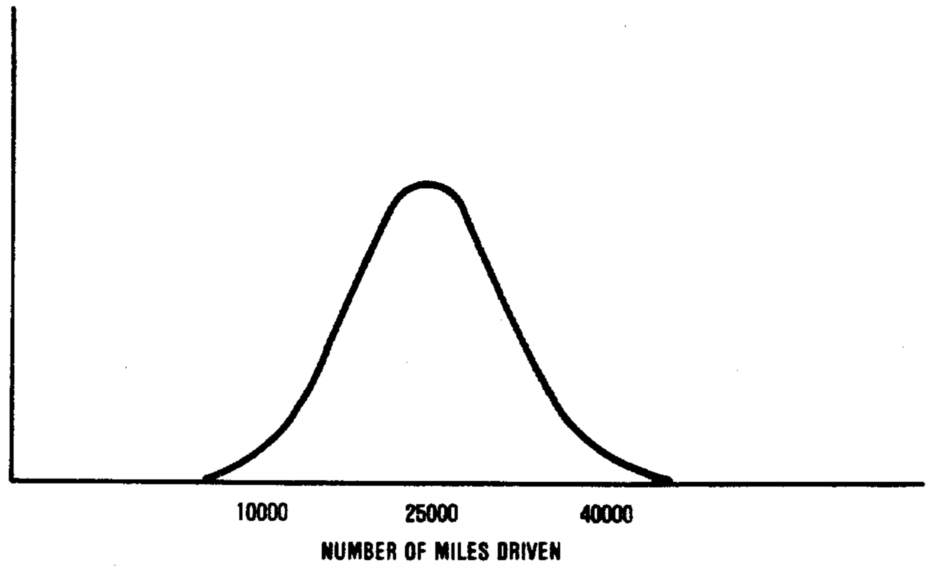


Figure 8-1. Limiting frequency distribution for number of miles driven before tire failure.

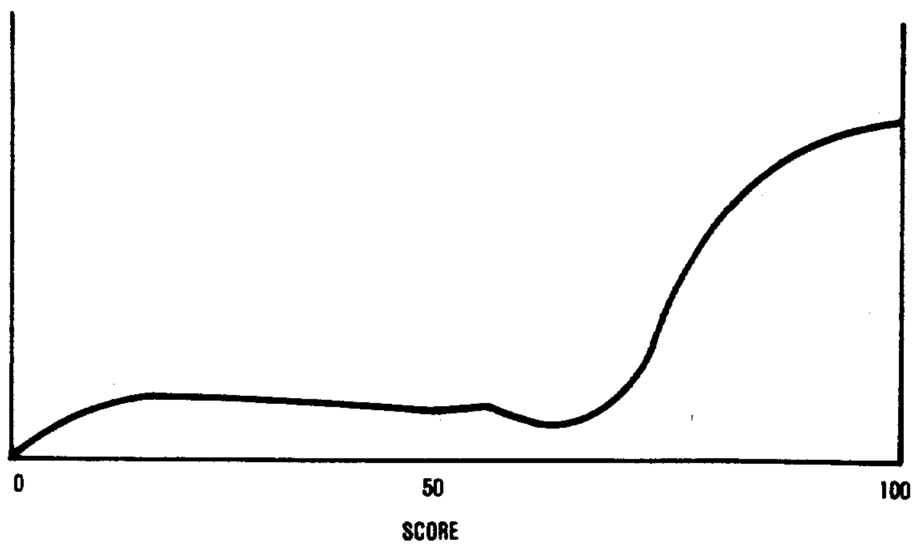


Figure 8-2. Limiting frequency distribution for reading score of individuals in Area I.

Measures of Central Tendency. A variety of measures exist for describing the "center" of a data set.

1. The Arithmetic Mean. The arithmetic mean is defined as the sum of the observations divided by the number of observations. The mean is a useful summary statistic when one wishes to include all observations, including extreme values in the summary measure. Suppose the values 7, 56, 2, 2, 8 represent the number of days during the past 2 months on which a sample of five individuals ate steak for dinner. The mean of the observations 7, 56, 2, 2, 8 is (7 + 56 + 2 + 2 + 8)/5 = 75/5 = 15 days. For a more thorough discussion of the arithmetic mean, see Yamane (1964, Chapter 3).

2. The Median. The median of a data set is the value of the middlemost observation (in the case of an even number of observations, the average of the two "middlemost" observations) when the observations are ranked in size order. The median is a useful summary statistic when it is felt that "extreme" observations distort, or are unrepresentative of, the underlying data. For the values listed above (i.e., 7, 56, 2, 2, 8), the median is 7 days since, after ranking these observations (e.g., 2, 2, 7, 8, 56), 7 appears as the middlemost observation. Note that the extreme value, 56, did not enter into the calculation of the median. For a discussion of the median, see Yamane (1964, Chapter 3).

3. The Mode. The mode is obtained in a manner analogous to that of qualitative variables. For a discussion of the mode, see Yamane (1964, Chapter 3).

Many other measures of central tendency exist, such as the harmonic mean, geometric mean, etc. See Yamane (1964, Chapter 3) for a discussion of these and other measures.

Measures of Variability. Although measures of central tendency summarize data in terms of their center, these statistics are in no way descriptive of the dispersion of the data. For example, although Figure 8-3 illustrates two sets of data with similar "centers," the data sets are quite different. If the data sets represent the diameters of a precision tool manufactured by two different machines, the graphs indicate that one machine produces tools of almost uniform quality while the second produces tools with considerably higher variability. Some commonly used measures of variability are discussed below.

1. The Range. The range of a data set is the difference between its highest and lowest values. For the values 7, 56, 2, 2, 8, the range is 56 - 2 = 54. The range is very sensitive to extreme values, and, like the mean, should not be used when extreme values are felt to be unrepresentative of the process or population under study.

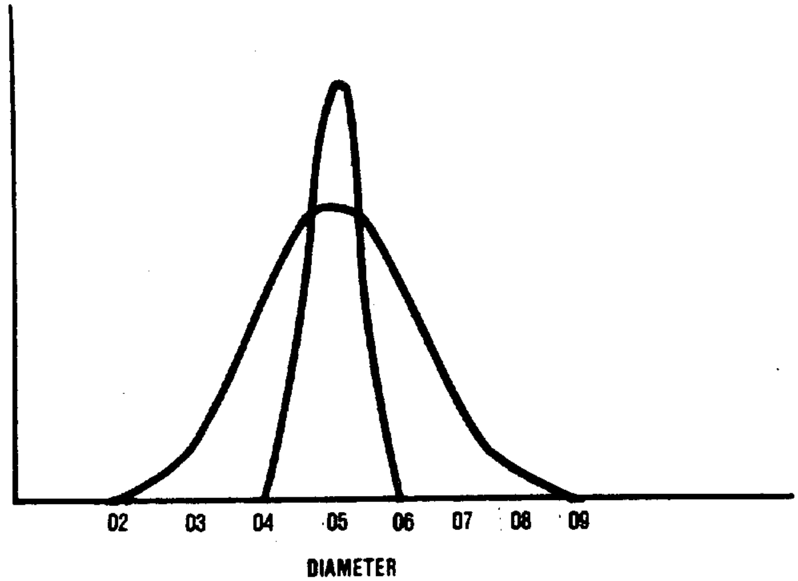


Figure 8-3. Two frequency distributions with similar means but different dispersion.

2. The Variance and Standard Deviation. The variance of a data set is the average of the squared deviations of each observation from the arithmetic mean. The standard deviation is the square root of the variance. Both of these measures of dispersion are especially useful when one wishes to make inferences beyond the observed data or to obtain a useful measure of sampling precision. For the values 7, 56, 2, 2, 8, the variance is $(7-15)^2 + (56-15)^2 + (2-15)^2 + (2-15)^2 + (8-15)^2/4 = (8)^2 + (41)^2 + (13)^2 + (13)^2 + (7)^2/4 = 2132/4 = 533$. The standard deviation is $\sqrt{533} = 23.1$. (Note: The denominator is usually taken as one less than the number of observations.)

Many other measures of dispersion exist, including mean absolute deviation, semi-interquartile range, etc. For a discussion of measures of dispersion, see Yamane (1964, Chapter 4).

Measures of Association

When more than one quantitative variable is measured, one frequently wishes to describe the type and strength and direction of relationship between those variables. To illustrate, Figure 8-4 presents a plot known as a scatter diagram that illustrates the observed data in Example 4. Note that the relationship between variables appears to be positive and linear. Frequently, product-moment correlation coefficients are utilized as measures of association between quantitative variables. For discussions of correlation coefficients and illustrations of their use, see Walker and Lev (1953, Chapters 10 and 11).

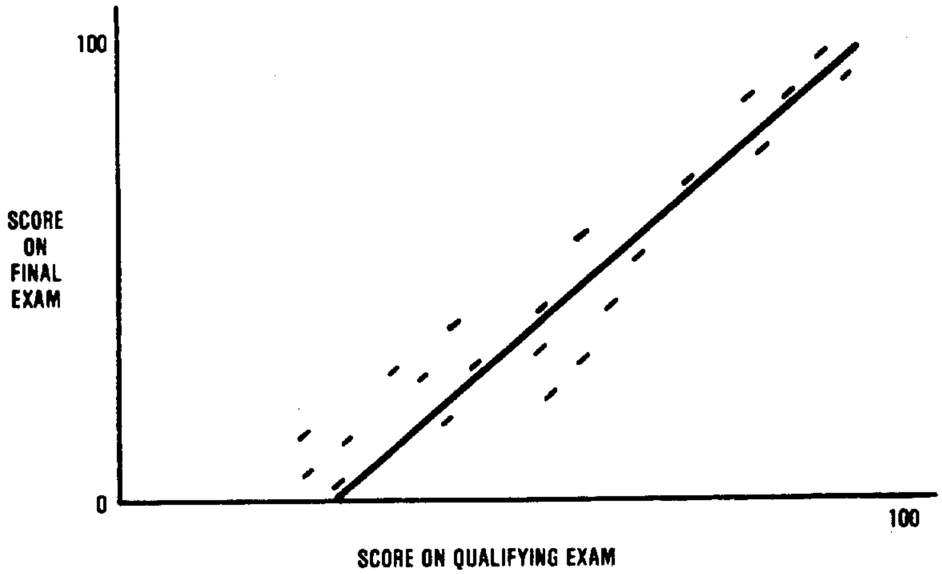


Figure 8-4. Scatter diagram of score on qualifying exam and score on final exam (Example 4).

Other Measures. There exist many other summary statistics for quantitative data. Measures of "skewness" describe the symmetry of a set of observations, while measures of kurtosis consider the "peakedness" of a frequency distribution.

This section was designed merely to introduce the topic of descriptive statistics. It is essential that references such as those cited previously be consulted before attempting to properly summarize any data.

Very often we desire to make statements, estimates, or decisions that go beyond the specific data that have been observed or analyzed. The next section deals with this subject of statistical inference.

Statistical Inference

It is often necessary to generalize findings to a larger domain than the individuals or units actually observed. If we define our population or universe to be that larger domain, and if we define the observations being analyzed to be the sample, we then wish to use the statistics and other information obtained from the sample to make statements about the characteristics of the entire population. This generalization from observed data to population of interest is called statistical inference. Specific areas of statistical inference include:

1. Estimation of Unknown Population Parameters (Note: A parameter is a summary measure of the units of a population whereas a statistic is a summary measure of the units of a sample).

2. Tests of Hypotheses.

3. Analysis of the Relationships Between Variables.

4. Forecasting and Decision Making.

5. Time-Series and Trend Analysis.

The above list is by no means exhaustive; rather, it is intended to provide the reader with some knowledge of the problems with which inferential methodology is concerned. It should be noted that general conclusions derived from sets of observations are necessarily uncertain. Statistical methodology provides techniques for both assessing the accuracy of our estimates and for judging the probability of making incorrect decisions within this climate of uncertainty.

Estimation

Statistical estimation is concerned with the problem of estimating one or more population parameters from the information contained in a sample. One problem that frequently arises in practice is that of estimating the mean of a variable in a population. Another common problem considers estimation of the proportion of individuals or units in a population with some predefined characteristic or membership in a specific category. To illustrate, Example 1 relates to the problem of estimating the average tread life over rough terrain for all tires of a specific type. The technique for analyzing this problem is typical of many statistical analyses; that is, a sample is drawn and summary statistics (obviously including the mean of the sample in this case) are computed from the observations in the sample. The value of the summary statistic (the mean tread life of tires in the sample) is then used as an estimate of the mean tread life of all tires produced. Considering the situation in Example 2, we may wish to estimate the proportion of all Marine recruits (not merely those in the sample) who could use the machine properly after reading the training manual. In this case, the statistic "proportion in the sample who used the machine properly after reading the training manual" is often used as the estimate.

The mere presentation of an estimate, however, avoids the issue of its accuracy or precision or error. Statistical estimation methodology provides techniques for constructing estimates that are, in general, as accurate as can be achieved from the sampling methods employed. Techniques have been developed for estimating the accuracy or error of statistical estimates. These techniques are based upon consideration of the sampling distribution of statistics; that is, the study of the frequency distribution of repeated sampling and estimation from a given population.

Based upon the Central Limit Theorem of statistics, statistical techniques based upon the normal distribution have been developed for judging the precision of estimates of population means and proportions when the sample size is sufficiently large (in most applications, at least 30 to

100) and drawn in a specific manner. Often an estimate and its associated precision are utilized to construct a confidence interval—an interval in which we are reasonably certain that our unknown population parameter is located. For example, rather than state that our estimate of the average tread life of tires is 23,570 miles--we might present our findings as "we are 99 percent 'confident' that the average tread life is between 23,260 and 23,880 miles." For a discussion of these and other estimation concepts, see Dixon and Massey (1969, Chapters 5-7).

Another estimation problem concerns not simply the estimation of the value of an unknown population parameter but, rather, the estimation of differences between a given parameter of two or more populations (i.e., the difference between the average tread life of tires produced by manufacturer A and manufacturer B). For a discussion of this and similar problems, see Dixon and Massey (Chapter 8).

Hypothesis Testing

Hypothesis testing is an aspect of statistical methodology concerned with determining whether an unknown population parameter is equal to a prespecified value (or class of values). Considering Example 2, we might be interested in determining whether the proportion of all Marine recruits who cannot utilize the machine properly after reading the training manual is 5 percent or less. Considering Example 1, we might truly be concerned not with estimating average tread life, but, rather, simply judging whether average tread life is at least 35,000 miles.

Ordinarily, hypothesis testing problems are denoted by specifying both a null hypothesis, or statement, and an alternative hypothesis about an unknown population parameter. Considering our tread life example, our null hypothesis might be "average tread life of all tires is 35,000 miles or more," while the alternative hypothesis might be "average tread life of all tires is less than 35,000 miles." Statistical techniques have been developed for use in deciding which of these hypotheses is correct. These methods enable the user to develop test procedures with definable probabilities of making incorrect decisions.

When one constructs a test of hypothesis, the following two decision errors are possible: (1) Type I Error: Concluding that the alternative hypothesis is true when, in fact, the null hypothesis is true; and (2) Type II Error: Concluding that the null hypothesis is true when, in fact, the alternative hypothesis is true.

The utilization of proper test construction methodologies in conjunction with appropriate sampling techniques allows the user to analyze the probability of making either of these errors. Carrying out a test of hypothesis is called a test of significance. For a discussion of significance testing, see Dixon and Massey (1969, Chapters 6-8).

Multivariate Analysis

Multivariate statistical techniques consider the analysis of several variables at once. This type of statistical analysis is performed when one wishes to assess the relationship between several variables. Frequently, this type of analysis is utilized when one wishes to search for "causes" (although an analysis of data itself is not sufficient to attribute causality) or for predictive or forecasting purposes. Examples of multivariate techniques include:

1. Correlation Analysis is frequently employed when quantitative variables are analyzed. For example, if our observations consist of pairs of quantitative observations as in Example 4 (i.e., 1st Exam Score, Final Exam Score), correlation analysis may be the appropriate methodology to analyze the relationship between these two variables.

2. Chi-Square Techniques are frequently used to estimate and test hypotheses about the relationships between qualitative or categorical variables.

3. Analysis of Variance (ANOVA) Techniques are often employed when one wishes to analyze the relationship between a quantitative variable and one or more qualitative or categorical variables. An ANOVA might be employed when one wishes to analyze the relationship between test-score (quantitative variable) and region (qualitative variable) and teaching methodology (qualitative variable).

4. Regression Techniques are utilized when one wishes to derive and analyze the relationship between one or more "predictor" or "independent" variable and a dependent variable. For example, one may wish to predict an individual's final exam grade (dependent variable) on the basis of his qualifying exam score (independent variable).

The multivariate techniques discussed thus far are but a handful of the wide variety available. It is critical that no techniques be utilized without a thorough understanding of the assumptions underlying the use of each method. See Snedecor and Cochran (1973); Yamane (1964); Dixon and Massey (1969); or Freund and Williams (1972) for discussion of some multivariate methods.

Time Series

Time series techniques deal with the analysis of the behavior of variables over time. The assessment of trends, cycles, and seasonal fluctuations are some of the questions addressed by time series methodology. See Yamane (1964, Chapters 12 and 13) and Neter and Wasserman (1973, Chapters 29-32) for discussions of time series methodology.

The concepts and methodologies presented in this section are merely an introduction to the type of problems analyzed through statistical inference.

Sampling Techniques

Thus far, we have not discussed techniques for obtaining the data needed for analysis. It is obvious that we wish to obtain and analyze data that are "representative" of the population of interest. Random sampling techniques have been developed that enable the user to draw samples likely to be representative of the population of interest. These techniques include:

1. Simple Random Sampling. Simple random sampling techniques draw individuals into the sample in a manner whereby every individual or the unit in the population has an equal chance of being selected. Furthermore, individuals are chosen independently. That is, the selection of an individual into the sample has no impact on the selection of any other individual or item into the sample. Therefore, if we wish to select a sample of all Marines using simple random sampling methodology, then we must select individuals so that every marine has an equal (and independent) chance of selection. The utilization of this and other random sampling techniques requires the user to have a frame, or listing, of the individuals or items in the population of interest.

2. Stratified Random Sampling. Stratified random sampling is a sampling technique in which the population is first divided into stratum, or subpopulations, and then a simple random sample is drawn from each stratum. For example, if we divide Marines into stratum based upon geographic location and then sample randomly from each of these stratum, such a scheme would constitute a type of stratified random sampling. Such sampling frequently results in estimates having increased precision. However, stratified sampling is often difficult to carry out in practice.

3. Cluster Random Sampling. Cluster random sampling is a sampling methodology whereby "groups" or "clusters" of individuals or items are selected as part of the sample at once rather than individually. For example, when drawing a sample from the population of all naval personnel serving on ships, we might draw a sample of ships (clusters) and consider all personnel on the chosen ships as members of the sample. Cluster sampling techniques are often the easiest and most inexpensive procedures to carry out. The results of these schemes, however, also tend to be the most difficult to analyze.

Other sampling techniques such as quota sampling, systematic sampling, and combinations of the above schemes are also employed.

It is crucial that data be obtained in accordance with accepted sampling techniques if one wishes to measure the accuracy of estimates, to analyze the probability of making errors, or to make generalizations of findings from sample to population. Furthermore, the specific sampling and estimation (or hypothesis testing) methodology employed enable the user to gauge the size of the sample (number of individuals or units) needed to obtain a desired precision. For discussions of sampling, see McCarthy (1957, Chapter 10) and Kish (1967). For discussions of a related subject, the design of experiments, see Snedecor and Cochran (1973).

Summary

This chapter has introduced some of the basic concepts of statistics—including descriptive techniques, statistical inference, and sampling methodology. Investigation of references such as those described in this chapter is essential prior to the utilization of any statistical technique.

REFERENCES

Churchman, C., & Ratoosh, P. Measurement-definitions and theories. New York: John Wiley & Sons, Inc. 1959.

Dixon, W. J., & Massey, F. J. Introduction to statistical analysis (3rd ed.). New York: McGraw-Hill, Inc., 1969.

Ellis, B. Basic concepts of measurement. Cambridge, ENG: Cambridge University Press, 1966.

Freund, J., & Williams, F. Elementary business statistics (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

Hamburg, M. Statistical analysis for decision making. New York: Harcourt, Brace, and World, Inc., 1970.

Kish, L. Survey sampling. New York: John Wiley & Sons, Inc., 1967.

McCarthy, P. J. Introduction to statistical reasoning. New York: McGraw-Hill, Inc., 1957.

Neter, J., Wasserman, W., & Whitmore, G. A. Fundamental statistics for business and economics (4th ed.). Boston: Allyn and Bacon, Inc., 1973.

Snedecor, G., & Cochran, W. Statistical methods (6th ed.). Ames, IA: Iowa State University Press, 1973.

Walker, H., & Lev, J. Statistical inference. New York: Henry Holt and Company, Inc., 1953.

Yamane, T. Statistics: An introductory analysis. New York: Harper and Row, Inc., 1964.

SECTION NINE--PERSONNEL PERFORMANCE TEST PLANNER'S CHECKLIST

This section provides a checklist that can be used by test planners to ensure that they have performed all necessary steps to conduct an effective test. The checklist covers necessary questions to be answered for pretest planning, pretest operations, test period, and posttest period. In some cases, comments are provided on the right-hand side of the following pages.

A. PRETEST PLANNING

A1. <u>General Organization of the Personnel Performance Test Plan (PPTP)</u>

a. When must it be completed?

b. Will it be developed in several stages or all at once?

c. Will it be included as one part of the overall Operational System Test (OST) plan or will it be a separate document?

d. How long and detailed must it be?

e. Who must review and approve it?

f. What information must the PPTP contain? (See Section Two)

A2. <u>Equipment/System Background Information</u>

a. Is a description of the equipment/system to be tested available?

(1) Sent along with the test plan requirement?

(2) Are other sources of equipment description specified?

(3) Is the description sufficiently detailed?

b. Is a description of personnel tasks required to operate/maintain the equipment available?

(1) Included with equipment description?

(2) Are other sources of personnel information specified?

(3) Is the description sufficiently detailed?

c. Any other relevant documentation available?

(1) Specifications, e.g., MIL-STD 1472B?

(2) Previous test plans, reports?

A3. Test Purpose

These purposes include:

a. Assess whether system development personnel-related objectives have been accomplished.

b. Ensure that personnel can accomplish required tasks satisfactorily.

c. Forecast how personnel will perform in actual operations.

d. Ensure that personnel elements are effectively integrated into the system.

e. Detect personnel-related deficiencies (e.g., inadequate equipment and procedures, training, technical data, supplies) so that improvements can be made.

f. Identify personnel resources (e.g., manpower) needed to support the operational system.

g. Validate the training program.

Needless to say, not all these purposes may apply to the same OST.

a. Which of these purposes applies to this test? What are the implications of these purposes?

If the purpose is to determine whether system development personnel-related objectives have been accomplished, it is necessary to know what those objectives were--in particular, the criteria for these objectives. The same is true of determining that personnel can do their jobs effectively: how does one know without quantitative criteria? To validate personnel training, again one must know what the training program was designed to accomplish: this means training criteria. To determine what personnel-related deficiencies exist, it is necessary to specify in measuring performance what errors consist of and how inadequate performance is defined.

b. What will be measured by personnel performance tests?

(1) Human engineering.

Have human engineering specifications (MIL-STD 1472B or any other) been levied against the system? Has previous (developmental) testing measured human engineering?

(2) Adequacy of operating procedures.

What does "adequacy" in this context mean? Are procedures up to date? Have they been tried out previously? How do they relate to Marine Corps objectives and operations?

(3) Maintenance operations.

Is maintenance being covered by any other section of the OST test organization? If so, it may not be part of the personnel performance test responsibility unless what the other section is doing fails to involve personnel elements. Is the test concerned with all aspects of maintenance or troubleshooting alone? Has a malfunction reporting procedure already been established? Which of the following maintenance procedures have been established for the system: (a) attempt to repair malfunctions during the test; (b) abort the test and return system to depot; (c) do not attempt to repair but proceed with the test; and (d) working around the malfunction. What information should be collected about maintenance (e.g., total down time, repair time, types of errors made)? Examine in detail implications of collecting personnel data in relation to maintenance; they may give the planner a headache.

(4) Correctness of technical manuals.

This aspect is not quite the same as evaluating the adequacy of operating procedures. Determining correctness of any technical manuals involves a great deal of very detailed work.

(5) Adequacy of training.

Training adequacy can be determined in various ways: Types of personnel errors may indicate lack of training; test personnel can be asked directly whether they feel their training was adequate and if not, in what ways it was deficient.

(6) Ability of test personnel to perform tasks.

Comments with regard to performance criteria apply especially to this aspect.

(7) Effect of special operating conditions on personnel performance.

For example, are personnel required to drive tanks at night as well as during the day, in swamps as well as on hard surfaces, etc.? Examine mission objectives to determine if very contrasting operating conditions exist for which it would be useful to collect data on personnel capability to perform under these conditions.

(8) Other (e.g., logistics).

A4a. <u>Relation of PPTP to System Operations</u>

(1) What operations will be performed as part of OST?

For various reasons (e.g., cost/time), not all the operations in which the system under test is ordinarily utilized may in fact be tested.

(2) Will personnel data be gathered on all OST operations?

If the system has many operations and personnel, the requirement to collect personnel data on all of these may impose a severe burden on data collectors.

(3) If not, on which ones?

If only selected system operations will be used to collect personnel data, it will be the test planner's responsibility to select these (if not already specified) on the basis of: (a) criticality to mission accomplishment; (b) frequency of performance (more frequently performed operations are, all other things being equal, more important to evaluate); and (c) difficulty of operation, if known (more difficult tasks will stress personnel more).

(4) Are there any special conditions in these operations that would impact on test planning?

(5) Will all tasks performed in specified OST operations be measured?

Each system operation to be measured may require a number of tasks to be performed. Some of these are more or less important, more or less easy to gather data on. The test planner must specify which of these tasks (if not all) must be measured; test observers need this information.

A4b. Measurement of Personnel Performance

(1) What criteria for successful performance of the tasks being measured exist? (List these.)

(2) Are they quantitative and in sufficient detail?

(3) If performance criteria for these tasks are not specified in available documentation, what possible other sources exist?

(4) Can criteria be developed by consensus of experienced personnel? (See Section Two)

A4c. Measurement Methods

(1) Will instrumentation be required? Compare advantages/disadvantages.

In general, unless the data desired can be secured in no other way, instrumentation is not a preferred measurement method because of cost, scheduling problems, the need for specialized equipment operators (and maintainers), and difficulty in using such equipment in a field setting.

(a) If so, is it available?

(b) Must it be procured? From where? What are the procurement procedures?

(c) Cost/schedule.

(2) Observation.

(a) What information will observers record?

(b) What data recording forms will be required?

(3) Interviews of test personnel.

(a) What information will be secured from interviews?

(b) What questions should be asked?

(4) Questionnaires.

(a) What information will be secured from questionnaires?

(b) What questions should be asked?

(5) Ratings.

(a) What information will be collected from ratings?

(b) Who will fill out the rating scales?

(c) Will rating scales have to be developed or are adequate ones available?

A4d. Test Personnel Required

Test personnel are those personnel who operate and maintain the equipment being evaluated.

(1) How many?

(2) When must they be available?

(3) Is special background required? If so, what?

(4) Rank/skill level?

It is not sufficient merely to specify that test personnel will have a given rank and military speciality. Since the personnel will vary in terms of their ability, the planner should ask whether they should come from the top 10 percent in ability, the middle (50%) in aptitude, or even lesser skilled personnel. Obviously, if test personnel are the "cream of the crop" of their speciality, they do not properly represent the great mass of military personnel; however, system performance in the test will be more efficient, since such personnel can more adequately compensate by their skill for any deficiencies the system may have. On the other hand, lower skilled personnel (e.g., the middle 50% in ability) will be more representative of the military population who will eventually have to use the system; but the system in their hands will not look "as good" as if it were operated by more effective personnel. The choice is a matter of philosophy: making the system look its best; or getting results that apply more directly to the overall Marine Corps population.

(5) Secured from what units?

(6) Will test personnel require training on equipment?

(a) If so, will training be given at the factory, by USMC, where, and for how long?

A4e. Test Observers Required.

(1) How many?

(2) What type? Rank/speciality area?

(3) What will their duties be?

It is particularly important to specify in precise detail the activities required of observers. If this is not done, the data recovered may be inaccurate or some may even be missed.

(4) What training will they be required to have? (Describe training.)

It is highly desirable that observers be given realistic training in their observational duties; if this is not done, the data they secure may be inaccurate or some may even be missed.

(5) Who will provide this training, where will it be given, and when?

(6) From where will observers be secured?

(7) To whom will they report?

It is presumed that test observers will report to the individual in charge of personnel performance testing but this should be specified.

A5. Testing Schedule

a. Will personnel performance data be gathered as an integral part of the overall OST?

Ordinarily this is or should be the situation, in which case the overall OST schedule determines the personnel performance test schedule.

b. Will special personnel performance tests be required?

Ordinarily such special-purpose personnel tests should not be necessary if personnel performance testing is fully integrated into all OST phases. However, it is conceivable that special questions relative to personnel may arise that cannot be satisfied in the normal course of OST. The test planner should examine OST operations to be performed before answering this question.

c. If so, what are these and how will they be conducted?

d. What will be the impact of such special tests on the overall test schedule?

A6. Data Analysis

a. What data will be secured?

Before testing begins (even before training of test observers begins), a detailed list of the data items to be collected, along with information on how and for what purpose they are to be collected should be drawn up so that everyone involved knows exactly what is needed. Statistical analysis cannot be specified before this list is developed.

b. What statistics will be applied?

The plan for statistical analysis of the data is an integral part of the PPTP.

c. How will the data be processed?

d. Who will perform the data analysis?

If the statistical analysis is to be performed by other than the usual USMC agencies, the performing agency should be identified. This includes any personnel who will handle/process the data between the test observer who collects it and the analyzing agency.

e. Are computer facilities necessary? Who will provide them?

A7. Final Test Report

a. What sections will the test report include?

USMC regulations may specify a particular format for the personnel performance test report, whether it is separate or included as part of the overall OST report. Whatever the case, personnel performance test report should have at least the following sections: purpose of test, methods used (including instrumentation (if used), data recording forms, questions asked, etc.), subjects, procedure for collecting data, results, conclusions, recommendations.

b. What is the schedule for the test report?

c. Who will write the test report?

B. PRETEST OPERATIONS

This refers to the period between the time the PPTP is written and the start of actual testing. It includes all the preparations (including observer training) for conducting personnel performance testing.

B1. Availability of Test Personnel

a. Have they arrived?

b. Have they received required training (including checking out on equipment)?

c. Do they have all required job aids (if job aids are necessary to task performance)?

d. Have they been instructed on the role they will play in OST?

All test personnel should be informed that, as a routine part of OST, their performance will be measured and that they will be interviewed, observed, and/or asked to fill out certain forms. They should be reassured that this evaluation is solely to check out the equipment.

B2. Availablity of Test Observers

a. Are they on-site?

b. Have they received required training as observers?

B3. Availability of Measurement Devices

a. If instrumentation is required, has it been received and checked out and are observers trained in its use?

b. Are all manual recording forms ready and have they been tried out as part of observer training?

c. Are all interview/questionnaire questions developed and tried out in observer training?

d. Is the test schedule up to date?

This will often change up to the start of testing because of delays in getting equipment ready.

C. TEST PERIOD

C1. Initial Checks on First Day's Weeks's) Results

It is highly desirable to check results of the start of testing because various problems often arise in testing which must be resolved. If these are not solved, much data collected may be inadequate or even lost.

a. Check with test observers:

(1) Any difficulties in collecting data experienced by observers or in performing by test personnel?

Observers can usually report on whether test personnel are experiencing difficulties that might interfere with data collection.

(2) Are any changes to measurement procedures, instrumentation, recording forms, or test schedule required?

(3) What is the effect of such changes on the PPTP?

(4) Will the desired number of data points be secured?

Results of the first week of testing should indicate whether it is possible to collect all the data specified in the PPTP. Changes, if any, in the overall OST operation (e.g., breakdown of equipment, reshuffling of personnel) may interfere with personnel performance data collection and may require corresponding changes in the data collection procedure.

(5) Are the desired data being secured?

(6) Do observers appear to know their jobs?

It is desirable to check on how well observers are performing because those who appear to be falling down on the job may have to be reindoctrinated or replaced.

b. Check with test management:

(1) Is the test on schedule?

(2) Are any changes anticipated in test operations that will impact on the PPTP?

C2. <u>Periodic Check During Test Operations Concerning Above Questions</u>

Periodic checks on data collection should be made because at any time the OST operation may be modified (because of equipment malfunction, scheduling delays) such that personnel performance data collection may have to be curtailed or otherwise modified.

C3. Final Check

a. Have all necessary data been secured and recorded?

If some necessary data are missing at the conclusion of OST, the personnel performance test planner will have to decide what can be done about this.

b. Have all data been transmitted to data analysts?

D. POSTTEST PERIOD

D1. Data Analysis

a. What is the schedule for data analysis? Is the analysis on schedule?

b. Are the data appropriate to the planned statistical analysis?

If, for various reasons, it is found that the data collected will not fit the planned statistical analysis, important decisions about changing the analysis format must be made.

c. Were sufficient data collected to satisfy test objectives?

d. Are the results relevant to the test objectives?

Inadequate planning may result in insufficient or irrelevant data being collected. If so, critical decisions must be made.

D2. Final Test Report

a. What is the schedule for the preliminary draft? For the final test report? Is the report writing on schedule?

b. Are the results/conclusions clearcut?

c. What recommenations can be made? Are they reasonable? What will their impact on the system be?

d. What system modifications are required:

(1) In hardware?
(2) In procedures?
(3) In training?
(4) In manning?

SECTION TEN--THE PERSONNEL PERFORMANCE TEST REPORT

This section outlines the major points to be included in a personnel performance test report.

Introduction

This section describes the major points to be included in a report describing the results of the personnel performance test. The Marine Corps has its own report format, as described in MCO 5000.11, Test and Evaluation of Systems and Equipment for Operating Forces of the Marine Corps; and the personnel performance test report described in this Section should conform to that directive. Within the constraints of the Marine Corps test report format, the items described herein should be included.

The personnel performance test report is a major vehicle for the transmission of information about the test and will reach a wide variety of interested agencies. It is important, therefore, that care be taken in its preparation.

The major categories which the test report should cover are:

1. Summary
2. Test Objectives
3. Test Method
4. Results
5. Conclusions
6. Recommendations
7. References
8. Appendices

Outline of the Personnel Performance Test Report

A. Summary of Test Report

A paragraph or two describing the highlights of the study with emphasis on:

1. Purpose of the personnel performance measurement.
2. When and where test was conducted.
3. Major results and conclusions.

B. Personnel Performance Test Objectives

1. This section should describe the objectives for which the personnel performance test was conducted. Specifically these objectives should have been to answer the following questions:

a. The determination of how well personnel perform with the new system.

b. The determination of whether personnel satisfy system requirements as far as their performance is concerned.

c. The problems that personnel experience as these reflect on various aspects of the system, e.g.:

(1) Human engineering of equipment design.
(2) Operating/maintenance procedures.
(3) Manning.
(4) Appropriate personnel background to perform duties.
(5) Training.
(6) Other (e.g., logistics, manuals, job aids).

C. <u>Test Method</u>

1. Test Personnel

This section descibes the characteristics of personnel acting as test subjects.

a. Definition of test personnel as those operating and maintaining the system during test exercises.

b. Selection of test personnel:

(1) Personnel selected from what units.
(2) Personnel background (e.g., military speciality, rank).

Indicate any personnel characteristics particularly important to the system (e.g., strength, aptitude).

(3) Number of subjects.
(4) Selection criteria.

Indicate the basis for determining how many personnel were selected as test subjects and the rationale for the selection criteria (e.g., the 95th percentile of scores in school training, ranking by commanding officer of their unit, selection on a random basis). If personnel were selected by tests or scores, what were these? Were there any constraints on personnel selection (e.g., small population) and what were these?

c. Special training received by test personnel (to operate/maintain test system).

2. Test Procedure

This section describes general test methodology and performance criteria.

a. Test was conducted over what time period? Using what facilities? As part of operational exercises or in the form of special tests? How was test conducted?

b. Tasks/operations for which personnel performance data were collected. List and, if reader is unlikely to be familiar with these, describe major functions/tasks performed for which data were collected. If not all tasks/operations were observed/measured, what was basis for selection? Indicate number of operating cycles (e.g., tank runs, rounds fired) on which data were collected.

c. Experimental Design

If a specific experimental design was used (e.g., repeated measures on the same subjects, special order of performing tasks such as alternating day/night exercises), describe at this point and indicate rationale for the design.

d. Specific variables tested (e.g., day vs. night operations, sandy vs. marshy terrain). Reason for being concerned about these variables.

e. Personnel performance criteria:

(1) For all major operational tasks performed, what quantitative criteria describe adequate personnel performance (e.g., allowable firing miss distance (2 feet); maximum time allowed for replacing X component (38 minutes))?

(2) Indicate source of criteria:

(a) Overall system requirements.
(b) System documentation (reference).
(c) Operational requirements determined by mission.
(d) Consensus of skilled experts.

(3) List any objective performance measures collected and categorize these by the criteria in section C2e(1). Define each measure employed (e.g., what is meant by error, response time, etc.?).

(4) Indicate any difficulties or problems in measuring these criteria. If so, what was done to resolve these problems?

f. Data collection methods:

(1) If observation was used, indicate:

(a) Who made the observations.
(b) How the observations were made.
(c) How observers were qualified to make these (e.g., training, experience).
(d) What observers were supposed to observe in relation to what system operations.
(e) If any observational data recording forms were used, place these in the appendix.

(2) If interviews were held with test personnel, describe:

(a) The general content of the interviews.
(b) When and where held.
(c) Who was interviewed (not in terms of specific names but in terms of categories of personnel).
(d) Average length of interview.
(e) Whether taped or manually recorded.

(3) If questionnaires were used, include the form used in appendix, and describe:

(a) The general content of the questionnaire.
(b) As with interviews, when or how frequently the questionnaire was employed.
(c) Who completed questionnaires.

(4) If rating scales were used, include them in appendix, and describe:

(a) The nature of the scales.
(b) The data they were supposed to produce.
(c) Who completed scales and how frequently.

(5) If instrumentation was used to collect objective measures, describe:

(a) The general nature of the instruments (e.g., time and events recorder, noise level measurement device).
(b) The particular measures it was used for.

Note. If the instrument is novel, it might be advisable to append a more detailed specification of its operating characteristics, including a photograph.

(6) If the experimental design of the study (see section C2c) involved a comparison of two or more conditions (e.g., performance under different climatic or terrain conditions), include a description of these special conditions. Any special conditions that were important to the test should be described in detail.

D. Results

1. Statistical Analysis

Referring back to the experimental design (section C2c) as the rationale for the analysis:

a. Describe the analyses performed (e.g., Analysis of Variance, *t*-tests, correlations).

b. Indicate the adequacy of the data collected, particularly any factors that might have affected the analysis, such as too little data, non-normal distribution, etc.

Note. If the manner in which the overall OST was performed influenced the quantity/quality of the data, indicate what this was.

2. Personnel Performance Effectiveness

a. Objective Measures.

This section refers back to section B1a and describes how well personnel have performed. It should include data gathered by instrumentation or by observation of quantitative indices (e.g., miss distance in firing at targets). It should include both operator and maintainer functions unless maintenance will be covered in a separate report or report section.

(1) Determine the statistical mean (average) and standard deviation performance in terms of specific measures for each major function/task as previously called out in section C2b. Compare the mean with any system-required personnel performance (the standard of accomplishment). Examine the variability (standard deviation) of the performance: Is the variability so great that the mean value is unreliable?

(2) Determine the statistical significance of differences between any conditions being compared (section C2d).

(3) List the performance values for each major function/task in tabular form. If these data are extensive, they should be included in a separate appendix.

(4) Where appropriate, categorize types of errors made by personnel and indicate their frequency.

Note: The statistical section of the report should be written by a qualified statistician or at least reviewed by him.

b. Subjective Data

Any subjective data (i.e., those gathered from observations, interviews, questionnaires, ratings, or critical incidents) that bear on how well personnel have performed or which explain their performance should be included here. Subjective data which can be described in quantitative terms (e.g., mean and standard deviation of ratings, the percentage of those responding yes and no to particular questions in interviews and questionnaires, the number of those observed to perform in particular ways or the frequency of their performance) should be listed in tabular form, where possible.

3. Equipment Characteristics

This section describes any human engineering equipment discrepancies that have been noted by test observers or by test personnel in interviews, questionnaires, or rating scales.

a. List each discrepancy per equipment and refer to appropriate section of MIL-STD 1472B (Department of Defense, 1974) for which it is a discrepancy. For example, "the noise level within the tank compartment is excessive, measuring peaks of 90db (see paragraph 5.8.3.2 of MIL-STD 1472B)."

b. Indicate the importance of the discrepancy in terms of its effect on test personnel and/or mission accomplishment, using a scale such as (1) minor--1, (2) moderately important--2, (3) extremely important--3. Indicate actual or possible effects on performance from the test data.

c. Where appropriate, include diagrams, photos, etc., illustrating the discrepancy (e.g., diagram of improperly laid out control panel).

4. Operating and/or Maintenance Procedures and/or Manuals

List any inaccuracies or changes required in procedures or manuals that were found as a result of test performance.

5. Training

The purposes of this section are to describe the adequacy of the training given test personnel to operate/maintain the new system and to indicate where further training is required. The training curriculum provided test personnel (see section C1c) should be examined in terms of how well personnel performed and how they felt about their training. Data will be derived from a number of sources:

a. Functions/tasks with inordinately high error rates or very delayed response times, where the cause of such errors appear to result from inadequate training.

b. Data secured from interviews and questionnaires in which questions were asked specifically about training (e.g., were there any functions/tasks for which not enough training was given or the training appeared to be inappropriate?).

6. Personnel Requirements

This section includes any deficiencies noted in:

a. Manning--the number of men required to operate/maintain the system (for example, if two men are specified but three are required or vice versa).

b. Special aptitudes noted that are required to perform system functions.

E. Conclusions

1. General

This section describes the answers to objectives in section B1. System personnel can or cannot operate/maintain the system to requirements. Manning is or is not appropriate for required tasks. Training is or is not adequate, etc.

2. Specific

Inadequacies were found in:

a. The following tasks (list and describe).
b. Human engineering (describe).

c. Personnel requirements (describe).
d. Training (describe).

These have the following effects on system operations (describe).

F. Recommendations

1. Changes to the system should be made with regard to:

a. Equipment design.
b. Procedures.
c. Personnel requirements.
d. Training.
e. Other.

2. Indicate which of the above modifications can be made by:

a. Equipment redesign.
b. Changes to procedures.
c. Training of personnel.
d. Logistics (e.g., spares, tools, etc.)

G. References

1. Military documents cited.
2. Civilian publications cited.

H. Appendices

1. Tabular data (e.g., statistical analyses, lists of errors made, important raw data).

2. Photos/diagrams of important items of equipment referred to in the body of the report.

3. Data collection forms, interview questions, etc.

SECTION ELEVEN--USEFUL REFERENCES

This section presents additional specifications and reference materials which the evaluator may find useful.

Specifications and Standards

MIL-H-46855A Human Engineering Requirements for Military Systems, Equipment and Facilities, 2 May 1972.

MIL-STD-1472B Human Engineering Design Criteria for Military Systems, Equipment and Facilities, 31 December 1974

MIL-STD-721B Definitions of Effectiveness Terms for Reliability, Maintainability, Human Factors, and Safety, 25 August 1966.

Reference Books

McCormick, E. J. Human factors engineering (3rd edition). New York: McGraw-Hill, 1970.

Meister, D., & Rabideau, G. F. Human factors evaluation in system development. New York: Wiley, 1965.

Van Cott, H. P., & Kinkade, R. G. Human engineering guide to equipment design. Washington, D. C.: U. S. Government Printing Office, 1972.

Woodson, W. E., & Conover, D. W. Human engineering guide for equipment designers (2nd edition). Berkeley, CA: University of California Press, 1966.

DISTRIBUTION LIST

Chief of Naval Operations (OP-987H), (OP-964D)
Chief of Naval Personnel (Pers-10c), (Pers-2B)
Chief of Naval Material (NMAT 04), (NMAT 08), (NMAT 08T244)
Director of Navy Laboratories
Commandant of the Marine Corps (5)
Commanding Officer, Naval Development and Training Center (Code 0120)
Director, Training Analysis and Evaluation Group (TAEG)
Army Research Institute for the Behavioral and Social Sciences
Military Assistant for Training and Personnel Technology, Office of the Under Secretary of Defense for Research and Engineering
Coast Guard Headquarters (G-P-1/62)