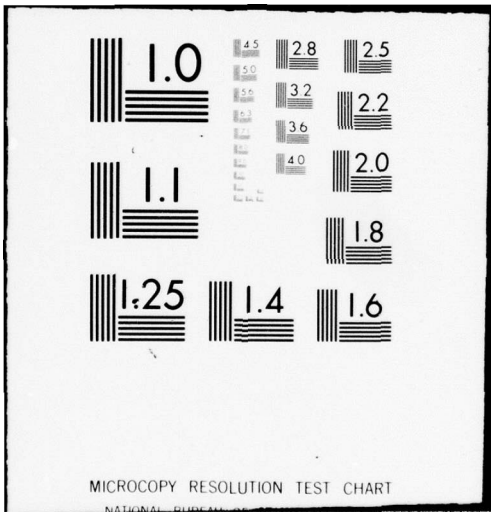
AD-A072 465 AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX F/G 5/9
THREE SETS OF TASK FACTOR BENCHMARK SCALES FOR TRAINING PRIORIT--ETC(U)
MAY 79 D C THOMSON, K GOODY

UNCLASSIFIED AFHRL-TR-79-8 NL

| OF |
ADA
072465

END
DATE
FILMED
9-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART

AFHRL-TR-79-8

(2) LEVEL II

AIR FORCE HUMAN RESOURCES LABORATORY

THREE SETS OF TASK FACTOR BENCHMARK SCALES FOR TRAINING PRIORITY ANALYSIS

By

David C. Thomson, Sq Ldr, USAF/RAAF Exchange
Kenneth Goody, Sq Ldr, USAF/RAAF Exchange

OCCUPATION AND MANPOWER RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

May 1979
Final Report for Period July 1976 – December 1978

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Occupation and Manpower Research Division, under project 7734, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Sq Ldr David C. Thomson (ORA) was the Principal Investigator for the Laboratory.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

RAYMOND E. CHRISTAL, Technical Director
Occupation and Manpower Research Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFHRL-TR-79-8	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)* THREE SETS OF TASK FACTOR BENCHMARK SCALES FOR TRAINING PRIORITY ANALYSIS	5. TYPE OF REPORT & PERIOD COVERED Final July 1976 – December 1978
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) David C. Thomson Kenneth Goody	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Occupation and Manpower Research Division Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62703F 77340705
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235	12. REPORT DATE May 1979 13. NUMBER OF PAGES 46
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

benchmark scales
consequences of inadequate performance
task delay tolerance
task difficulty
task factor ratings
task factors
training course design
training priority
training priority index

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The Occupation and Manpower Research Division of the Air Force Human Resources Laboratory has established that task training priority is a function of task factors and that, within a specialty, training priority ratings can be duplicated mathematically from task factor ratings. Because these ratings are measured on a different scale for each specialty, this methodology cannot be generally established and applied. To overcome this limitation, a series of benchmark scales was developed for the measurement of task factors against common frames of reference. This report presents the development of the benchmark scales for specialties with an electronic and a mechanical aptitude requirement. It then reports on the validation of the three sets of benchmark scales (Administrative/General, Electronic, and Mechanical). The benchmark scales were shown to have advantages and disadvantages compared with the traditional relative scales.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

404 415

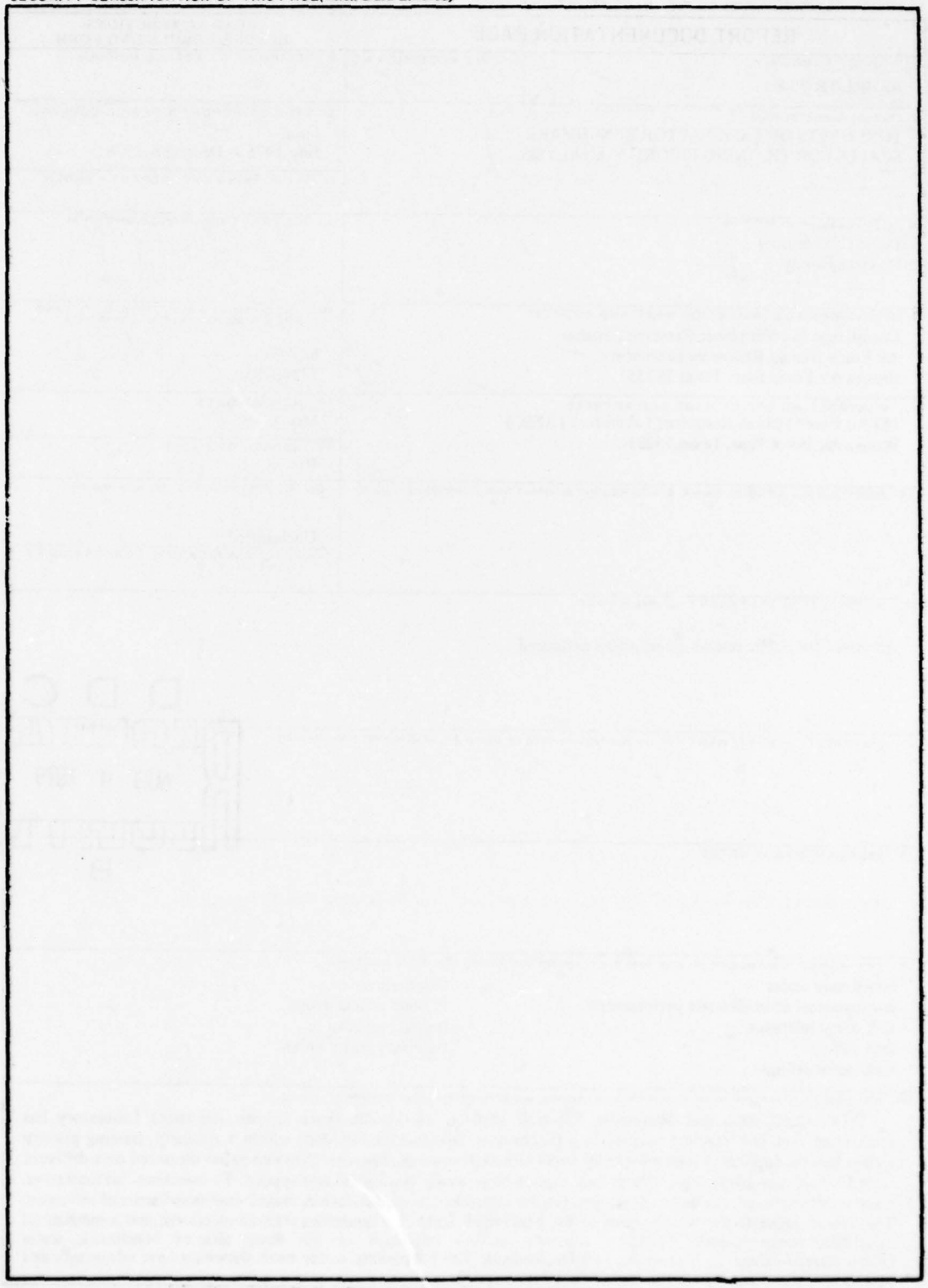


TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

THREE SETS OF TASK FACTOR BENCHMARK SCALES FOR TRAINING PRIORITY ANALYSIS

I. INTRODUCTION

The Occupation and Manpower Research Division of the Air Force Human Resources Laboratory (AFHRL) is engaged in research on the development of procedures for determining task training priorities (Christal, 1970; Mead, 1976). One element of this research is the development of benchmark scales for measuring task factors that contribute to training priority decisions. The type of benchmark scale employed is a 9-point scale on which each level is illustrated by three typical tasks that belong at that level. Scales have been developed for three task factors: Probable Consequences of Inadequate Performance, Task Delay Tolerance, and Task Difficulty. In all, three series of scales have been developed: one for specialties with an Administrative or a General (A/G) aptitude requirement, a second for specialties with an Electronic (E) aptitude requirement, and a third associated with a Mechanical (M) aptitude requirement. This report will discuss the method used in developing the benchmark scales, describe the initial validation and final rater agreement studies, and provide all three series of the benchmark scales along with recommended instructions for their use.

II. BACKGROUND

The basic concept of the present task training priority research was conceived and reported by Christal (1970). Three papers presented at the 17th Annual Conference of the Military Testing Association (MTA) documented achievements up to 1976. The first of these papers (Christal & Weissmuller, 1975), now also available as a technical report (Christal & Weissmuller, 1976), described eight new programs that were introduced into the Comprehensive Occupational Data Analysis Programs (CODAP) system to enable investigators to manipulate and analyze task factor data. The second MTA paper (Mead, 1975) reported the results of a training priority study that demonstrated the feasibility of mathematically duplicating training priority ratings in terms of a number of task factors. The third MTA paper (Goody & Watson, 1975) introduced the benchmark scales as a means to permit measurement of task factors against common frames of reference for various specialties. It was envisaged that a limited number of regression equations using benchmark scale task factor data could be computed, each applying across a number of specialties and predicting task training priorities.

Goody's (1976a) technical report gave a comprehensive overview of the training priority research effort at AFHRL and reported the initial work and techniques used in developing the set of A/G task factor benchmark scales. During 1977 and 1978, the final A/G, E, and M benchmark scale validations were completed and the results are the subject of this report. The use of all nine scales in predicting training emphasis will be the subject of a separate report.

III. THE DEVELOPMENT OF THE SCALES

The first step in developing each series of scales was to gather rating data for a reasonably large inventory of tasks, hereafter referred to as the aptitude area task inventory. After these data were refined and analyzed, they were used to select the 27 tasks for each of the benchmark scales. The following paragraphs elaborate on the aptitude area task inventories developed, the rating process and data analysis, and finally, the procedure used for selecting the tasks for the benchmark scales.

The aptitude area task inventory for each series contained tasks typically performed in a variety of specialties with the appropriate aptitude requirements, and each task statement was labelled with the Air Force Specialty (AFS) that normally performs the task. The tasks selected were performed by some airmen at the journeyman level in the AFS nominated. An attempt was made to avoid tasks so specialized that only a person from the AFS who performs the tasks could understand them. Finally, the tasks selected for the general inventory had to be ones which exhibited wide variability of response. The important characteristics of the three aptitude area task inventories are displayed in Table 1.

Table 1. **Aptitude Area Task Inventories Distribution of Tasks and Specialties by Aptitude Area**

Aptitude Area	Number of Tasks	Number of Specialties Represented
Administrative/General (A/G)	438	52
Electronic (E)	392	50
Mechanical (M)	323	42

Each aptitude area task inventory was then rated on each of the existing three task factor 9-point relative scales (Probable Consequences of Inadequate Performance, Task Delay Tolerance, and Task Difficulty). The raters were selected randomly from the first-line supervision level (7-skill level) of all the represented specialties within the appropriate aptitude area. Different raters were used for each factor to avoid induced correlation. Potential confusion was of special concern in rating the Task Delay Tolerance factor which employed an inverted scale compared to the other two factor scales; i.e., level 1 was the most demanding level on the scale. About 120 ratings (N) were sought on each factor, such large numbers being necessary to provide confidence in the stability of the means obtained by having raters rate tasks from other specialties. First-line supervisors were selected to provide an optimum blend of general experience and first-hand knowledge of journeyman level tasks.

The first step in the data analysis was to identify and delete "divergent" raters, using the techniques reported by Goody (1976b). A divergent rater is one whose ratings are substantially different from those of the rest of the group. This is usually caused by the rater not adhering to the rating instructions, either deliberately or through lack of understanding, or by the rater employing a different rating policy from the majority of raters.

Having refined the data set, the degree of interrater agreement between the raters was measured using the intraclass correlation technique reported by Lindquist (1953). Because the tasks were rated relative to each other rather than on an absolute scale, the adjustment option, discussed by Christal and Weissmuller (1976), was used to convert each rater's scores to a common mean of 5.0 and a standard deviation of 1.0. The adjusted interrater agreements for single raters (R_{11}) and for rater groups (R_{kk}) are reported in Table 2. Because high interrater agreement coefficients were obtained, the task means were stable enough to use as a basis for selecting tasks for the benchmark scales.

The data for each factor for each aptitude area were then analyzed separately to select the tasks for the corresponding benchmark scale. The purpose of this analysis was to select three representative tasks to illustrate, by example, each level on the scale. Each of the three tasks in each set had about the same mean ratings on the factor involved, but their mean ratings were appreciably different from those of the tasks chosen to represent the levels on either side.

The basic tool used to select the tasks for the scale was a CODAP printout containing the task statements for all the tasks in the inventory, and the mean and standard deviation of all

Table 2. **Interrater Agreements (Adjusted) for Ratings on the Aptitude Area Task Inventories**

Task Factor and Aptitude Area	N	K	R_{11}	R_{kk}
Consequences of Inadequate Performance				
Administrative/General	116	115.38	.507	.992
Electronic	115	112.15	.472	.990
Mechanical	125	124.25	.474	.991
Task Delay Tolerance				
Administrative/General	120	118.27	.442	.989
Electronic	124	122.35	.366	.986
Mechanical	116	115.59	.370	.985
Task Difficulty				
Administrative/General	117	116.33	.478	.991
Electronic	128	125.02	.486	.992
Mechanical	121	117.50	.475	.991

Note. — N = number of raters in the sample.
K = average number of raters per task for the whole inventory.

ratings on each task for the task factor involved. The tasks were arranged in descending order of task mean rating. Four lines were drawn across the upper half of the printout–the first one-fourth standard deviation above the unweighted mean of the task means (AVEMN); the next, one-half standard deviation above that; the third, one-half standard deviation higher again; and the fourth, still another one-half standard deviation higher. A corresponding set of four lines was drawn across the lower half of the printout–the first, one-fourth standard deviation below the AVEMN; and so on. The inventory tasks were thus divided into nine groups; the seven in the middle each spanned one-half standard deviation of task mean ratings, and the other two comprised the upper and lower tails. These nine groups correspond to the nine levels of the task factor for which the scale was being developed. The three tasks for level 5 were then selected from around the middle of the middle group (around the AVEMN), and those for each of the other levels from their corresponding group. In each case, the tasks selected were near each other and in the half of their group farthest from the AVEMN. This ensured maximum, yet fairly uniform, separation between levels, and homogeneity within levels. Tasks with high standard deviations were avoided since they indicated low rater agreement on those tasks.

As an example, a portion of the mechanical aptitude area task inventory ordered on descending Task Difficulty mean value, is shown in Figure 1. Tasks marked with an asterisk were selected for the M Benchmark Task Difficulty Scale as they have low individual standard deviations, provide a wide coverage of the various specialties, and would appear to be meaningful to all airmen of the mechanical aptitude AFSs. A flow chart, depicting the development phases of the three task factor mechanical scales, is displayed in Figure 2.

The results of this initial phase of the development process was a list of 27 task statements for each factor, each set purporting to define one of nine graduated levels of that factor. An example of one of the nine scales so developed with its instructions is presented in Appendix A. The scales permit the rating of tasks relative to the tasks in the benchmark scale rather than relative to other tasks in the inventory that is being rated.

Task Number	Task Statement	Task Mean	Task SD	Group Width	Group/Level
					Groups 8/9 (29 tasks)
33[a]	Install thrust-reversing system on jet engines (Engine M)	6.15	.575		
37[a]	Operate a/c inflight refuelling system (Flight Engineer)	6.10	.683	1/2SD	Group 7 (41 tasks)
43[a]	Draw sketches or plans of parts to be machined (M Mech.)	6.07	.819		
44	Conduct spectrometric analysis	6.06	.739		
71	Refuel a/c with engines running or Ground power connected	5.73	.835		
72	Repair ignition systems on recip. engines	5.71	.619		
73[a]	Mix caustic solutions (Cryogenic Fluid Spec.)	5.71	.774	1/2SD	Group 6 (70 tasks)
78[a]	Identify and splice priority circuits (Cable Splic. Spec.)	5.70	.631		
79[a]	Remove, replace, adjust components on bomb doors	5.70	.534		
161	Inspect metal surfaces for cracks using fluorescent pene.	5.09	.686		
162[a]	Bleed, adjust, service a/c brake systems (A/C Mainten. Spec.)	5.07	.550		
169[a]	Test aviation fuel for water (Fuel Spec.)	5.01	.685	1/2SD	Group 5 (59 tasks)
177[a]	Erect poles using power equip. (Out. Wire and Antenna Main.)	4.95	.621		
178	Perform preacceptance inspections of a/c loads etc.	4.95	.593		
244	Secure a/c for severe weather conditions	4.31	.642		
245[a]	Operate a/c cargo loading equip. (Aircargo Spec.)	4.31	.554		
249[a]	Install ice cream refrigerators (Refrig. and Air condit. Spec.)	4.27	.740	1/2SD	Group 4 (54 tasks)
252[a]	Center brake shoes etc. (General Purpose Vehicle Repair)	4.25	.584		
254	Load aerospace ground equip on a/c, trucks or trailers	4.24	.663		
279[a]	Lash or tie cargo on truck, trailer (Vehicle Oper/Dispat.)	3.98	.585		
282	Operate motorized hoists	3.98	.619		
283[a]	Erect or use scaffolds/ladders (Protect. Coating Spec.)	3.97	.655	1/2SD	Group 3 (38 tasks)
290[a]	Refill mobile fuel units (Fuel Specialist)	3.83	.739		
291	Remove safety locking devices from a/c before flight	3.76	.968		
	[a]Tasks selected for the M Task Difficulty Benchmark Scale.				Groups 2/1 (32 tasks)

Figure 1. **Mechanical Aptitude Area Task Inventory Task Difficulty Means.**

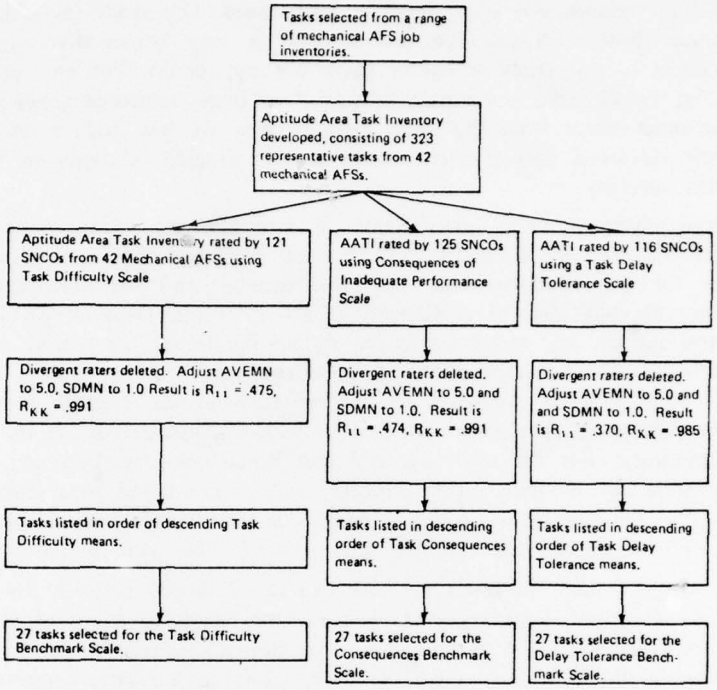


Figure 2. **Phases in the development of the three mechanical benchmark factor scales.**

IV. INITIAL VALIDATION STUDY

As already noted, the aptitude area inventories used to develop the scales contained tasks performed in a large number of specialties. For each aptitude area, 13 of these specialties were selected for the initial validation study, and the tasks associated with those specialties were extracted to form a smaller inventory of tasks (about 110) that will be referred to as the validation inventory for the corresponding aptitude area.

For each factor, a set of validation raters was selected by randomly picking first level supervisors from throughout the aptitude area. These respondents rated all tasks in the appropriate validation inventory against the corresponding benchmark scale. They will be referred to as "benchmark-general" ratings to distinguish them from the original ratings which shall be called the "relative-general" ratings. The only difference between the gathering of the benchmark and the relative ratings, apart from the type of scale used, is that the relative-general raters rated the full set of aptitude area inventory tasks, which included the interspersed subset of validation inventory tasks.

Another set of ratings was gathered during the initial validation study. Samples of first-level supervisors were selected randomly from each of the 13 specialties chosen to be presented in the validation inventory. These respondents rated only the tasks in the validation inventories that were associated with their own specialty. These ratings were against the benchmark scales and will be referred to as the "benchmark-specialist" ratings to distinguish them from the other two sets.

As in the developmental phase, the first stage of data processing was to identify and delete divergent raters. Because the benchmark-general raters rated the complete validation inventory, the standard deletion procedure was used for them. However, each benchmark-specialist rated only a

few tasks and the true variance was quite small in many cases. This made the task of identifying divergent raters much more difficult. The analyst took a very conservative approach, leaving possible divergent raters in this study whenever there was any doubt. For example, a rater who gives the same rating to all tasks is normally discarded as being noncooperative; in this case, if the variance of the other raters from the same specialty was also low, such a rater was retained. Specialist raters were discarded only if their ratings showed illogical relationships to the bulk of the ratings from their specialty.

Having removed divergent raters, coefficients of interrater agreement, using the intraclass correlation technique reported by Lindquist (1953), were computed for each task factor within each aptitude area for the relative-general, benchmark-general, and benchmark-specialist ratings. Table 3 presents the raw and adjusted coefficients of interrater agreement (R_{11}) obtained in each case. For the relative ratings, any raters eliminated during the initial phase were also omitted in computing these statistics. These relative interrater agreements are different from those reported in Table 2; those in Table 2 were computed on all the tasks in the original aptitude area task inventories while the relative rating statistics of Table 3 were computed only on the tasks included in the validation inventory. For the relative-general and benchmark-general ratings, coefficients of interrater agreement were also computed after adjusting each rater's scores to a common mean and standard deviation. For the benchmark-specialist ratings, however, adjusting the scores would be meaningless, as each of the raters rated only the few tasks from his own specialty.

Because the average number of raters per task (K) varied greatly between the three types of ratings, R_{kk} statistics have not been reported. Instead, the predicted value of R_{kk} that would result from 20 raters rating each task is presented. This allows a comparison to be made between the group reliabilities of the types of ratings. In every case the interrater agreement for general ratings, made using the benchmark scales, consistently exceeds the interrater agreements for ratings made using the relative scales under parallel circumstances. This suggests that the benchmark scales can produce reliable and stable ratings. Furthermore, with only one exception, the interrater agreement coefficients for the benchmark-specialist ratings also consistently exceed the corresponding coefficients for the benchmark-general ratings. This suggests that specialists rating only their own tasks, and using the benchmark scales, can provide stable ratings for the task factors.

Using the validation inventory task means, zero-order correlation coefficients were computed between the three types of ratings for each of the task factors for each of the aptitude areas. The results are tabulated as Table 4. With only one exception, the correlations between the relative-general ratings and benchmark-general ratings exceeds .9 for each aptitude area. Hence the use of relative and benchmark scales does result in the tasks being ranked in the same order. But the correlations between the benchmark-specialist ratings and the benchmark-general ratings, although high, are lower than the relations between the two sets of ratings provided by the general raters. Thus, while raters drawn from a variety of specialties can agree on task factor ratings on an inventory of tasks, their agreement with specialist ratings of the same set of tasks is not as high. Assuming the specialists are most familiar with the tasks in their specialties, they should be able to give the more correct rank order of those tasks for each task factor. The lower correlations are presumably the result of slight inaccuracies in the rank ordering of tasks by the benchmark-general and relative-general raters. A flow chart at Figure 3 provides an example of the major steps in the initial validation of the electronic consequences scale.

Finally, the initial validation study addressed the question of whether the benchmark scales apply across specialties. To test this, the mean ratings (AVEMN) provided by the benchmark-general raters were compared with those from the benchmark-specialist raters for the same sets of tasks for the three A/G scales. In all three cases, the specialists' mean ratings were usually displaced towards the more demanding end of the scale as demonstrated in Table 5. That is, in comparison with a general group of raters, a specialist, when using the benchmark scales to

Table 3. **Interrater Agreements for Tasks in Validation Inventories**

Task Factor and Aptitude Area	K	Raw[a]		Adjusted[a]	
		R_{11}	$R_{20,20}$	R_{11}	$R_{20,20}$
Consequences of Inadequate Performance					
Administrative/General Series (127 Tasks)					
Relative-General	115.38	.428	.937	.569	.964
Benchmark-General	29.80	.559	.962	.606	.969
Benchmark-Specialist	12.98	.462	.945	–	–
Electronic Series (110 Tasks)					
Relative-General	112.15	.248	.868	.394	.929
Benchmark-General	30.00	.401	.930	.472	.947
Benchmark-Specialist	15.11	.437	.939	–	–
Mechanical Series (99 Tasks)					
Relative-General	124.25	.245	.866	.369	.921
Benchmark-General	26.00	.345	.913	.407	.932
Benchmark-Specialist	12.80	.361	.919	–	–
Task Delay Tolerance					
Administrative/General Series (127 Tasks)					
Relative-General	118.27	.403	.931	.481	.949
Benchmark-General	24.95	.431	.938	.493	.951
Benchmark-Specialist	12.19	.435	.939	–	–
Electronic Series (110 Tasks)					
Relative-General	122.35	.207	.839	.287	.890
Benchmark-General	29.85	.286	.889	.390	.927
Benchmark-Specialist	13.64	.440	.940	–	–
Mechanical Series (99 Tasks)					
Relative-General	115.59	.215	.846	.292	.892
Benchmark-General	28.00	.293	.892	.346	.914
Benchmark-Specialist	12.67	.359	.918	–	–
Task Difficulty					
Administrative/General Series (127 Tasks)					
Relative-General	116.33	.386	.926	.485	.950
Benchmark-General	29.98	.462	.945	.541	.959
Benchmark-Specialist	13.80	.539	.959	–	–
Electronic Series (110 Tasks)					
Relative-General	125.02	.333	.909	.441	.940
Benchmark-General	34.97	.441	.940	.537	.959
Benchmark-Specialist	14.78	.579	.965	–	–
Mechanical Series (99 Tasks)					
Relative-General	117.50	.285	.889	.393	.928
Benchmark-General	29.98	.374	.923	.435	.939
Benchmark-Specialist	14.76	.408	.932	–	–

[a] $R_{20,20}$ = Project reliability of the mean rating that would result if $_K$ were equal to 20 in all cases.

Table 4. **Correlations Between the Three Types of Validation Inventory Task Means for Each Factor for Each Aptitude Area**

Task Factors and Aptitude Area	XY	XZ	YZ
Consequences of Inadequate Performance			
Administrative/General Series	.971	.909	.897
Electronic Series	.916	.843	.836
Mechanical Series	.912	.760	.767
Task Delay Tolerance			
Administrative/General Series	.944	.822	.790
Electronic Series	.875	.923	.787
Mechanical Series	.909	.632	.724
Task Difficulty			
Administrative/General Series	.961	.805	.870
Electronic Series	.954	.809	.828
Mechanical Series	.927	.809	.828

Note. — X = Relative-General Ratings.
Y = Benchmark-General Ratings.
Z = Benchmark-Specialist Ratings.

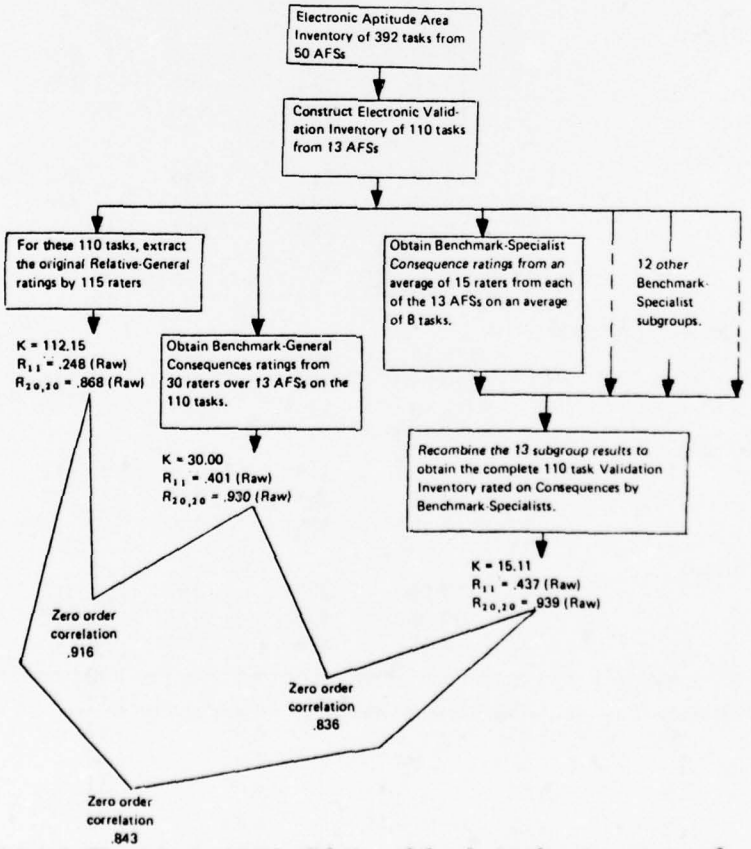


Figure 3. **Phases in the initial validation of the electronic consequences of inadequate performance benchmark scale.**

Table 5. **Inflation of Task Factor Ratings by Specialist Raters When Compared with General Raters When Using the A/G Benchmark Scales**

AFSC	Number of Tasks	Benchmark General		Benchmark Specialists		Number of Spec Raters
		SD	Mean	Mean	SD	
Consequences of Inadequate Performance						
29150	11	1.66	4.45	4.86	1.90	13
39150	6	1.61	4.96	5.32	1.42	11
43330	5	1.28	5.20	6.25	1.27	12
51150	4	1.04	3.54	3.65	1.57	12
57150	14	1.13	7.18	7.34	1.17	12
63150	7	1.17	5.31	6.12	1.33	13
64550	13	1.62	4.11	4.90	2.00	8
70250	19	1.24	2.75	3.81	1.56	12
79150	7	.87	2.86	4.11	1.99	12
90150	9	1.45	6.39	6.37	1.50	15
90250	9	2.56	5.05	5.32	2.61	12
92250	18	.52	6.97	7.52	1.08	11
99120	5	.88	3.57	4.82	1.69	10
Total/Average	127	2.05	4.95	5.56	2.06	153
Task Delay Tolerance						
29150	11	.89	4.92	5.04	1.69	12
39150	6	1.08	6.24	5.94	1.29	8
43330	5	1.21	5.17	3.80	1.39	8
51150	4	.77	6.70	5.97	1.55	10
57150	14	.92	1.86	2.13	.97	12
63150	7	1.29	4.81	4.24	1.60	16
64550	13	1.55	6.11	4.98	1.77	13
70250	19	1.28	6.75	5.35	1.55	8
79150	7	1.00	7.81	5.11	2.13	10
90150	9	1.60	3.16	2.61	1.47	12
90250	9	1.92	3.80	4.63	2.34	11
92250	18	.66	4.18	4.54	1.18	14
99120	5	.97	7.33	5.71	1.36	15
Total/Average	127	2.07	5.05	4.50	1.88	149
Task Difficulty						
29150	11	1.42	3.73	3.93	1.67	15
39150	6	1.16	5.48	5.45	1.37	16
43330	5	.87	4.76	5.57	1.24	14
51150	4	1.74	4.29	4.23	2.19	17
57150	14	1.21	4.40	5.60	1.65	13
63150	7	1.21	4.62	4.57	1.60	14
64550	13	.83	3.20	3.67	1.16	14
70250	19	.82	2.46	2.86	1.08	12
79150	7	1.40	4.39	5.04	1.90	16
90150	9	1.19	5.26	5.09	1.52	13
90250	9	2.43	4.25	4.49	2.58	8
92250	18	.64	4.74	4.98	1.08	14
99120	5	1.35	4.63	4.88	2.11	18
Total/Average	127	1.47	4.11	4.48	1.73	184

rate tasks from his own specialty, tends to indicate that the difficulty of doing the task is higher, the consequence of not performing the task satisfactorily is more serious, and the time delay permitted before the task must be done is smaller.

To test the hypothesis that the inflation was uniform across specialties, an analysis of covariance was permformed on the three A/G factors. The benchmark-general ratings were used as a control variable, and the significance of the difference in specialist ratings between specialties was tested. In all three cases, the F statistic was significant at the 5 percent level, indicating that the inflation effect is probably not uniform across specialties. Figure 4 displays this result demonstrating graphically the non-uniform nature of the "inflation" of ratings by specialists. Further analysis showed that the inflation was not related to the aptitude requirements of the 13 specialties.

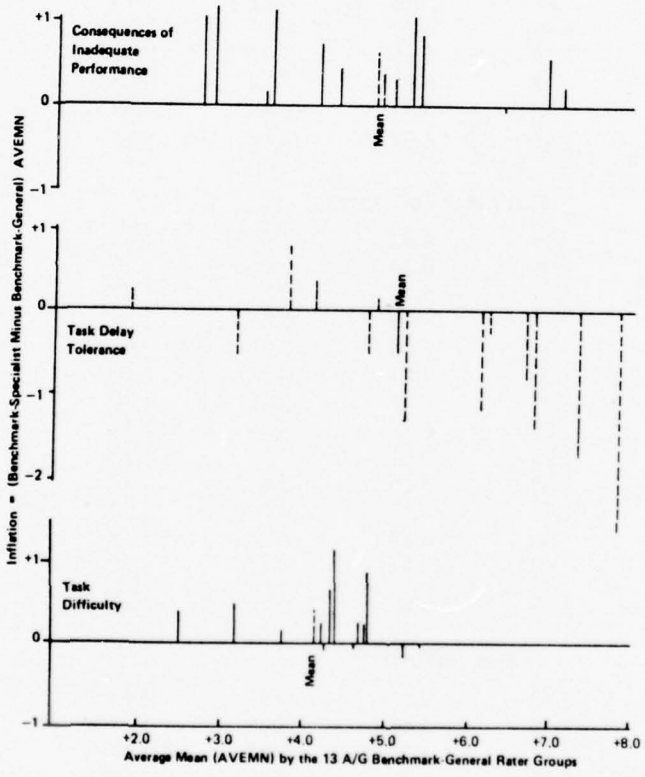


Figure 4. **Inflation of ratings by A/G specialist raters.**

V. FINAL RATER AGREEMENT STUDY

For the final rater agreement study, a minimum of two AFSs from each of the aptitude areas were selected for the purpose of comparing the benchmark (B) scales and the relative (R) scales. Raters from skill levels 5, 7, and 9 were asked to rate the complete job inventories from their career ladder on at least one task factor. Table 6 shows the distribution of raters used in the final rater agreement study by AFS, task factor, and type of rating scale used.

Table 6. **Number of Raters by Factor and Type of Rating Scale for 11 AFS in Final Rater Agreement Study**

Air Force Specialty		Minimum Aptitude Requirement	Number of Raters					
			Consequences		Delay Tolerance		Task Difficulty	
			Bench.	Ralative	Bench.	Ralative	Bench.	Relative
293X3	Radio Operator	A60	51	45	49	50	49	78
651X0	Procurement	A70	67	61	71	63	59	101
531X5	Non-Destructive Inspection	G50	61	–	67	–	67	55
906X0	Medical Administration	G60	77	105	87	104	101	78
304X4	Ground Radio Communication Equip	E80	66	60	57	58	55	122
304X0	Radio Relay Equip	E80	39	35	39	50	44	89
423X4	Pneudraulic Repair	E,M40	60	–	71	–	69	73
552X5	Plumbers	M40	69	82	66	62	69	116
423X1	Environmental Systems	M40	52	33	52	34	52	77
427X5	Airframe Repair	M40	71	63	77	65	74	75
631X0	Fuel Specialists	G,M40	71	–	71	–	74	75

Prior to gathering the final rater agreement study data, some small changes were made to the layout of the benchmark scales, to the accompanying instructions and to the definition of the Task Difficulty factor. The layout of the scales was simplified and shortened by deleting AFS code numbers, shortening the long definition of the factors on each scale and by deleting the repetition of the instructions on the scales. The main instructions on how to use the scales were shortened from two pages to one-half page. The definition of Task Difficulty was slightly altered by changing the emphasis upon the measurement of task difficulty in terms of the need for lengthy, systematic training to an emphasis upon measuring the factor in terms of the time needed to learn to do a task satisfactorily.

These changes were considered necessary to prevent rater fatigue and confusion and to clarify the essential emphasis of each scale.. The effects of the changes were studied in the electronic specialty: Ground Radio Communication Equipment Repairman (AFSC 304X4). One group of raters used the simplified benchmark scales and another group used the older benchmark scales and instructions. The pertinent comparative statistics are listed in Table 7 and show that there are no significant differences ($p = .05$) and thus, presumably, no deterioration in the performance of the raters. Hence the modifications to the scales and instructions were retained for the remaining surveys. The final modified versions of all nine scales with their accompanying instructions are displayed as Appendixes B to J. (In the survey, the benchmark scale was presented as a separate card, thus allowing a rater to refer to the scale throughout the rating process.)

The final rater agreement study sought to answer a series of questions; Can the benchmark scales be used to obtain reliable task factor ratings? Do raters using the benchmark scales converge on the same vector as they do using the relative scales? How do the benchmark scales compare with the relative scales for efficiency? How well do specialists use the benchmark scales to rate tasks from their own specialty? Finally, are the benchmark factor measurements comparable across specialties?

The first step in preparing to answer these questions was to identify, by a fixed selection rule, divergent raters and remove them from the samples. This ensured that biased samples of raters, due to subjective analyst decisions, did not exist; and thus, the remaining groups of raters' performances could be compared. From the CODAP program REXALL, raters were selected as being divergent if the correlation between each individual's mean ratings and the mean ratings for

Table 7. **Simplified and Old Benchmark Scales Compared on Various Statistics for theAFS 304X4 Ground Radio Communications Equipment Ladder**

Task Factor	Raw R_{11}	Raw R_{5050}	Zero Order Correl.	No. of Raters	Percent Raters Deleted	Aver Mean	SD Mean	Aver SD	SDSD
Consequences of Inadequate Performance									
Simplified Benchmark	.130	.882	.938	63	5	4.358	.738	1.788	.157
Old Benchmark	.114	.866		51	6	4.476	.723	1.820	.143
Task Delay Tolerance									
Simplified Benchmark	.154	.901	.957	49	14	5.131	.834	1.830	.217
Old Benchmark	.242	.941		49	13	5.198	.897	1.529	.240
Task Difficulty									
Simplified Benchmark	.257	.945	.978	55	0	5.268	1.082	1.761	.202
Old Benchmark	.264	.947		56	7	5.171	1.076	1.714	.232

Note. — Aver Mean = Mean of all task means.
SD Mean = Standard Deviation of the task means.
Aver SD = Mean of the task standard deviation.
SDSD = Standard deviation of the task standard deviation.

the total sample was not significant.[1] Interrater reliability coefficients and other pertinent statistics for the benchmark and relative samples for each factor, *after divergent raters* have been removed, are displayed in Tables 8, 9, and 10.

The interrater reliability coefficients were then adjusted to suppress the effect of rater response set. In all cases, F ratios indicated that the corresponding R_{11} values were significantly above zero to warrant their further use.

To address the first question, which deals with the ability of the benchmark scales to yield reliable ratings, differences in the corresponding benchmark and relative R_{11} values were evaluated by testing the unit normal deviates. The process was to first convert each R_{11} value to a Fisher Z score, then compute unit normal deviates by dividing the difference between the benchmark Z and the relative Z by the standard error of their difference (Haggard, 1958). Details of this reliability test are given in Appendix K.

The results of the significance tests of differences between benchmark and relative R_{11} values are presented in Table 11. At a probability of 0.05, the benchmark R_{11} values are significantly higher than the relative R_{11} values in 14 comparisons, not significantly different in 10

[1] The student t score for each rater is calculated from:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$ (McNemar, 1969)

where r = correlation between the sample's mean ratings and each individual rater's mean ratings
N = number of tasks rated by each individual rater
N − 2 = degrees of freedom

Table 8. **Comparisons Between Benchmark and Relative Raters on the Task Difficulty Factor for 11 AFS Ladders**

Specialty/Aptitude		Type of Scale	Raw R_{11}	K	Raw R_{5050}	Raters Deleted, Percent	Aver[a] Mean	SD[a] Mean	SDSD[a]	Zero Order Correl.
293X3	Radio Operator A60	B	.255	27.99	.945	16	4.840	.964	.208	.585
		R	.148	38.48	.896	24	4.351	.862	.453	
651X0	Procurement A70	B	.280	54.07	.951	5	4.881	1.032	.157	.929
		R	.225	60.73	.935	22	4.926	.918	.194	
906X0	Medical Administration G60	B	.308	93.11	.957	2	4.340	1.100	.193	.943
		R	.325	45.76	.960	6	4.610	1.130	.221	
531X5	Non-Destructive Inspection G50	B	.222	57.30	.935	6	5.123	.961	.180	.728
		R	.136	41.96	.887	16	4.866	.769	.342	
304X4	Ground Radio Communications Equipment E80	B	.257	49.25	.945	0	5.268	1.082	.202	.921
		R	.259	63.28	.946	11	4.880	.833	.175	
304X0	Radio Relay Equipment E80	B	.270	39.74	.949	2	5.230	.956	.171	.780
		R	.181	62.07	.917	11	4.986	.710	.198	
423X4	Pneudraulic Repair E,M40	B	.189	55.42	.921	10	4.787	.832	.196	.818
		R	.216	52.80	.932	11	5.141	.823	.187	
552X5	Plumbers M40	B	.227	64.29	.936	3	4.291	.974	.205	.893
		R	.249	73.52	.943	34	5.079	.916	.184	
423X1	Environmental Systems M40	B	.156	43.41	.902	6	4.681	.743	.169	.924
		R	.215	49.67	.932	14	4.987	.773	.193	
427X5	Airframe Repair M40	B	.231	63.13	.938	11	5.103	.999	.185	.934
		R	.228	48.74	.936	28	5.023	.928	.214	
631X0	Fuel Specialists G,M40	B	.237	65.71	.940	7	4.298	.868	.255	.914
		R	.242	41.76	.941	25	5.146	.959	.206	

[a]Based or adjusted to a (1–9) point scale.

Table 9. **Comparisons Between Benchmark and Relative Raters on the Task Delay Tolerance Factor for Eight AFS Ladders**

Specialty/Aptitude		Type of Scale	Raw R_{11}	K	Raw R_{5050}	Raters Deleted, Percent	Aver[a] Mean	SD[a] Mean	SDSD[a]	Zero Order Correl.
293X3	Radio Operator A60	B	.314	32.63	.958	12	5.113	1.298	.308	.918
		R	.248	30.04	.943	24	4.472	1.066	.286	
651X0	Procurement A70	B	.181	56.08	.917	17	5.368	.969	.184	.907
		R	.157	48.78	.903	19	4.315	.808	.240	
906X0	Medical Administration G60	B	.171	80.21	.912	5	5.483	.856	.198	.947
		R	.143	88.97	.893	5	4.583	.723	.161	
531X5	Non-Destructive Inspection G50	B	.192	46.24	.922	25	5.260	.980	.216	–
304X4	Ground Radio Communications Equipment E80	B	.154	45.96	.901	14	5.131	.834	.217	.730
		R	.066	45.45	.779	16	4.364	.516	.191	
304X0	Radio Relay Equipment E80	B	.177	36.41	.915	0	5.589	.899	.256	.897
		R	.199	36.55	.926	20	4.813	.916	.230	
423X4	Pneudraulic Repair E,M40	B	.216	56.84	.932	11	4.990	.959	.303	–
552X5	Plumbers M40	B	.126	45.10	.878	28	4.923	.957	.210	.466
		R	.098	43.51	.845	27	4.546	.630	.203	
423X1	Environmental Systems M40	B	.127	42.10	.880	10	5.391	.858	.222	.723
		R	.147	20.92	.896	29	5.543	.980	.276	
427X5	Airframe Repair M40	B	.212	57.45	.931	21	5.289	1.205	.255	.847
		R	.165	51.31	.908	20	4.664	.880	.231	
631X0	Fuel Specialists B,M40	B	.209	61.17	.929	11	4.319	1.052	.223	–

[a]Based or adjusted to a 1 to 9 point scale.

Table 10. **Comparisons Between Benchmark and Relative Raters on the Consequence of Inadequate Performance Factor for Eight AFS Ladders**

	Specialty/Aptitude	Type of Scale	Raw R_{11}	K	Raw R_{5050}	Raters Deleted, Percent	Aver[a] Mean	SD[a] Mean	SDSD[a]	Zero Order Correl.
293X3	Radio Operator A60	B	.225	33.97	.935	8	4.779	1.057	.221	.913
		R	.218	33.01	.933	20	4.545	.851	.231	
651X0	Procurement A70	B	.099	53.90	.846	12	4.441	.669	.143	.890
		R	.103	56.60	.851	5	4.662	.573	.159	
906X0	Medical Administration G60	B	.104	72.03	.853	3	4.590	.702	.137	.847
		R	.109	95.37	.860	5	4.050	.553	.144	
531X5	Non-Destructive Inspection G50	B	.150	54.09	.898	7	4.994	.917	.164	–
304X4	Ground Radio Communications Equipment E80	B	.130	57.68	.882	5	4.358	.738	.157	.922
		R	.120	54.58	.872	7	4.621	.639	.162	
304X0	Radio Relay Equipment E80	B	.251	36.27	.944	5	4.346	1.015	.224	.873
		R	.091	32.79	.834	3	4.931	.630	.174	
423X4	Pneudraulic Repair E,M40	B	.240	50.97	.941	10	5.290	1.078	.178	–
552X5	Plumbers M40	B	.076	56.91	.804	16	4.331	.647	.264	.817
		R	.132	72.82	.883	6	4.236	.690	.155	
423X1	Environmental Systems M40	B	.208	44.44	.929	6	5.194	.952	.190	.943
		R	.143	27.42	.893	6	4.968	.857	.266	
427X5	Airframe Repair M40	B	.242	55.60	.941	20	4.547	1.246	.235	.809
		R	.139	59.32	.890	5	4.950	.756	.183	
631X0	Fuel Specialists G,M40	B	.248	53.01	.943	20	4.838	1.209	.163	–

[a]Based or adjusted to a 1 to 9 point scale.

Table 11. **Comparison of Standardized R_{11} Values Derived from Benchmark and Relative Scale Data**

	Specialty/Task Factor	Benchmark Standard R_{11}	Relative Standard R_{11}	Benchmark F_{11}	Relative F_{11}	Benchmark K	Relative K	Number of Tasks	$\frac{Z_B - Z_R}{\sigma Z_{B-R}}$	Significantly different R_{11}s (p - .05)
				Task Difficulty						
292X3	Radio Operator	.3573	.2206	16.5574	11.8910	27.99	38.48	345	3.0179	Yes
651X0	Procurement	.3947	.3208	36.2554	30.1530	54.07	61.73	328	1.6494	No
531X5	Non-Destructive Inspec.	.3422	.1850	30.8139	10.5258	57.30	41.96	230	8.0254	Yes
906X0	Medical Administration	.4601	.4275	80.3416	35.1704	93.11	45.76	813	11.6662	Yes
304X4	Ground Radio Comm. Equip.	.4123	.3352	35.5447	32.8984	49.25	63.28	730	1.0342	No
304X0	Radio Relay Equipment	.3942	.2590	26.8599	22.6941	39.74	62.07	322	1.4917	No
423X4	Pneudraulic Repair	.2920	.2969	23.8557	23.2930	55.42	52.80	575	.2830	No
552X5	Plumbers	.3327	.3102	33.0522	34.0607	64.29	73.52	407	−.3002	No
423X1	Environment Systems	.2797	.3120	17.8595	23.5259	43.41	49.67	736	−3.6924	Yes
427X5	Airframe Repair	.3573	.3018	36.1012	22.0709	63.13	48.74	252	3.8546	Yes
631X0	Fuel Specialists	.3453	.3397	35.6517	22.4841	65.71	41.76	374	4.4018	Yes
				Consequences of Inadequate Performance						
293X3	Radio Operator	.3686	.3142	20.8310	16.1257	33.97	33.01	345	2.3352	Yes
651X0	Procurement	.2222	.2174	16.3954	16.7243	53.90	56.60	328	−.1777	No
906X0	Medical Administration	.2298	.2580	22.4860	34.1509	72.03	95.37	813	−5.9141	Yes
304X4	Ground Radio Comm. Equip.	.2814	.2650	23.5852	20.6842	57.68	54.58	730	1.7548	No
304X0	Radio Relay Equipment	.4026	.2467	25.4440	11.7370	36.27	32.79	322	6.8193	Yes
552X5	Plumbers	.1586	.2651	11.7271	27.2722	56.91	72.82	407	−8.4255	Yes
423X1	Environmental Systems	.3053	.2767	20.5260	11.4885	44.44	27.42	736	7.7445	Yes
427X5	Airframe Repair	.3821	.2741	35.3834	23.3924	55.60	59.32	252	3.2430	Yes
				Task Delay Tolerance						
293X3	Radio Operator	.3700	.3254	19.8800	15.4922	32.15	30.04	345	2.2718	Yes
651X0	Procurement	.2762	.2464	22.4017	16.9510	56.08	48.78	328	2.4928	Yes
906X0	Medical Administration	.2581	.2510	28.9074	30.8161	80.21	88.97	813	−.9050	No
304X4	Ground Radio Comm. Equip.	.2822	.1286	19.0713	7.7081	45.96	45.45	730	12.0870	Yes
304X0	Radio Relay Equipment	.2825	.2681	15.3369	14.3880	36.41	36.55	322	.5634	No
552X5	Plumbers	.1681	.1553	10.9107	8.9965	49.05	43.51	407	1.9199	No
423X1	Environmental Systems	.2427	.2171	14.4911	6.7992	42.10	20.92	736	10.0650	Yes
427X5	Airframe Repair	.3300	.2458	29.2916	17.7240	57.45	51.31	252	3.9349	Yes

comparisons, and significantly lower in three comparisons. Thus, it can at least be said that the benchmark scales can be used by raters to give consistent and reliable ratings of the three task factors. Two of the three instances in which the relative scales provided significantly more reliable data than the benchmark scales involved AFSs with mechanical aptitude requirements. An in-house study of the mechanical benchmark scales was undertaken to shed light on the apparent anomaly.

In that study, 20 psychologists from the Occupation and Manpower Research Division of AFHRL were asked to indicate whether they thought that groups of 5, 7 and 9-skill-level persons from nine mechanical AFSs would have sufficient general mechanical knowledge and experience to be able to use the mechanical benchmark scales and instructions. For example, would these experienced maintenance personnel be able to relate levels of delay tolerance of tasks from their own job inventory with the delay tolerance of tasks on the benchmark scale? The psychologists' general opinion was that non-aircraft-maintenance personnel would not have sufficient general knowledge and experience to be able to reliably relate to the wide variety of tasks on the Consequences scale and would also have real problems using the Delay and Difficulty scales.

Further investigation of the individual tasks on the mechanical Consequences scale revealed that only four of the 27 tasks were likely to have any real meaning to the Plumbers AFS, and these were at benchmark level 6 or below. For the General Purpose Vehicle AFS, the Consequences scale was thought to have 10 meaningful tasks, but again these tasks were concentrated in the lower levels. Thus the 9-level, 27-task M benchmark Consequences scale was effectively reduced to only a few levels and a small number of tasks for non-aircraft mechanical AFSs. In comparison, Aircraft Loadmaster and Airframe Repair AFS personnel, should find at least 14 and 16 tasks, respectively, on the Consequences scale that would be of real meaning to them; these tasks being spread evenly throughout the nine levels of the scale.

The significantly high intraclass reliability coefficients achieved by the aircraft-associated maintenance specialties using the mechanical consequences benchmark scale, supports these findings. Thus, the three mechanical benchmark scales should be administered with discretion and preferably only used by aircraft-associated maintenance specialties.

To answer the question: "Do raters using the benchmark scales converge on the same vector as they do using the relative scales"; zero-order correlation coefficients between the relative and benchmark means vectors were computed for each factor. Those coefficients are reported in Tables 8, 9, and 10. All but seven of the 23 correlation coefficients are above .85, and only two are below .73. In those two cases, Radio Operator task difficulty and Plumbers task delay tolerance, the low relative scale intraclass reliabilities (refer to Table 11) contributed to the poor correlations.

The USAF Occupational Measurement Center has shown that a major split in job types exists within the 293X3 Radio Operator career ladder: a portion of these personnel operate airborne radio equipment and are also required to do some equipment troubleshooting, while the remainder operate ground equipment. It is thought that this split in job types has resulted in biased task difficulty data. Raters' perception of task difficulty, defined in terms of the relative amount of time to learn to do a task, could result in widely varying ratings—which would be particularly noticeable when a relative scale is used. Thus a lower relative scale intraclass reliability coefficient is derived. In comparison, radio operators' perception of the consequences of inadequate performance and task delay tolerance factor would not be so diverse, irrespective of the distinct job groups within the career ladder. All 293X3 personnel, when rating these factors, do not require actual task performance by the rater for that rater to be able to understand the responsibilities associated with all tasks, and thereby give a valid, reliable rating.

The poor correlation coefficient for 552X5 Plumbers task delay tolerance is thought to be due to differences in the benchmark and relative rater samples. The disproportionately high percentage of raters that had to be deleted from both benchmark and relative samples (28 and 27 percent, respectively) indicates that the raters themselves had widely diverse thoughts about the

delay tolerance associated with individual tasks. One 7-skill rater exemplified this when he commented in his task delay tolerance survey booklet "without any situations in which to apply the tasks, a priority ranking is nothing more than a preference ranking–the situation and not the task determines the priority (delay tolerance) of getting the task done." However, that particular rater rated 98% of all the tasks–the same number as his contemporaries did, on the average. Contributing to the Plumbers' poor task delay tolerance factor correlation coefficient were the inherent problems of the M scale when used by a non-aircraft-maintenance specialty.

The general conclusion that the benchmark scales are measuring devices that permit raters to rank tasks in the same order as raters using the traditional relative scales is acceptable in light of the large number of sufficiently high correlation coefficients between the benchmark and relative means vectors.

To answer the question on relative efficiency of the scales, the percentage of raters deleted was considered the primary statistic to be evaluated. The percentages of raters deleted, listed in Tables, 8, 9 and 10, show a lower percentage for a large majority of benchmark scale results. The average benchmark scale percentage of deletion is 10.1%, while the average relative scale percentage of deletion is 15.4%. These results, although based on only a few specialties within each aptitude area (Tables 8, 9, and 10), warrant the conclusion that the use of benchmark scales to gather task factor data is more cost efficient since a smaller sample size is generally required.

The question of whether the specialists can use the benchmark scales to rate tasks from their own specialty is answered by comparing a number of statistics on Tables 8, 9, and 10. The best indication that specialists are able to understand and use the benchmark scales is the intraclass correlation coefficients. These coefficients have already been discussed and shown to be generally superior to the relative scale agreement coefficients. Other characteristics to consider when answering this question are the tendency of the raters to use the full range of the scales and the variance of the task ratings. The statistic that best expresses the use of the full 9-point scale is the standard deviation of the task means (SDMN). For the benchmark scales, the difficulty, delay and consequences average SDMNs are, respectively, 0.9556, 0.9877, and 0.9300 which are significantly higher ($p = .05$) than the corresponding relative scale average SDMNs of: 0.8745, 0.8090, and 0.6936. A statistic that describes the concordance of the ratings as the full range of the scale is used is the standard deviation of the task standard deviations (SDSD). The benchmark difficulty, delay, and consequences average SDSDs are, respectively, 0.1929, 0.2355, and 0.1887. These are comparably as low as the relative scale average SDSDs of 0.2333, 0.2258, and 0.1843. Although the initial validation study showed that specialists tend to inflate their ratings when using the benchmark scales, those same raters will also show a better interrater agreement, will tend to more often use the high and low ends of the scales, and will provide ratings that have acceptably low task mean variance.

The final question as to whether the benchmark factor measurements can be compared across specialties is difficult to answer conclusively because of the low number of specialties studied. It has already been shown in the initial validation study that specialists tend to inflate ratings of tasks from their own specialty, and yet are able to reliably rank order these tasks. But the variable inflationary effects of their ratings preclude across-specialty comparisons when 5, 7 and 9-skill raters are used. This finding was borne out by the fact that the average of benchmark task means (AVEMN) within aptitude areas showed no significant correlation with the minimum aptitude requirement of the AFSs studied. This also supports the position that the benchmark scales, as they presently exist, should not be used by 5, 7 and 9-skill level personnel to obtain cross-specialty comparative task factor information. A possible means of eliminating inflationary effects of task ratings is to require that tasks across specialties be observed and rated against the scales by unbiased technical observers. That process is used in the current AFHRL aptitude requirements research work unit and is successful in the sense that highly reliable cross-specialty comparisons of task difficulty are being made.

VI. CONCLUSIONS AND RECOMMENDATIONS

Benchmark scales for measuring task difficulty, probable consequences of inadequate performance, and task delay tolerance in the electronic, mechanical, and administrative/general aptitude areas have now been developed. The methodology has been established and can be used to create new scales as the old ones become outdated.

In using benchmark scales, raters have to use technical knowledge outside their past and current job experiences. Nevertheless, the evidence indicates that the benchmark scales will allow experienced raters to provide better interrater agreement than do the relative scales. Furthermore, the desired level of stability of the means is obtained more efficiently as fewer deletions and thus fewer raters are necessary when the benchmark scales are used. But when raters are not familiar with a large majority of the tasks listed on the benchmark scales, then the task factor data is suspect. This was borne out in this study when non-aircraft-maintenance personnel could not relate to the mechanical benchmark scales and, as a result, unreliable data were obtained. The conclusion then is that this particular series of benchmark scales can and should be used in lieu of the relative scales by experienced raters when task factor data are needed to assist in making within-specialty training priority decisions.

This study has also shown that specialists, when rating their own job inventories, tend to inflate their ratings of the task factors and that this inflation is not consistent. Unless an adjustment is made or precautions are taken to guard against inflated ratings, this particular series of benchmark scales should therefore not be used when ratings of the task factors are to be compared across specialties.

REFERENCES

Christal, R.E. Implications of Air Force occupational research for curriculum design. In B.B. Smith & J. Moss, Jr. (Eds.), *Report of a seminar: Process and techniques of vocational curriculum development.* Minnesota Research Coordinating Unit for Vocational Education, University of Minnesota, Minneapolis, Minnesota, April 1970, 27–61.

Christal, R.E., & Weissmuller, J.J. New CODAP programs for analyzing task factor information. *Proceedings of the 17th Annual Conference of the Military Testing Association,* Indianapolis, September 1975, 266–282.

Christal, R.E., & Weissmuller, J.J. *New CODAP programs for analyzing task factor information.* AFHRL-TR-76-3, AD-A026 121. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, May 1976.

Goody, K. *Task factor benchmark scales for training priority analysis: Overview and development phase for administrative/general aptitude area.* AFHRL-TR-76-15, AD-A025 847. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1976. (a)

Goody, K. *Comprehensive Occupational Data Analysis Programs (CODAP): Use of REXALL to identify divergent raters.* AFHRL-TR-76-82, AD-A034 327. Lackland AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, October 1976. (b)

Goody, K., & Watson, W.J. Task factor benchmark scales for use in determining training priority. *Proceedings of the 17th Annual Conference of the Military Testing Association,* Indianapolis, September 1975, 559–565.

Haggard, E.A. *Intraclass correlation and the analyses of variance.* New York: Dryden Press Inc., 1958, 25–26.

Lindquist, E.F. *Design and analysis of experiments in psychology and education.* Boston: Hougton Mifflin, 1953, 359–361.

McNemar, Q. *Psychological Statistics.* New York: John Wiley and Sons Inc., 1969, 156–157.

Mead, D.F. *Establishing job task training priorities.* Paper presented at the First International Learning Technology Congress, Washington, D.C., 20-24 July 1976.

Mead, D.F. Determining training priorities for job tasks. *Proceedings of the 17th Annual Conference of the Military Testing Association,* Indianapolis, September 1975, 551–558.

Occupational Survey Report on Radio Operator Career Ladder by USAF Occupational Measurement Center. Lackland AFB, Texas, 1975, 3.

APPENDIX A: CONSEQUENCES OF INADEQUATE PERFORMANCE

Explanation

1. Fold out the back cover of this booklet.

2. Read the Definition and Explanation of Consequence of Inadequate Performance at the top of the fold-out. You will notice that consequences has two aspects, risk of personal injury on the one hand and monetary loss on the other. We are concerned with both aspects.

3. Now read the Benchmark Scale below the Definition and Explanation on the fold-out page. Notice that it consists of task statements describing tasks performed in a wide variety of specialties. They are divided into nine sets of three, each set defining by example a level of Consequences of Inadequate Performance. The probable consequences of inadequate performance of the three tasks at Level 1 are not very serious. For the three tasks at Level 2 the consequences of inadequate performance are a little more serious, but still not very serious. The degree of seriousness of inadequate performance continues to increase gradually level by level until at Level 9 the outcome of inadequate performance would almost certainly be disastrous. If some of the tasks on the scale seem to be out of place, remember that we are concerned with the most probable outcome rather than either the most pessimistic or the most optimistic outcome. For example, inadequate performance of three tasks at Level 7 could result in disaster; on the other hand, in different circumstances, the consequences could be slight. The most probable consequences are very serious, but not disastrous.

4. Probably none of the tasks in the scale is performed in your specialty. This doesn't matter. These tasks form a reference point for each level of the scale to assist you in rating tasks in your own specialty. We have found that NCOs like yourself can use them for this purpose.

5. We want you to rate the tasks in this booklet on Consequences of Inadequate Performance, using the nine levels as defined in the Benchmark Scale. Before starting, read through the Instructions on the opposite page, and re-read the definition and scale on the fold-out.

Consequences of Inadequate Performance

Instructions

1. Keep the definition and scale folded out where you can constantly refer to it as you progress through the booklet.

2. Consider each task in turn.

   a. What is the most probable outcome of inadequate performance of the tasks? Is life or health in danger? Will it cost money to rectify? How serious are these consequences?

   b. Select the set of three tasks in the Benchmark Scale that you feel have about the same degree of Consequences of Inadequate Performance.

c. The level corresponding to this set of three task statements is your rating for the task under consideration. Indicate this rating in the field provided to the right of the task statement by darkening the circle containing the corresponding number.

(1) Use a No 2 medium lead pencil to mark your ratings. Do not use any other type of writing tool.

(2) If you change your mind on a rating, or accidently mark the wrong circle, neatly erase the incorrect mark and mark the correct circle.

(3) Confine your mark to one circle only for each task.

(4) Rate as many tasks as you can. Omit only those on which you have absolutely no knowledge.

(5) Rate relative to the tasks in the benchmark scale rather than the other tasks in your career ladder. Do not be concerned if you feel you are favoring either end of the scale.

3. When you have finished rating, turn to the background information page of this booklet. Complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

4. Return the completed booklet to the CBPO.

CONSEQUENCES OF INADEQUATE PERFORMANCE

DEFINITION

Consequences of inadequate performance is a measure of the seriousness of the **probable** consequences of inadequate performance of a task. It is measured in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours of work, etc.

BENCHMARK SCALE

Level 1 – Least Serious Consequences of Inadequate Performance

Deliver newspaper to local distribution points (Information Specialist AFSC 79150)
Clean flight planning room (Command and Control Specialist AFSC 27450)
Fold or count hospital linen (Medical Service Specialist AFSC 90250)

Level 2

Tatto identification on Air Force working dogs (Veterinary Specialist AFSC 90850)
Arrange, mark and display property for best sales results (Materiel Facilities Specialist AFSC 64750)
Enter daily work assignments on time cards (Administration Specialist AFSC 70250)

Level 3

Draw up work rosters for taxi operators or drivers on large AF base (Programs and Work Control Specialist AFSC 55530)
Compute selling price for processed meat and meat produce (Meatcutter AFSC 61250)
Compute quantity of earth to be removed or used for fill (Geodetic Surveyor AFSC 22250)

Level 4

Finish and polish gold alloy inlays, crowns or fixed partial dentures (Dental Laboratory Specialist AFSC 98250)
Operate keypunch machine to keypunch data cards (Personnel Specialist AFSC 73250)
Perform normal satellite photography sequence (Aerospace Control & Warning Systems Operator AFSC 27650)

Level 5

Reload computer after power failures or fluctuations (Communications Center Specialist AFSC 29150)
Detect theft of money or stock from commissaries or supply service outlets (Supply Services Specialist AFSC 61150)
Measure and record auditory acuity or hearing sensitivity (Aeromedical Specialist AFSC 90150)

Level 6

Quell disturbances involving military personnel (Security Specialist AFSC 81150)
Prepare aircrew navigation kits (Air Operations Specialist AFSC 27150)
Take and record pulses, temperatures and respirations (Medical Service Specialist AFSC 90250)

Level 7

Fit cargo parachutes to airdrop cargo (Aircrew Life Support Specialist AFSC 92250)
Analyze radarscope photographs to identify targets or evaluate target condition (Imagery Interpreter AFSC 20650)
Sterilize surgical instruments or supplies (Operating Room Specialist AFSC 90252)

Level 8

Apply first aid at scene of accident or incident (Security Specialist AFSC 81150)
Render missile safe for maintenance or verify missile safing (Missile Safety Specialist AFSC 24150B)
Alert direction finding (DF) stations when aircraft emergencies occur (Radio Operator AFSC 29353)

Level 9 – Most Serious Consequences of Inadequate Performance

Defend AF installations against attack by hostile forces or sabateurs (Security Specialist (Military Dog Qualified) AFSC 81150A)
Assist patient to maintain proper airway during surgery (Operating Room Specialist AFSC 90252)
Rescue personnel from aircraft or aerospace vehicle (Fire Protection Specialist AFSC 57150)

USE OF THE SCALE

1. For each task in turn, think of the **probable** consequences of inadequate performance. Think in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours or work, etc.
2. Decide which set of three tasks in the above scale have about the same consequences of inadequate performance.
3. The level indicated for this set of three tasks is your measure of the consequences of inadequate performance for the task under consideration.

APPENDIX B: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Consequences of Inadequate Performance. **Consequences of Inadequate Performance is here defined as the seriousness of the probable consequences of inadequate performance in doing a task.** It is conceptually measured in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours, etc.

2. Take out the loose colored card and you will find the definition of Consequences of Inadequate Performance repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of Senior NCOs who agreed that each group of three tasks in the Scale **have approximately the same probable seriousness** if those tasks were inadequately performed. The Scale thus contains tasks; such as "assisting patients to maintain proper airway during surgery;" that have a high probable seriousness if inadequately performed and would be rated as an 8 or 9. At the other extreme certain tasks; involving cleaning, drawing up rosters, verifying records, etc.; may have little probable seriousness if inadequately performed. Thus that type of task would be given a low rating.

3. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider would have approximately the **same probable seriousness if inadequately performed** as the task you are considering. The level of this group of three benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

4. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

5. Return the completed booklet to the CBPO.

CONSEQUENCES OF INADEQUATE PERFORMANCE

(Administrative/General)

DEFINITION

Consequences of inadequate performance is a measure of the **probable** consequences of inadequate performance of a task.

BENCHMARK SCALE

Level 9 – Most Serious Consequences of Inadequate Performance

Defend AF installations against attack by hostile forces or sabateurs (Security Specialist (Military Dog Qualified))
Assist patient to maintain proper airway during surgery (Operating Room Specialist)
Rescue personnel from aircraft or aerospace vehicle (Fire Protection Specialist)

Level 8

Apply first aid at scene of accident or incident (Security Specialist)
Render missile safe for maintenance or verify missile safing (Missile Safety Specialist)
Alert direction finding stations when aircraft emergencies occur (Radio Operator)

Level 7

Fit cargo parachutes to airdrop cargo (Aircrew Life Support Specialist)
Analyze radarscope photographs to identify targets or evaluate target condition (Imagery Interpreter)
Sterilize surgical instruments or supplies (Operating Room Specialist)

Level 6

Quell disturbances involving military personnel (Security Specialist)
Prepare aircrew navigation kits (Air Operations Specialist)
Take and record pulses, temperatures and respirations (Medical Service Specialist)

Level 5

Reload computer after power failures or fluctuations (Communications Center Specialist)
Detect theft of money or stock from commissaries or supply service outlets (Supply Services Specialist)
Measure and record auditory acuity or hearing sensitivity (Aeromedical Specialist)

Level 4

Finish and polish gold alloy inlays, crowns or fixed partial dentures (Dental Laboratory Specialist)
Operate keypunch machine to keypunch data cards (Personnel Specialist)
Perform normal satellite photography sequence (Aerospace Control & Warning Systems Operator)

Level 3

Draw up work rosters for taxi operators or drivers on large AF base (Programs and Work Control Specialist)
Compute selling price for processed meat and meat produce (Meatcutter)
Compute quantity of earth to be removed or used for fill (Geodetic Surveyor)

Level 2

Tattoo identification on Air Force working dogs (Veterinary Specialist)
Arrange, mark and display property for best sales results (Materiel Facilities Specialist)
Enter daily work assignments on time cards (Administration Specialist)

Level 1 – Least Serious Consequences of Inadequate Performance

Deliver newspaper to local distribution points (Information Specialist)
Clean flight planning room (Command and Control Specialist)
Fold or count hospital linen (Medical Service Specialist)

APPENDIX C: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Task Delay Tolerance. **Task Delay Tolerance is here defined as** the measure of how much delay can be tolerated between the time an airman becomes aware the task is to be performed and the time he must commence doing it. Must he commence immediately or does he have time to consult a manual, seek guidance, or even be taught how to do the task?

2. Take out the loose colored card and you will find the definition of Task Delay Tolerance repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of Senior NCOs who agreed that each group of three tasks in the Scale have approximately the **same amount of delay before the tasks must be performed.** The Scale thus contains tasks; such as "issue scramble orders to fighter aircraft;" that have the least delays and must be done immediately. Those tasks would be rated as a 1 or 2. At the other extreme certain tasks; involving review and research, cleaning and washing, etc.; may have large delays permitted. This type of task would be given high ratings of 8 or 9.

3. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider have approximately the **same delay tolerance** as the task you are considering. The level of this group of three Benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

4. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

5. Return the completed booklet to the CBPO.

TASK DELAY TOLERANCE

(Administrative/General)

DEFINITION

The Task Delay Tolerance of a task is a measure of how much delay can be tolerated between the time the airman becomes aware the task is to be performed and the time he must commence doing it.

BENCHMARK SCALE

Level 9 – Most Tolerant of Delay – Do when ready

Review or select books or publications for unit library (Administration Specialist)
Research and write feature stories in Air Force publications (Information Specialist)
Clean teeth of animals (Veterinary Specialist)

Level 8

Write item identification descriptions and specifications for catalogues (Procurement Specialist)
Interview or hire civilian personnel (Supply Services Specialist)
Prepare and analyze work flow process charts (Management Engineering Specialist)

Level 7

Monitor workload reporting systems (Manpower Specialist)
Brief personnel on state or local motor traffic laws (Safety Specialist)
Draw up work rosters for taxi operators or drivers on large Air Force base (Programs and Work Control Specialist)

Level 6

Operate computer remote inquiry terminals (Computer Operator)
Purge or clear chemical lines in film developing machines (Still Photographic Laboratory Specialist)
Service and maintain dental high-speed drilling equipment (Dental Laboratory Specialist)

Level 5

Identify military vehicles, installations or activities in visual photographs (Intelligence Operations Specialist)
Proofread or correct teletype tape or page copies (Communications Center Specialist)
Prepare daily weather maps (Weather Forecaster Specialist)

Level 4

Question suspects or witnesses (Security Specialist)
Perform colony counts on bacteria to estimate type and level of infection (Medical Laboratory Specialist)
Maintain proper temperature of food storage areas (Cook)

Level 3

Inspect runway for foreign objects (Air Operations Specialist)
Administer anaesthesia in dental surgery (Dental Specialist)
Adjust airborne radio receivers to obtain readable signals (Radio Operator)

Level 2

Quell disturbances involving military personnel (Security Specialist)
Identify tablets, capsules or liquids involved in poisoning cases (Pharmacy Specialist)
Operate safety console at missile control center during hazardous operations (Missile Safety Specialist)

Level 1 – Least Tolerance of Delay – Must do immediately

Use artificial respiration to restore breathing of accident or fire victims (Fire Protection Specialist)
Issue scramble order to fighter aircraft (Command and Control Specialist)
Assist during treatment of cardio-respiratory failure in operating room (Operating Room Specialist)

APPENDIX D: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Task Difficulty. **TASK DIFFICULTY is here defined as the time needed to learn to do task satisfactorily.** Take out the loose colored card and you will find the definition of Task Difficulty repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of senior NCOs who agreed that each group of three tasks in the Scale required approximately **the same amount of time to learn** to do those tasks satisfactorily. The Scale thus contains tasks, such as "cleaning display cases," that are very easy and require only a short time to learn. Such tasks should be rated low task difficulty. At the other extreme, certain complex tasks; such as performing deep roentgen therapy or operating missile safety consoles; take much longer to learn and could therefore be rated with a high task difficulty.

2. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider would require approximately the **same time to learn** as the task you are considering. The level of this group of three Benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

3. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

4. Return the completed booklet to the CBPO.

TASK DIFFICULTY

(Administrative/General)

DEFINITION

Task Difficulty is defined as the time needed to learn to do a task satisfactorily.

BENCHMARK SCALE

Level 9 – Most Difficult to Learn

Determine chemical composition of foreign made drugs (Pharmacy Specialist)
Perform deep roentgen therapy on tumor or cancer patients (Radiology Specialist)
Assist during treatment of cardio-respiratory failure in operating room (Operating Room Specialist)

Level 8

Operate safety console at missile control center during hazardous operations (Missile Safety Specialist)
Differentiate between actual targets and electronic countermeasures or decoys (Electronic Warfare Countermeasures)
Determine axis of attack for air-to ground attack missions (Intelligence Operations Specialist)

Level 7

Administer Intermittent Positive Pressure Breathing therapy (Medical Service Specialist)
Analyze computer stops for possible hardware malfunction (Supply Systems Specialist)
Determine position of aircraft by analysis of radarscope photographs after mission (Imagery Interpreter Specialist)

Level 6

Control or extinguish structural fires (Fire Protection Specialist)
Calculate number or amount of each food item to be prepared for therapeutic diet (Diet Therapy Specialist)
Prepare injured personnel for evacuation by litter or ambulance (Aeromedical Specialist)

Level 5

Verify labels or instructions for handling radioactive substances (Materiel Facilities Specialist)
Inspect buildings for termites or other wood destroyers (Entomology Specialist)
Prepare comparative productivity charts for work centers (Management Engineering Specialist)

Level 4

Complete and submit Radiation Exposure Registration Form (Environmental Health Specialist)
Assemble shelter manager kits (Disaster Preparedness Specialist)
Maintain imprest or petty cash account (Procurement Specialist)

Level 3

Challenge or identify unknown persons in vicinity of correctional facility (Corrections Specialist)
Deliver passenger manifests and allied documents to international border clearance authorities (Air Passenger Specialist)
Act as armed escort for personnel transferring funds (Security Specialist)

Level 2

Distribute administrative orders within unit (Administration Specialist)
Schedule health examinations for meat cutting personnel (Meatcutter)
Count property in warehouse bins or shelves (Inventory Management Specialist)

Level 1 – Least Difficult to Learn

Collect food trays or serving units from patients in hospital wards (Medical Service Specialist)
Clean display cases, furniture or fixtures in commissary (Supply Services Specialist)
Stamp time of receipt on incoming messages (Communications Center Specialist)

APPENDIX E: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Consequences of Inadequate Performance. **Consequences of Inadequate Performance is here defined as the seriousness of the probable consequences of inadequate performance in doing a task.** It is conceptually measured in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours, etc.

2. Take out the loose colored card and you will find the definition of Consequences of Inadequate Performance repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of Senior NCOs who agreed that each group of three tasks in the Scale **have approximately the same probable seriousness** if those tasks were inadequately performed. The Scale thus contains tasks; such as "locating mines using a mine detector;" that have a high probable seriousness if inadequately performed and would be rated as an 8 or 9. At the other extreme certain tasks; involving cleaning, drawing up rosters, verifying records, etc.; may have little probable seriousness if inadequately performed. Thus that type of task would be given a low rating.

3. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider would have approximately the **same probable seriousness if inadequately performed** as the task you are considering. The level of this group of three benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

4. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

5. Return the completed booklet to the CBPO.

CONSEQUENCES OF INADEQUATE PERFORMANCE

(Electronic)

DEFINITION

Consequences of inadequate performance is a measure of the seriousness of the probable consequences of inadequate performance of a task.

BENCHMARK SCALE

Level 9 – Most Serious Consequences of Inadequate Performance

Arm or disarm explosive-operated emergency egress systems in aircraft (Aircrew Egress Systems Repairman)

Install or remove nuclear warheads on missiles (Nuclear Weapons Specialist)

Locate mines using mine detector (Munitions Disposal Specialist)

Level 8

Test explosive components of Shrike Air/Ground Missile AGM45A (Missile Systems Maintenance Specialist)

Dispose of air munitions at safe disposal site by detonation, burning or neutralization (Munitions Maintenance)

Perform emergency destruct sequence on missile (Instrumentation Mechanic)

Level 7

Load aircraft and secure cargo (Flight Engineer Specialist)

Test for radiation leakage around X-ray equipment or radio isotope storage areas (Biomedical Equipment Maintenance Repairman)

Perform or practice pole-top rescues (Electrical Power Line Specialist)

Level 6

Operate and test safety devices on aircraft loading equipment such as MJ-1 or MJ-4 bomb lifts (Aerospace Ground Equipment Repairman)

Diagnose causes of malfunction or failure of automatic flight control system (Automatic Flight Control Systems Specialist)

Remove or replace components of aircraft speed brake system (Aircraft Pneudraulic Repairman)

Level 5

Lubricate bearings of powerhouse generators (Electrical Power Production Specialist)

Synchronize indicator sweep with antenna rotation (Air Traffic Control Radar Repairman)

Assemble, wire or connect component parts during computer installation (Electronic Computer Systems Repairman)

Level 4

Fabricate antenna systems and transmission lines (Electronic Warfare Systems Specialist)

Inspect and clean powerhouse smoke stacks (Plant Operator)

Test or replace field windings on relays or motors (Communications and Relay Center Equipment Repairman)

Level 3

Repair tape or wire recording systems (Avionic Communications Specialist)

Write technical training notes or lesson plans (Aircraft Control and Warning Radar Repairman)

Allocate maintenance tasks to work areas (Aircraft Maintenance Scheduling Specialist)

Level 2

Splice magnetic tape (Electronic Computer Systems Repairman)

Draw up work rosters for taxi operators or drivers on large AF base (Programs and Works Control)

Verify education level of applicants for AF enlistment (Recruiter)

Level 1 – Least Serious Consequences of Inadequate Performance

Perform area beautification (Air Traffic Control Radar Repairman)

Clean and vacuum simulator (Flight Simulator Specialist)

Clean and tip soldering irons (Flight Facilities Equipment Specialist)

APPENDIX F: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Task Delay Tolerance. **Task Delay Tolerance is here defined as** the measure of how much delay can be tolerated between the time an airman becomes aware the task is to be performed and the time he must commence doing it. Must he commence immediately or does he have time to consult a manual, seek guidance, or even be taught how to do the task?

2. Take out the loose colored card and you will find the definition of Task Delay Tolerance repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of Senior NCOs who agreed that each group of three tasks in the Scale have approximately the **same amount of delay before the tasks must be performed.** The Scale thus contains tasks; such as "emergency shutdowns;" that have the least delays and must be done immediately. Those tasks would be rated as a 1 or 2. At the other extreme certain tasks; involving maintenance of records, revision of technical orders or indices, cleaning and washing equipment, etc.; may have large delays permitted. This type of task would be given high ratings of 8 or 9.

3. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider have approximately the **same delay tolerance** as the task you are considering. The level of this group of three Benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

4. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

5. Return the completed booklet to the CBPO.

TASK DELAY TOLERANCE

(Electronic)

DEFINITION

The Task Delay Tolerance of a task is a measure of how much delay can be tolerated between the time an airman becomes aware the task is to be performed and the time he must commence doing it.

BENCHMARK SCALE

Level 9 – Most Tolerance of Delay – Do when ready

Clean or paint missile facilities or equipment (Missile Systems Maintenance Specialist)

Wash, clean or inspect maintenance vehicles (Flight Facilities Equipment Specialist)

Write test questions (Avionic Inertial and Radar Navigation Systems Specialist)

Level 8

Revise technical orders or indices (Weather Equipment Repairman)

Inventory bench stock, equipment or supplies (Flight Facilities Equipment Specialist)

Maintain electrical storage battery records (Telephone Switching Equipment Repairman)

Level 7

Clean parts or components using solvents (Avionic Navigation Systems Specialist)

Locate part or stock numbers in federal supply catalogs (Precision Measurement Equipment Laboratory Specialist)

Prepare or maintain Explosive Ordinance Disposal reports (Munitions Disposal Specialist)

Level 6

Change oil in antenna drive assemblies (Air Traffic Control Radar Repairman)

Analyze computer logic diagrams (Electronic Computer Systems Repairman)

Trace underground power cables using cable test set (Electrical Power Line Specialist)

Level 5

Tighten bolts or nuts to specified torques (Missile Systems Analyst Specialist)

Troubleshoot aircraft radio switching systems (Avionic Communications Specialist)

Perform operational tests on angle-of-attack or side-slip transmitters (Integrated Avionics Component Specialist)

Level 4

Test or check safety devices such as valves, regulators, or alarms on biomedical equipment (Biomedical Equipment Maintenance Repairman)

Load nuclear bombs, warheads or reentry vehicles onto transport aircraft (Nuclear Weapons Specialist)

Repair or adjust aircraft cockpit latches or locks (Aircrew Egress Systems Repairman)

Level 3

Perform inflight analysis of malfunctions in automatic tracking radar (Auto Tracking Radar Repairman)

Target or retarget guided missiles (Missile Systems Analyst Specialist)

Install nuclear weapon fusing systems (Weapons Mechanic)

Level 2

Perform nuclear bomb safety checks (Nuclear Weapons Specialist)

Monitor aircraft engine instruments during flight (Flight Engineer Specialist)

Check aircraft for armament safety (Weapons Control Systems Mechanic)

Level 1 – Least Tolerance of Delay – Must do immediately

Conduct emergency shutdown of missile launch facility (Missile Systems Analyst Specialist)

Render aircraft emergency egress systems safe after crash (Aircrew Egress Systems Repairman)

Perform emergency shutdowns of high pressure boilers (Plant Operator)

APPENDIX G: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Task Difficulty. **TASK DIFFICULTY is here defined as the time needed to learn to do a task satisfactorily.** Take out the loose colored card and you will find the definition of Task Difficulty repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of senior NCOs who agreed that each group of three tasks in the Scale required approximately **the same amount of time to learn** to do those tasks satisfactorily. The Scale thus contains tasks, such as "replacing light bulbs," that are very easy and require only a short time to learn. Such tasks should be rated low task difficulty. At the other extreme, certain complex tasks involving troubleshooting, calibrations, and alignments, etc., take much longer to learn and could therefore be rated with a high task difficulty.

2. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider would require approximately the **same time to learn** as the task you are considering. The level of this group of three Benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

3. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

4. Return the completed booklet to the CBPO.

TASK DIFFICULTY

(Electronic)

DEFINITION

Task Difficulty is defined as the time needed to learn to do a task satisfactorily.

BENCHMARK SCALE

Level 9 – Most Difficult to Learn

Isolate malfunctions in radar computer circuitry (Auto Tracking Radar Repairman)

Isolate malfunctions in guided missile MK 6 re-entry vehicle (Nuclear Weapons Specialist)

Perform terrain avoidance radar alignment tests on B52 Bomb Navigation System (Bomb Navigation Systems Mechanic)

Level 8

Troubleshoot ground radio communications systems and identify defective equipment (Ground Radio Communications Equipment Repairman)

Calibrate and certify airborne navigational aid test sets (Precision Measurement Equipment Laboratory Specialist)

Draw circuit, schematic or wiring diagrams (Instrumentation Mechanic)

Level 7

Repair weapons release computer system components on F4 E (Weapons Control Systems Mechanic)

Repair closed-circuit TVs in flight simulator (Flight Simulator Specialist)

Troubleshoot aircraft stall warning systems (Avionics Instrument Systems Specialist)

Level 6

Test minimum performance of electronic altimeter systems (Avionic Navigation Systems Specialist)

Perform aircraft engine starts, run-ups or shut-downs (Flight Engineer Specialist)

Use electrical wiring diagrams to locate defective switchgear components (Electrical Power Production Specialist)

Level 5

Clean, align or replace lenses or prisms in optical systems (Biomedical Equipment Maintenance Repairman)

Arm or de-arm guns during aircraft gun loading or unloading (Weapons Mechanic)

Locate and repair leaks in refrigerating systems (Refrigeration and Air Conditioning Specialist)

Level 4

Test, recharge or repair aircraft batteries (Aircraft Electrical Repairman)

Clean, inspect and adjust burners of forced-air heating systems (Heating Systems Specialist)

Remove or install aircraft external fuel tanks (Aircraft Maintenance Specialist)

Level 3

Load Aerospace Ground Equipment on aircraft, trucks, or trailers (Aerospace Ground Equipment Repairman)

Clean 50-caliber machine guns (Defensive Fire Control Systems Mechanic)

Inspect and service battery-powered emergency lighting units (Electrician)

Level 2

Label or tag equipment with identification, condition and status tags (Electrical Warfare Systems Specialist)

Notify customers of circuit failures or restorals (Tele-communications Systems Control Repairman)

Clean computer magnetic tapes using tape-cleaning machine (Computer Operator)

Level 1 – Least Difficult to Learn

Replace dessicants (drying agents) in electronic equipment (Avionic Communications Specialist)

Replace flood or security light bulbs (Electrical Power Line Specialist)

Remove snow or frost from equipment or facilities (Electronic Computer Systems Repairman)

APPENDIX H: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Consequences of Inadequate Performance. **Consequences of Inadequate Performance is here defined as the seriousness of the probable consequences of inadequate performance in doing a task.** It is conceptually measured in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours, etc.

2. Take out the loose colored card and you will find the definition of Consequences of Inadequate Performance repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of Senior NCOs who agreed that each group of three tasks in the Scale **have approximately the same probable seriousness** if those tasks were inadequately performed. The Scale thus contains tasks; involving work with munitions; that have a high probable seriousness if inadequately performed and would be rated as an 8 or 9. At the other extreme certain tasks; involving cleaning, painting, maintaining records, etc.; may have little probable seriousness if inadequately performed. Thus that type of task would be given a low rating.

3. ***Your job is to rate each task in this booklet relative to this Scale.*** Take each task in the booklet and choose from the Scale that group of three tasks that you consider would have approximately the **same probable seriousness if inadequately performed** as the task you are considering. The level of this group of three benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

4. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

5. Return the completed booklet to the CBPO.

CONSEQUENCES OF INADEQUATE PERFORMANCE

(Mechanical)

DEFINITION

Consequences of inadequate performance is a measure of the seriousness of the **probable** consequences of inadequate performance of a task.

BENCHMARK SCALE

Level 9 -- Most Serious Consequences of Inadequate Performance

Remove explosive stores from crashed aircraft (Weapons Mechanic)
Assemble and fill high-explosive bombs (Munitions Maintenance)
Determine if mines are booby trapped (Munitions Disposal Specialist)

Level 8

Hoist personnel into airborne helicopter (Helicopter Mechanic)
Remove or install aircraft emergency egress system seat catapults or rocket motors (Aircrew Egress Systems Repairman)
Check aircraft control surface travel, and rig flight control systems (Aircraft Maintenance Specialist)

Level 7

Inspect aircraft landing gear control and safety warning systems (Aircraft Electrical Repairman)
Perform initial emergency procedures for propellant leak in the propulsion system rocket engine of Minuteman missile (Missile Mechanic (Minuteman))
Perform emergency jettison of cargo (Aircraft Loadmaster)

Level 6

Spot-weld parts to airframe (Airframe Repair Specialist)
Secure engine test stands before test-running engine (Reciprocating Engine Mechanic)
Adjust aircraft steering system units (Aircraft Pneudraulic Repairman)

Level 5

Test cables for shorts, crosses, grounds or opens (Cable Splicing Specialist)
Operate aircraft loading controls such as doors, ramps or winches (Air Cargo Specialist)
Test materials for hardness, tensile strength, or other desired properties (Metal Machinist)

Level 4

Waterproof vehicles for deep fording (Vehicle Operator/Dispatcher)
Operate raw sewage lagoons or sewage operation ponds (Environmental Support Specialist)
Install floors or ceilings in buildings (Carpentry Specialist)

Level 3

Detect trends or developing problems from operationally ready rate reports (Maintenance Analysis Specialist)
Move damaged or salvaged aircraft or components (Construction Equipment Operator)
Change cutting edge on dozer, scraper or grader blades (Pavements Maintenance Specialist)

Level 2

Maintain individual training records (Airframe Repair Specialist)
Camouflage trucks or other vehicles (Vehicle Operator/Dispatcher)
Check power steering fluid level in staff cars (General Purpose Vehicle Repairman)

Level 1 -- Least Serious Consequences of Inadequate Performance

Organize and participate in AF commissioning ceremonies (Recruiter)
Use brushes to paint wooden surfaces (Protective Coating Specialist)
Clean and polish telephone instrument cases (Telephone Equipment Installer Repairman)

APPENDIX I: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Task Delay Tolerance. **Task Delay Tolerance is here defined as** the measure of how much delay can be tolerated between the time an airman becomes aware the task is to be performed and the time he must commence doing it. Must he commence immediately or does he have time to consult a manual, seek guidance, or even be taught how to do the task?

2. Take out the loose colored card and you will find the definition of Task Delay Tolerance repeated ahd a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of Senior NCOs who agreed that each group of three tasks in the Scale have approximately the **same amount of delay before the tasks must be performed.** The Scale thus contains tasks; such as "emergency shutdowns" and "the manual release of airdrop cargo," that have the least delays and must be done immediately. Those tasks would be rated as a 1 or 2. At the other extreme certain tasks; involving maintenance of records, painting, cleaning and washing equipment, etc.; may have large delays permitted. This type of task would be given high ratings of 8 or 9.

3. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider have approximately the **same delay tolerance** as the task you are considering. The level of this group of three Benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

4. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

5. Return the completed booklet to the CBPO.

TASK DELAY TOLERANCE

(Mechanical)

DEFINITION

The Task Delay Tolerance of a task is a measure of how much delay can be tolerated between the time the airman becomes aware the task is to be performed and the time he must commence doing it.

BENCHMARK SCALE

Level 9 – Most Tolerance of Delay – Do when ready

Conduct radio or TV programs to support the AF recruiting program (Recruiter)
Paint antenna supports (Outside Wire and Antenna Maintenance Repairman)
Remove paint from building exteriors using blowtorch, scrapers or sanders (Protective Coating Specialist)

Level 8

Erect steel towers or tanks (Construction Equipment Operator)
Form curved wooden parts by gluing or laminating (Carpentry Specialist)
Review miles or hours per gallon of fuel reports (Maintenance Analysis Specialist)

Level 7

Install telephones (Telephone Equipment Installer Repairman)
Cut sections of safety glass, each 4 ft X 4 ft, and grind, level and smooth edges (Airframe Repair Specialist)
Replace automobile engine pistons, fitting piston rings, pins and bearings using honers and aligners (Special Vehicle Repairman)

Level 6

Test, recharge or repair aircraft batteries (Aircraft Electrical Repairman)
Test-operate hydraulic or mechanic jacks (Aerospace Ground Equipment Repairman)
Transfer, handle and store refrigerant fluids (Refrigeration and Air Conditioning Specialist)

Level 5

Use electrical wiring diagrams to locate defective switchgear components (Electrical Power Production Specialist)
Lash or tie cargo on truck or trailer (Vehicle Operator/Dispatcher)
Remove or replace pneudraulic pressure switches (Aircraft Pneudraulic Repairman)

Level 4

Check engine oil or service engine oil systems (Flight Engineer Specialist)
Anchor landing mat planks for emergency runways (Pavements Maintenance Specialist)
Fill or test fullness of aircraft emergency oxygen bottles (Aircrew Egress Systems Repairman)

Level 3

Determine placement of dangerous cargo in aircraft (Aircraft Loadmaster)
Open or close Minuteman missile launcher closure (Missile Mechanic (Minuteman))
Maintain proper water level in high pressure steam-generating boilers (Plant Operator)

Level 2

Compute aircraft descent and landing data during flight (Flight Engineer Specialist)
Remove safety locking devices from aircraft control systems or surfaces before flight (Aircraft Maintenance Specialist)
Maintain ice level in special packages such as blood containers (Air Cargo Specialist)

Level 1 – Least Tolerance of Delay – Must do immediately

Perform emergency shutdowns of high-pressure boilers (Plant Operator)
Initiate recall of base personnel if emergency occurs (Programs and Works Control)
Manually release airdrop cargo over drop zone (Air Cargo Specialist)

APPENDIX J: EXPLANATION AND INSTRUCTIONS

1. You are asked to rate each of the tasks in this booklet to indicate Task Difficulty. **TASK DIFFICULTY is here defined as the time needed to learn to do a task satisfactorily.** Take out the loose colored card and you will find the definition of Task Difficulty repeated and a Benchmark Scale of tasks. Scan the Benchmark Scale and you will notice that it contains a variety of tasks performed by members of specialties other than your own. This Scale was developed by a large number of senior NCOs who agreed that each group of three tasks in the Scale required approximately **the same amount of time to learn** to do those tasks satisfactorily. The Scale thus contains tasks, such as "replacing air conditioner filters," that are very easy and require only a short time to learn. Such tasks should be rated low task difficulty. At the other extreme, certain complex tasks involving installing nuclear weapon fusing systems, and performing operational checks; etc., take much longer to learn and could therefore be rated with a high task difficulty.

2. **Your job is to rate each task in this booklet relative to this Scale.** Take each task in the booklet and choose from the Scale that group of three tasks that you consider would require approximately the **same time to learn** as the task you are considering. The level of this group of three Benchmark Scale Statements is your rating of the task under consideration. Mark this number (1 through 9) in the column to the right of the task statement. Please attempt to rate all tasks.

3. When you have finished rating all tasks, complete the background information block and enter an estimate of the time you spent and any comments you feel may help us in our research.

4. Return the completed booklet to the CBPO.

TASK DIFFICULTY

(Mechanical)

DEFINITION

Task Difficulty is defined as the time needed to learn to do a task satisfactorily.

BENCHMARK SCALE

Level 9 – Most Difficult to Learn

Prepare charges for emergency destruction of nuclear weapons using non-electric initiation methods (Munitions Disposal Specialist)
Remove B52 aircraft wing from aircraft (Airframe Repair Specialist)
Install nuclear weapon fusing systems (Weapons Mechanic)

Level 8

Self-test and operate atomic weapon test set (Missile Mechanic (Minuteman))
Perform operational checks on missile firing systems (Weapons Mechanic)
Use explosives to blast construction obstacles (Pavements Maintenance Specialist)

Level 7

Draw sketches or plans of parts to be machined (Metal Machinist)
Operate aircraft inflight refuelling system during flight (Flight Engineer Specialist)
Install thrust-reversing system on jet engines (Jet Engine Mechanic)

Level 6

Identify and splice priority circuits or working cables (Cable Splicing Specialist)
Remove, replace or adjust components of bomb door systems (Aircraft Pneudraulic Repairman)
Mix caustic solutions (Cryogenic Fluids Production Specialist)

Level 5

Bleed, adjust and service aircraft brake systems (Aircraft Maintenance Specialist)
Erect poles using power equipment (Outside Wire and Antenna Maintenance Repairman)
Test aviation fuel for water or other contamination (Fuel Specialist)

Level 4

Install ice-cream refrigerators (Refrigeration and Air Conditioning Specialist)
Operate aircraft cargo loading equipment such as fork lifts or cargo loaders (Air Cargo Specialist)
Center brake shoes or adjust brake anchor pins on service vehicles (General Purpose Vehicle Repairman)

Level 3

Refill mobile fuel units (Fuel Specialist)
Lash or tie cargo on truck or trailer (Vehicle Operator/Dispatcher)
Erect or use scaffolds or ladders (Protective Coating Specialist)

Level 2

Remove or replace troop seats in aircraft (Flight Engineer Specialist)
Inspect, clean or replace air conditioner filters (Refrigeration and Air Conditioning Specialist)
Stand by or operate fire extinguishers during helicopter startup (Helicopter Mechanic)

Level 1 – Least Difficult to Learn

Mow or edge grassed areas (Pavements Maintenance Specialist)
Remove or install batteries in motor vehicles (General Purpose Vehicle Repairman)
Construct sandbag revetments (Construction Equipment Operator)

APPENDIX K: TEST FOR SIGNIFICANT DIFFERENCE IN INTERRATER RELIABILITY COEFFICIENTS

Convert standardized interrater reliability coefficient into Fisher Z:

$$Z = \frac{1}{2} \log_e \left[\frac{1 + (K - 1)R_{11}}{1 - R_{11}} \right] = \frac{1}{2} \log_e F_{11}$$

$$\sigma_z^2 = \frac{K}{2(N - 2)(K - 1)}$$

where: R_{11} = interrater reliability coefficient

K = average number of raters per task

N = number of tasks rated by the rater

F_{11} = F ratio of variances between and within classes

Z = Fisher Z score

σ_z = standard error of the Fisher Z score

The test for significant difference in the Fisher Z scores is[2]:

$$\frac{Z_B - Z_R}{\sigma_{Z_{B-R}}} = \frac{\log_e \sqrt{\frac{F_B}{F_R}}}{\sqrt{\frac{1}{2(N - 2)} \left[\frac{K_B}{K_B - 1} + \frac{K_R}{K_R - 1} \right]}}$$

where: Subscript B refers to Benchmark Samples

Subscript R refers to Relative Samples

[2]When $K_B = K_R = 2$

$$\frac{Z_B - Z_R}{\sigma_{Z_{B-R}}} = \frac{\log_e \sqrt{\left[\frac{1 + R_B}{1 - R_B} \right] \Big/ \left[\frac{1 + R_R}{1 - R_R} \right]}}{\sqrt{\frac{2}{N - 2}}}$$

which is the appropriate test for significant difference between two uncorrelated Pearson Product Correlation Coefficients (R). McNemar, 1969).

☆U.S. GOVERNMENT PRINTING OFFICE: 1979- 671-056/83