

N85-11544

DECISION TREE RATING SCALES FOR WORKLOAD ESTIMATION:

THEME AND VARIATIONS

WALTER W. WIERWILLE
JULIE H. SKIPPER
CHRISTINE A. RIEGER

VEHICLE SIMULATION LABORATORY
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
BLACKSBURG, VIRGINIA 24061

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

DECISION TREE RATING SCALES FOR WORKLOAD ESTIMATION:
THEME AND VARIATIONS

Walter W. Wierwille
Julie H. Skipper
Christine A. Rieger*

Vehicle Simulation Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

SUMMARY

The Modified Cooper-Harper (MCH) scale has been shown to be a sensitive indicator of workload in several different types of aircrew tasks (Wierwille and Casali, 1983). The study to be described in this paper was undertaken to determine if certain variations of the scale might provide even greater sensitivity and to determine the reasons for the sensitivity of the scale. The MCH scale, which is a 10 point scale, and five newly devised scales were examined in two different aircraft simulator experiments in which pilot loading was treated as an independent variable. The five scales included a 15 point scale, computerized versions of the MCH and 15 point scales, a scale in which the decision tree was removed, and one in which a 15 point left-to-right format was used.

The results of the study indicate that while one of the new scales may be more sensitive in a given experiment, task dependency is a problem. The MCH scale on the other hand exhibits consistent sensitivity and remains the scale recommended for general use. The MCH scale results are consistent with earlier experiments also. This paper presents the results of the rating scale experiments and also describes the questionnaire results which were directed at obtaining a better understanding of the reasons for the relative sensitivity of the MCH scale and its variations.

INTRODUCTION

It has gradually become recognized that rating scales, properly designed and tested, represent a sensitive and economical means for estimating mental workload. They can be used in a systematic manner to obtain a single numerical response, which estimates the magnitude of the multidimensional construct of mental workload.

One of the most popular and widely accepted scales is the so-called Cooper-Harper scale (Cooper and Harper, 1969). This scale incorporates an unusual decision tree and descriptors directed at handling qualities, stability, and workload. The scale is well suited for estimation of workload in manual control systems. For example, Wierwille and Connor (1983) showed that the scale was quite sensitive to changes in turbulence level and longitudinal stability in an instrument landing task. Variations of the original scale have also appeared, but they too have been directed primarily

* Now with Hughes Aircraft Co., Ground Systems Group, Fullerton, CA

toward manual control applications (North and Graffunder, 1979; O'Connor and Buede, 1977; Siefert, Daniels, and Schmidt, 1972; and Wolfe, 1982). More recently, Wierwille developed a modification of the scale, called the Modified Cooper-Harper (MCH), which could be universally applied in mental workload estimation, regardless of the type of loading imposed by the task (Wierwille and Casali, 1983)*. In particular, the scale was designed to provide a global measure of mental workload in tasks having loading along communications, mediational, and perceptual dimensions. The scale was subsequently tested and found to be experimentally sensitive and valid in three independent simulator experiments.

Because the MCH scale had already been tested and found adequate, questions could be asked regarding the reasons for its sensitivity and regarding improvements that might be made. Thus, another study was undertaken in which the MCH scale was systematically varied in an effort to gain greater insight. Specifically, the MCH scale and five variations emphasizing major design aspects were used in this study. The six rating scales were then used in two different experiments, one involving mediational (cognitive) loading and one involving communications loading. The results are reported in this paper.

METHOD

Thirty six pilots (30 private and 6 student) participated, each participating in both experiments. Four pilots were females, and 32 were males. The pilots were tested for hearing and vision using standard tests. They were paid for their participation.

The aircraft simulator used for the two flight task experiments was a modified Singer-Link GAT-1B moving base, simulator. The simulator had three degrees of physical motion--yaw, pitch, and roll. For both experiments, the simulator was equipped with translucent blinders to eliminate outside distractions. The ambient illumination was held constant. A lapel microphone and speaker system were installed in the simulator cockpit so that the subjects could communicate with the "tower" (experimenter). To assure that the subjects were continually providing input control to the simulator, mild, random wind gusts were introduced into the simulator flight dynamics. For the mediational experiment the simulator was additionally equipped with a Kodak Ektagraphic slide projector (Model 260) mounted in front of the simulator windscreens. To computerize two of the six rating scales, a TRS-80 Model III micro-computer was used. The rating scales were programmed in BASIC, and the subject ratings were performed on the TRS-80 computer in a reduced glare setting.

Six rating scale designs were used in both the communications and the mediational experiments. The first rating scale was the Modified Cooper-Harper (MCH) rating scale described earlier. The MCH scale has a 3-3-3-1, decision tree scale structure. The second rating scale, COMPMCH, was a computerized version of the MCH scale. The TRS-80 was used to administer the MCH scale to the subjects on a decision-by-decision basis. The subjects

* Figure 1 of Wierwille and Casali (1983) shows the MCH scale.

were only permitted to deal with one primary decision at a time. Thus, the subjects did not know where each primary decision would lead on the rating scale. (A typical computer frame of the COMPMCH scale is illustrated in Figure 1). The computer implemented scale was used to discover whether or not the decision tree logic of the MCH scale was being utilized or if the subjects were merely rating on the basis of the category descriptors and numerical values. After each computer rating, the subjects were asked by the computer if they were satisfied with their rating. If they were not satisfied, the program repeated the procedure for rating. When the subjects were satisfied with their rating, the rating value was recorded. To investigate the possibility of additional rating scale categories increasing the sensitivity of the MCH scale, the third rating scale, MCH+ (Figure 2), expanded the MCH scale to a 15 point decision-tree rating scale. One additional category was added to the first three rating groups and two additional categories were added to the last rating group, giving a 4-4-4-3 scale structure. The COMPMCH+ scale, the fourth rating scale, was a computerized version of the MCH+ scale and was implemented in the same manner as the COMPMCH scale.

In the fifth rating scale, the PBMCH (performance-based MCH) scale (Figure 3), the primary decision hierarchy was changed by manipulating the tree structure. The PBMCH decision tree flow was from left to right and the first decision was concerned with the errors of the subjects in performing the instructed task. This scale was used in an attempt to improve the sensitivity of the MCH scale by modifying the decision tree logic of the scale requiring an assessment of the subjects' errors first in the rating process. Finally, the sixth rating scale, the NDT (no decision tree) scale (Figure 4), removed the visual decision tree structure from the MCH scale to find out how the visual tree affected the sensitivity of the MCH scale. The NDT scale presents the MCH rating information in a tabular format.

Identical experimental designs were used in both the communications and mediational experiments. Data were analyzed as a rating scale by load (6x3) design. Load presentation order was completely counterbalanced. Each subject used only one rating scale, which was the same scale for both experiments. Six subjects used each scale, resulting in a total of 36 subjects. Thus, rating scale was a fixed-effects between-subjects variable and load level was a fixed-effects within-subject variable. Experience level was controlled by dividing the 36 subjects into sextiles according to flight hours and then selecting one subject from each sextile for each rating scale.

COMMUNICATIONS EXPERIMENT

The communications experiment task and protocol were identical to those used by Casali and Wierwille (1983) in an experiment comparing many different kinds of workload estimation techniques. The reader is referred to this earlier experiment for a detailed description of the task. Briefly, the aircraft control and communications requirements were performed simultaneously in the task. After reaching altitude, subjects maintained straight and level flight in mild turbulence until instructed to make changes.

For the communications aspect, the subjects listened to an 8-minute tape recorded message that was played over the cockpit speaker system. The taped

communications scenario was a "tower" controller with a male voice. The subjects were required to attend to two components of the taped scenario. The first component consisted of pilot commands. In the commands, the subjects were asked to change and report aircraft parameters (e.g. change altitude, heading, and radio frequency, and report airspeed, aircraft model, altitude, and heading). In the second component of the taped scenario, the subjects were presented with strings of randomly constructed aircraft call signs. Each call sign consisted of two international phonetic letters and two single digits (e.g. Alpha-Four-Bravo-One). Out of the randomly presented call signs the subjects were instructed to respond "now" to their specific call sign "One-Four-India-Echo" and to any of 5 permutations of the call sign which always featured "one" in the first position of the call sign. Thus, the subjects had six target call signs to listen for, each beginning with "one", as a cue to listen to what followed.

The communications load was varied in this experiment by manipulating the presentation rate of the target call signs and the non-target permutations of "One-Four-India-Echo." The three load levels were: low, 1 target every 12 seconds with 0 non-target permutations; medium, 1 target every 5 seconds with 30% permutations; and high, 1 target every 2 seconds with 40% permutations.

The experiment began with a practice flight which contained equal portions of all three communications load levels. The data run flights then followed --one at each load level. After each of the experimental flights, the simulator was placed in autopilot control and the subjects left the simulator to make a rating on their respective rating scale. They then completed a questionnaire. The questionnaire was administered to allow the subjects to describe the factors on which their ratings were based. After the final experimental flight the subjects landed the simulator and were dismissed. (They returned later the same day to participate in the mediational experiment. After completion of both experiments, they were debriefed, paid, and dismissed.)

In addition to the ratings, all verbal responses of the subjects were recorded and later scored for errors of omission, errors of commission, and reaction times.

COMMUNICATIONS EXPERIMENT RESULTS

The main statistical analysis results for the communication experiment are presented in Table 1. The rating scale scores for each rating scale were first subjected to a one-way analysis of variance. An α -level of 0.01 was specified to account for the fact that six different rating scale ANOVA's were performed. Mean values, in terms of Z-scores for each rating scale, were also computed and appear in the table. For those ANOVAs resulting in significance at $p < 0.01$, Duncan's multiple comparisons were carried out.

The results of the tests indicate that the MCH, COMPMCH, and PBMCH scales resulted in significant ANOVA's. All three scales increased monotonically with load. Furthermore, the three scales exhibited similar sensitivity, with the MCH showing slightly greater sensitivity than the other two.

Two multivariate analyses were performed on the voice response measures, at $\alpha = 0.05$, to test the data for performance variations due to either the rating scale groups or the pilot experience level groups. The three measures used were errors of omission, errors of commission, and response times. The Wilk's U-likelihood ratio statistic F-approximation is reported. The results showed that there were no statistically significant performance variations among the rating scale groups ($F(15, 77) = 1.40, p = 0.1663$) nor the pilot experience level groups ($F(15, 77) = 1.61, p = 0.0895$).

To obtain general information regarding the effects of pilot experience level and load presentation order on the ratings of the subjects, converted and collapsed raw score data were analyzed in two separate ANOVAs. The results indicated that neither the pilots' experience levels ($F(5, 30) = 2.43, p = 0.0579$) nor the load presentation orders ($F(1, 35) = 0.43, p = 0.5173$) affected the ratings of the pilots. The experience level results were analyzed further using regression, but the additional analyses did not provide significant findings.

The responses to the questionnaire presented to the subjects indicated a shift in tone from positive to negative as the load levels progressed from low to medium to high. A Chi-square analysis on a 2 x 3 contingency table, response type by load levels, resulted in $\chi^2 = 68.326, p < 0.0001$, confirming the change of tone due to load in the responses of the subjects. Typical response classifications were "time-sharing", "aircraft control", and "recognition of target call signs".

Finally, it is worth mentioning that the MCH scale results in this experiment were virtually identical to the MCH scale results obtained in the earlier (Casali and Wierwille, 1983) study. This indicates a high degree of repeatability for the MCH scale.

MEDIATIONAL EXPERIMENT

This experimental task and protocol were also identical to an earlier experiment in which mediational activity was emphasized (Wierwille, Rahimi, and Casali, 1984) and in which many different workload techniques were evaluated. The reader is referred to this earlier experiment for a detailed description of the task. Briefly, the overall task consisted of two components: straight and level flight in mild turbulence (within specified tolerances), and solution of navigation problems. Subjects performed the tasks simultaneously with instructions indicating equal priority.

The navigation task of solving wind triangle problems was used to interject mediational loading into the basic flight task. Wind vector triangles depicted on slides involved solving for the effects of wind direction and velocity on the path and speed of an aircraft. The slides contained both a problem triangle and a reference triangle. The reference triangle provided numerical values associated with the triangle legs and the angles corresponding to the problem triangle.

The difficulty of the navigation problems was manipulated by varying the question type, the numbers used in the mental calculation of the problems, and

the orientation of the reference triangles. Depending upon the question type, the problems required triangle comparison, triangle comparison followed by an addition or subtraction, or triangle comparison followed by an addition or subtraction and a subsequent division. For all load levels, the slide presentation rate was held constant at a rate of one slide per 25 seconds. Subjects expressed their answers verbally. These responses were recorded for later use in computing response time and number of correct responses. It is important to note that the subjects did not implement the solutions to the navigation problems. They maintained constant altitude, heading, and airspeed throughout each flight.

The general flight procedures for the mediational experiment were the same as for the communications experiment. In particular, one practice and three data flights were performed, and subjects left the simulator while in autopilot to make their ratings and questionnaire responses.

MEDIATIONAL EXPERIMENT RESULTS

The main results of the mediational experiment are presented in Table 2. The table includes individual ANOVA's at a corrected α level of 0.01, standardized mean (Z-score) values for each rating scale, and Duncan's multiple comparisons tests for those scales having significant ANOVA's.

The results indicate that only the PBMCH scale was not significant at $p < 0.01$. All of the scales exhibited monotonic increases with load. In terms of the Duncan's tests, sensitivity among those scales demonstrating significance could be ranked as follows: Most sensitive, MCH+; next most sensitive, COMPMCH and NDT; next most sensitive, MCH and COMPMCH+. However, all five scales are actually quite sensitive, considering the small sample size and strict criterion used.

To provide substantiation of the results obtained with the rating scale data, a MANOVA was performed using both mean response time and percentage of errors on the navigation problems for each experimental flight as dependent measures. When using the F-approximation of Wilks U-statistic to compare the groups of subjects assigned to each rating scale condition, there was no significant main effect of rating scale, $F(10, 58) = 1.49$, $p = 0.1684$. This result indicates that no differences in primary task performance were associated with subject assignment. The lack of a rating scale main effect suggests that conclusions regarding the sensitivity of the scales are based on true scale differences rather than group differences in primary task performance.

A second MANOVA was conducted to determine whether there was a main effect of experience level on mean response time and percent error in the mediational task. The F-approximation to Wilk's U-statistic revealed no significant differences in task performance associated with experience level, $F(10, 58) = 0.49$, $p = 0.8894$.

Using the standardized ratings for the three load presentations--first, second, or third, a one-way analysis of variance revealed no significant

differences attributed to load level presentation order, $F(2,70) = 0.37$, $p = 0.6942$. A one-way ANOVA on the sum of the standardized ratings across the load levels for each subject indicated no significant effects of experience level on the summed ratings, $F(15,30) = 1.33$, $p = 0.2815$.

The questionnaire responses to the low, medium, and high load levels were sorted into comments which were "positive" or favorable in tone and "negative" or unfavorable in tone. A Chi-square test revealed significant differences in the frequencies of the favorable and unfavorable responses across the load levels, $\chi^2 = 55.94$, $p = 0.0001$. Favorable comments occurred most often at the low load level, while unfavorable ones occurred most often at the high load level. Based on categories which were derived by sorting, it seems that the major factors which influenced the subjects' ratings were the amount of time available, the difficulty of the task, and their assessment of how well the task requirements were met.

In terms of comparison of the MCH scale results of this experiment with those of the earlier mediational experiment (Wierwille, Rahimi, and Casali, 1984), it was found that again the two were virtually identical.

CONCLUSIONS DRAWN FROM THE RESULTS OF THE TWO EXPERIMENTS

Several conclusions can be readily drawn by comparing the information contained in Tables 1 and 2. First, in terms of global sensitivity, only the MCH and the COMPMCH exhibited significance in both experiments at the $p < 0.01$ level. This finding indicates that none of the other scales possess as high a general sensitivity as the MCH scale and its computerized version. All of the other scales exhibited sensitivity in only one experiment. While the MCH+ scale and NDT scale exhibited slightly higher sensitivities than the MCH in the mediational experiment, these two scales could not be counted on to provide better results than the MCH in other types of experiments.

The table also shows that the MCH scale and COMPMCH scale are about equal in sensitivity. Apparently, computerizing the scale, such that a subject is forced to use the tree structure, has no effect on the sensitivity of the scale. In the communications experiment the MCH scale is slightly more sensitive, and in the mediational experiment the COMPMCH is slightly more sensitive. On balance, however, they have the same sensitivity.

It should be noted that each given subject used only one rating scale. Thus, the ratings for the MCH+ scale, for example, were performed by the same group of subjects in both experiments. Therefore, one cannot attribute the differences in scale sensitivity across experiments to individual differences in subject groups. All other peripheral statistical tests support the conclusion that all of the scales except the MCH and COMPMCH are task dependent.

Other conclusions can also be drawn. Does increasing the number of categories from 10 to 15 as in the MCH+ scale (Figure 2) improve sensitivity? The answer appears to be "not consistently". While the MCH+ is somewhat more sensitive in the mediational experiment, it is substantially less sensitive in the communications experiment. For the computerized version of the 15

category scale (the COMPMCH+), sensitivity is about the same as the MCH in the mediational experiment and much lower than the MCH in the communications experiment. The conclusion is that 15 categories is not generally as good as 10 categories.

Does revision of the scale to produce a left-to-right decision tree with 15 categories (the PBMCH, Figure 3) improve sensitivity? The answer to this question is "no". The PBMCH is not as sensitive as the MCH in either of the two experiments.

Finally, does a tabular format, with the decision tree removed (the NDT, Figure 4) improve sensitivity? The answer in this case is again "not consistently". While the NDT is slightly more sensitive than the MCH in the mediational experiment, it is much less sensitive than the MCH in the communications experiment.

In regard to the questionnaire responses, it was found that pilots do rate on the basis of concepts similar to those which researchers tend to think should be included in workload. While wording did vary, the subjects tended to rate on the basis of time pressure, difficulty, assessed performance, and problems of time sharing. Their comments changed in tone and frequency as expected with load level.

In general then, conflicting results between the two experiments indicate that sensitivity of most rating scales varies in subtle ways. However, the MCH scale and its computerized version are consistently sensitive and reliable. Furthermore, pilots' ratings appear to be based on factors similar to those which researchers currently consider important.

ACKNOWLEDGEMENTS

The authors wish to thank Ms. Sandra Hart, NASA Ames Research Center, who served as grant technical monitor for this project (NAG2-17). The authors also wish to thank Drs. John G. Casali, Robert D. Dryden, and Harry L. Snyder for their helpful suggestions.

REFERENCES

- Casali, J. G. and Wierwille, W. W. A comparison of rating scale, secondary task, physiological, and primary task workload estimation techniques in a simulated flight task emphasizing communications load. Human Factors, 25 (6), December, 1983, 623-641.
- Cooper, G. E. and Harper, R. P., Jr. The use of pilot rating in the evaluation of aircraft handling qualities. Moffett Field, CA: National Aeronautics and Space Administration, Ames Research Center, NASA TN-D-5153, April, 1969.
- North, R. A. and Graffunder, K. Evaluation of a pilot workload metric for simulated VTOL landing tasks. Proceedings of the Human Factors Society Twenty-Third Annual Meeting, Boston, Massachusetts, 1979, 357-361.

- O'Conner, M. F. and Buede, D. M. The application of decision analytic techniques to the test and evaluation phase of the acquisition of a major air system. McLean, Virginia: Decisions and Designs, Inc., Report TR 77-3, 1977.
- Seifert, R., Daniels, A. F., and Schmidt, K. A method of man-display/control system evaluation. Proceedings of the AGARD Conference on Guidance and Control Displays, No. 96, AGARD-CP-96, February, 1972, 8-1--8-9.
- Wierwille, W. W. and Connor, S. A. Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. Human Factors, 25 (1), February, 1983, 1-16.
- Wierwille, W. W. and Casali, J. G. A validated rating scale for global mental workload measurement applications. Proceedings, 27th Annual Meeting of the Human Factors Society, Norfolk, Virginia, October, 1983, Vol. 1, 129-133.
- Wierwille, W. W., Rahimi, M., and Casali, J. G. Evaluation of sixteen measures of mental workload using a simulated flight task emphasizing mediational behavior. Human Factors, 25, 1984. (In press, special issue, Human Factors in Aviation Psychology.)
- Wolfe, J. D. Crew workload assessment: Development of a measure of operator workload. Minneapolis, Minnesota: Honeywell, Inc. Technical Report No. AFFDL-TR-78, 1978.

TABLE 1. Communications Experiment Results

Rating Scale	ANOVA F (2, 10) P	Standardized Mean Scores*			
		L	M	H	
MCH	0.0035	-1.011	0.395	0.616	
COMPCH	0.0015	-0.510	0.000	0.510	
MCH+	0.1347	-0.594	-0.019	0.613	
COMPCH+	0.0433	-0.484	0.132	0.352	
PBMCH	0.0012	-0.983	0.046	0.937	
NDT	0.0237	-0.522	-0.095	0.617	

*Means with a common underline do not differ at $p < 0.01$ using Duncan's multiple comparisons. These comparisons were performed only on rating scale scores demonstrating significance at $p < 0.01$ in the individual ANOVAs.

TABLE 2. Mediatlional Experiment Results

Rating Scale	ANOVA F (2, 10) P	Standardized Mean Scores*			
		L	M	H	
MCH	0.0039	-0.758	-0.188	0.946	
COMPCH	0.0030	-0.667	-0.200	0.867	
MCH+	0.0001	-1.023	-0.440	1.068	
COMPCH+	0.0064	-0.808	-0.195	0.613	
PBMCH	0.3800	-0.279	-0.082	0.361	
NDT	0.0001	-0.845	-0.307	1.152	

*Means with a common underline do not differ at $p < 0.01$ using Duncan's multiple comparisons. These comparisons were performed only on rating scale scores demonstrating significance at $p < 0.01$ in the individual ANOVAs.

MINOR BUT ANNOYING DIFFICULTY: MODERATELY HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE. -4-

MODERATELY OBJECTIONABLE DIFFICULTY: HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE. -5-

VERY OBJECTIONABLE BUT TOLERABLE DIFFICULTY: MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE. -6-

TYPE THE NUMBER OF THE RATING OF YOUR CHOICE, THEN PRESS ENTER.?

Figure 1. Sample frame of the COMPMCH rating scale.

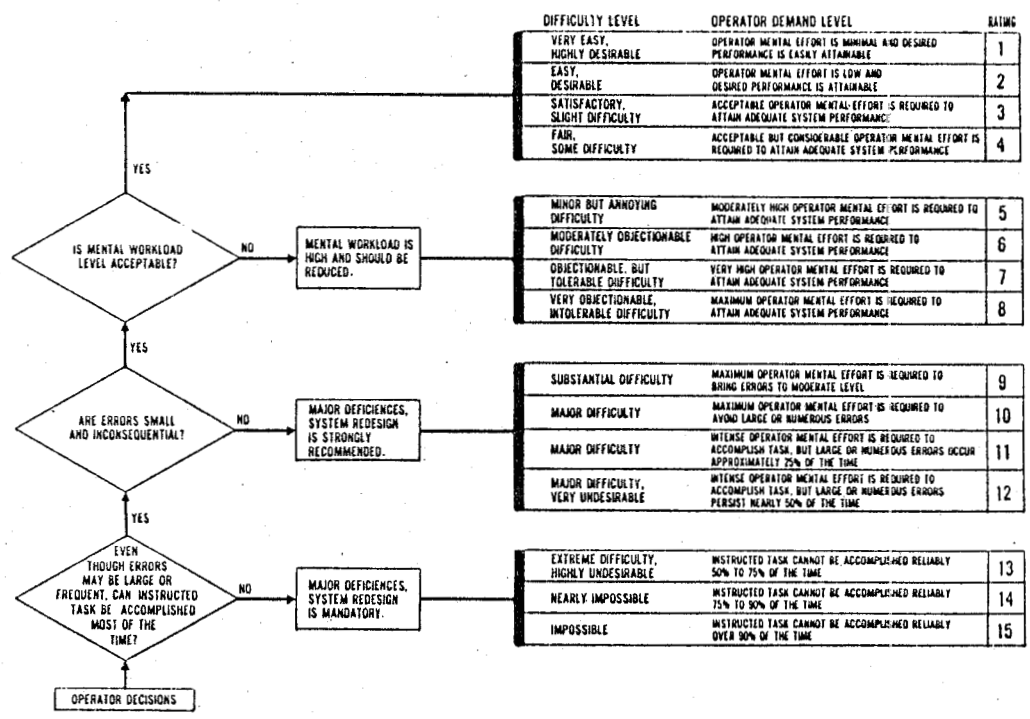


Figure 2. MCH+ rating scale (reduced in size).

ORIGINAL PAGE IS OF POOR QUALITY

ORIGINAL PAGE IS
OF POOR QUALITY

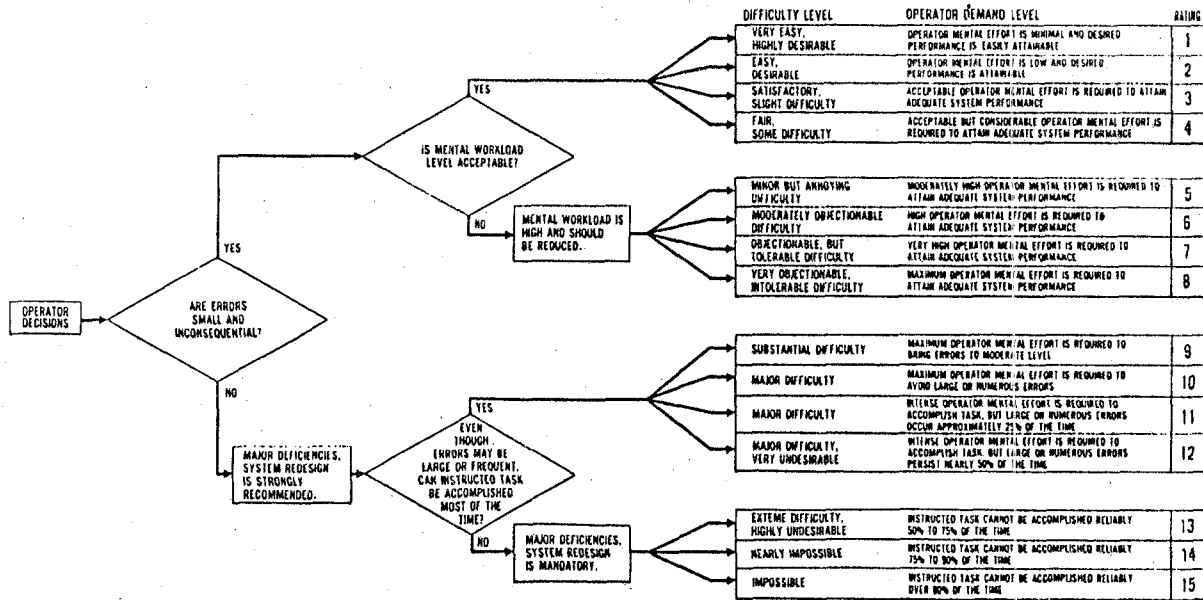


Figure 3. PBMCH rating scale (reduced in size).

	OPERATOR DECISIONS			DIFFICULTY LEVEL	OPERATOR DEMAND LEVEL	RATING
	TASK IS ACCOMPLISHABLE MOST OF THE TIME	ERRORS ARE SMALL AND INCONSEQUENTIAL	MENTAL WORKLOAD IS ACCEPTABLE			
	YES	YES	YES	VERY EASY, HIGHLY DESIRABLE	OPERATOR MENTAL EFFORT IS MINIMAL AND DESIRED PERFORMANCE IS EASILY ATTAINABLE	1
	YES	YES	YES	EASY, DESIRABLE	OPERATOR MENTAL EFFORT IS LOW AND DESIRED PERFORMANCE IS ATTAINABLE	2
	YES	YES	YES	FAIR, MOD DIFFICULTY	ACCEPTABLE OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	3
MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.	YES	YES	NO	MINOR BUT ANNOYING DIFFICULTY	MODERATELY HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	4
	YES	YES	NO	MODERATELY OBJECTIONABLE DIFFICULTY	HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	5
	YES	YES	NO	VERY OBJECTIONABLE BUT TOLERABLE DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE	6
MAJOR DEFICIENCIES, SYSTEM REDESIGN IS STRONGLY RECOMMENDED	YES	NO	---	MAJOR DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO BRING ERRORS TO MODERATE LEVEL	7
	YES	NO	---	MAJOR DIFFICULTY	MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO AVOID LARGE OR NUMEROUS ERRORS	8
	YES	NO	---	MAJOR DIFFICULTY	INTENSE OPERATOR MENTAL EFFORT IS REQUIRED TO ACCOMPLISH TASK, BUT FREQUENT OR NUMEROUS ERRORS PERSIST	9
MAJOR DEFICIENCIES, SYSTEM REDESIGN IS MANDATORY.	NO	---	---	IMPOSSIBLE	INSTRUCTED TASK CANNOT BE ACCOMPLISHED RELIABLY	10

Figure 4. NDT rating scale (reduced in size).