

THE EFFECTS OF MODIFICATION OF A DECISION TREE RATING
SCALE
USED FOR MENTAL WORKLOAD ESTIMATION IN A COMMUNICATIONS TASK

by

Julie Hamilton Skipper

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

in

Industrial Engineering and Operations Research

APPROVED:

Dr. W. W. Wierwille, Chairman

Dr. J. G. Casali

Dr. H. L. Snyder

August, 1983
Blacksburg, Virginia

THE EFFECTS OF MODIFICATION OF A DECISION TREE
RATING SCALE USED FOR MENTAL WORKLOAD ESTIMATION
IN A COMMUNICATIONS TASK

by

Julie Hamilton Skipper

(ABSTRACT)

Of the subjective rating scales utilized in workload research the Cooper-Harper scale is the most popular. This decision tree rating scale and modifications of the scale have provided sensitive measurements by discriminating among various mental workload levels. This research is an investigation into the characteristics of a Modified Cooper-Harper rating scale (MCH) developed by Wierwille and Casali (1983). Six rating scale designs emphasizing major characteristics which might cause the MCH scale to be a sensitive measure of mental workload were used in this study. The aims of the research were to discover what modifications of the MCH might make it more sensitive.

A communications task developed by Casali and Wierwille (1983) was manipulated to present 36 subject pilots, both

private and student, with three communications loading levels. The pilots were distributed into the six rating scales by experience level. Six different experience levels were represented in each of the rating scale groupings. Using the communications loading, the presence of a decision tree in the scales appeared to improve the scale's ability to discriminate among loading levels. The expansion of the MCH scale to 15 categories decreased the sensitivity of the MCH rating scale. The standard 10-point MCH rating scale was the most consistent of the six rating scales and attained a high ability to discriminate among loading levels.

Finally, a companion mediational task study by Rieger (1983) using the same six rating scales resulted in substantially different results, suggesting that subjective rating scales are extremely task dependent.

ACKNOWLEDGEMENTS

The sponsors of this research were NASA Ames Research Center, Moffett Field, California, and Virginia Polytechnic Institute and State University, Blacksburg, Virginia. Thanks are extended to Mrs. Sandra Hart, grant monitor.

Special thanks are extended to Dr. Walter W. Wierwille for his patience and guidance throughout all stages of this research. Without his valuable advice and encouragement the completion of the research would not have been possible. I would also like to thank my committee members for their helpful comments in conducting this research.

I would also like to extend my appreciation to Christine A. Rieger for her invaluable assistance in conducting this research and her ever present support and encouragement.

Finally, I thank my husband, Lee R. Skipper, for his patience, understanding, and support throughout the last year and a half.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
	<u>page</u>
INTRODUCTION	1
SUBJECTIVE RATING SCALES AND THEIR DEVELOPMENT	4
Rating Scales	6
Ranking Methods	11
Attitude Scales and Scaling Techniques	12
Semantic Differential Scale	15
Other Scales	16
Multidimensional scaling	17
Magnitude scaling	18
RATING SCALES IN MENTAL WORKLOAD ESTIMATION	20
Mental Workload	20
Cooper-Harper Rating Scale	22
Decision Trees and Their Logic	28
Conjoint Scaling Methods	29
Semantic Differential	35
Other Scales	36
RESEARCH OBJECTIVES	39
Objectives	39
METHOD	42
Experimental Design	42
Apparatus	44
Flight simulator	44
Microcomputer	46
Rating Scale Designs	46
Communications Task	49
Primary Task Measures	52
Errors of omission	52
Errors of commission	52

Reaction time	52
Subjects	53
Procedure	54
Pre-testing	54
Communications task	55
RESULTS	58
Sensitivity Analyses	58
Conversion to z-Scores	59
Overall Sensitivity ANOVA	59
Individual Sensitivity ANOVAs	60
Duncan's Multiple Comparison Tests	62
Additional Information	68
Primary Task Measures	68
Pilot Experience Level Analyses	72
Load Presentation Order Analysis	77
Comparison of MCH Ratings with Previous Experiment's Ratings	79
Questionnaire Data Presentation	80
DISCUSSION	86
Sensitivity of the Individual Rating Scales	86
Additional Sensitivity Information	88
Primary Task Measures	89
Experience Level and Load Presentation Order	89
Comparison of the Two MCH Data Sets	90
Questionnaire Responses	91
Comparison of the Communications Results with a Mediational Study	92
SUMMARY	97
REFERENCES	100

Appendix

	<u>page</u>
A. MULTI-DESCRIPTOR RATING SCALE	111
B. PHOTOGRAPHS OF EXPERIMENTAL EQUIPMENT AND STATIONS	119
C. RATING SCALE DESIGNS	124
D. INTRODUCTION TO THE WORKLOAD EXPERIMENT	135

Participant's Informed Consent	139
E. RATING SCALE INSTRUCTIONAL SETS	141
MCH Rating Scale Instructions	142
Overview	142
Important definitions	142
Rating scale steps	144
How you should think of the rating	146
COMPMCH Rating Scale Instructions	148
Overview	148
Important definitions	148
Rating scale steps	150
How you should think of the rating	153
MCH+ Rating Scale Instructions	155
Overview	155
Important definitions	155
Rating scale steps	157
How you should think of the rating	159
COMPMCH+ Rating Scale Instructions	161
Overview	161
Important definitions	161
Rating scale steps	163
How you should think of the rating	166
NDT Rating Scale Instructions	168
Overview	168
Important definitions	168
Rating scale steps	170
How you should think of the rating	171
PBMCH Rating Scale Instructions	173
Overview	173
Important definitions	173

Rating scale steps	175
How you should think of the rating	177
F. COMMUNICATIONS TASK INSTRUCTIONS	179
Communications	180
Communications Task Instructions	184
G. QUESTIONNAIRE	189
Questionnaire	190
H. Z-SCORE TRANSFORMATION EQUATIONS FOR MEAN SCORES .	191
I. UNEQUAL-N ANOVA SUMMARY TABLE FOR SENSITIVITY ANALYSIS	193
VITA	195

LIST OF TABLES

<u>Table</u>	<u>page</u>
1. ANOVA Summary Table for the Overall Experimental Design	61
2. Summary Tables for the Individual Rating Scales . .	63
3. Duncan's Tests for the Statistically Significant Rating Scales	64
4. Duncan's Tests with a Less Conservative Testing Criterion	69
5. MANOVA Summary Table for RS Performance Effects . .	71
6. MANOVA Summary Table for EL Performance Effects . .	73
7. ANOVA Summary Table for Experience Level Effects . .	75
8. ANOVA Summary Table for Load Presentation Order Effects	78
9. ANOVA Summary Table for Comparison of the MCH Data Sets	81
10. Questionnaire Negative Responses	84
11. Questionnaire Positive Responses	85
12. Z-score Transformation Equations for Mean Scores .	192
13. Unequal-N ANOVA Summary Table for Sensitivity Analysis	194

LIST OF FIGURES

<u>Figure</u>	<u>page</u>
1. Cooper-Harper (1969) Rating Scale	23
2. Wierwille and Casali (1983) Modified Cooper-Harper Rating Scale	26
3. Wolf's (1977) Modified Cooper-Harper Rating Scale .	27
4. MOAT: Helm and Donnell (1978) Conjoint Measurement Rating Scale	32
5. WCI/TE: Donnell (1979) Conjoint Measurement Rating Scale	33
6. Smit and Wewerinke (1978) Semantic Differential Rating Scale	37
7. Experimental Design Layout Model	43
8. Mean Z-scores and Duncan's Test for the MCH Scale .	65
9. Mean Z-scores and Duncan's Test for the COMPMCH Scale	66
10. Mean Z-scores and Duncan's Test for the PBMCH Scale	67
11. Plot of Mean Values for Experience Levels	76
12. Plotted Means for MCH Data Sets	82
13. Casali and Wierwille Multi-Descriptor Rating Scale	117
14. Simulator Cockpit--Instrumentation and Labels . .	120
15. Experimental Station for the Communications Flights	121
16. Photograph of a CRT Screen for the COMPMCH Rating Scale	122

17.	Photograph of a Pilot Making a Typical Computer Rating	123
18.	Modified Cooper-Harper (MCH) Rating Scale	125
19.	Computerized Modified Cooper-Harper Rating Scale .	126
20.	15 Point Modified Cooper-Harper (MCH+) Rating Scale	129
21.	Computerized 15 Point Modified Cooper-Harper (COMPMCH+) Rating Scale	130
22.	Tabular Modified Cooper-Harper (NDT) Rating Scale	133
23.	Performance Based Modified Cooper-Harper (PBMCH) Rating Scale	134

INTRODUCTION

In science and technology, the human is becoming an increasingly important link in a system. As systems become more complex, knowledge concerning human capacities and abilities to interact with specific systems is vital to system design. Measures concerning maximum and minimum thresholds; sensory, muscular, and perceptual strengths and weaknesses; and mental workload capabilities are of critical importance. One industry in which this is increasingly apparent is the aircraft industry, in the design of new aircraft. In fact, the Code of Federal Regulations for the Federal Aviation Administration (FAA) requires that for safe aircraft operation the minimum flight crew must be established - considering the workload of individual crew members (Parts 1 to 59, 1982); yet, the term "workload" has no formal definition in the field of workload estimation.

To obtain information concerning workload four major categories of mental workload measurement have been defined. They are physiological, spare mental capacity, primary task, and subjective opinion measures (Williges and Wierwille, 1979). Of primary importance in this report are the

subjective opinion measures and the psychometric and/or analytic scales which have been developed to measure subjective opinion.

Subjective opinion measures elicit an individual's opinion of the workload he or she experiences in performing a task by either direct or indirect queries. Thus, subjective opinion measures are extremely relevant in obtaining information from the operator regarding the system-imposed mental workload.

The primary barrier in obtaining an adequate or accurate measure of mental workload is that it is multi-dimensional and it is driven by the psychological and physiological attributes of a specific task. One means of obtaining a measure involving several dimensions is through integration or summation of the task effects. Of the four categories of mental workload measurement, subjective opinion measures may provide the best means of integrating the experience of mental workload into a single relevant measure (Gartner and Murphy, 1976). Some physiological, primary task, and spare mental capacity measures can provide meaningful measures, but none of them contains the degree of integration performed by the human rater. However, it is suggested that the physiological, primary task, and spare mental capacity measures be combined with the subjective

ratings to obtain additional information on mental workload (Childress, Hart, and Bortolussi, 1982; Eggemeier, 1980).

Due to the success and support of the subjective opinion rating scales, this research is an investigation into the development and establishment of subjective opinion and attitude scales and their promising characteristics in evaluating mental workload. In reviewing subjective measurement techniques, psychological scaling techniques will be introduced first. Then, techniques specifically used in mental workload research will be described.

SUBJECTIVE RATING SCALES AND THEIR DEVELOPMENT

Subjective measurements are usually applicable to situations in which there are no known corresponding physical measurements. In fact, one of the main requirements of a truly subjective metric is that it should be entirely independent of all physical measurement (Thurstone, 1954). As a result, experimentation can be conducted using subjective measures with aesthetic objects and other types of stimuli with which no physical measurements are associated. Areas in which subjective measures are highly applicable include social attitudes, moral issues, marketing research, economics, and the measurement of values.

The principles upon which subjective measures have been established are psychophysical concepts, including methods of equal appearing intervals, rank ordering, paired comparisons, magnitude estimation, and triads. When using the rating scales developed through psychophysical concepts, it is assumed that the rater is able to discriminate between or among stimuli and that the discrimination is in the same general region of the subjective continuum for the same

stimuli (Thurstone, 1954). This is important when asking a human subject to act as a measuring instrument by eliciting his or her opinion. The data which are obtained by subjective opinion are then analyzed to determine the properties of the responses. "In general, psychological scaling is a formalization of the behavioral responses representing a psychological magnitude of interest. As such, it represents the realm in the measurement process of establishing relation between objects" (Helm, 1981, p. 5).

Various psychometric methods have been developed for eliciting subjective opinions. Of main importance in this review are psychological rating scales and attitude measures which could be used for the estimation of mental workload. Of the possible psychometric rating scales available--numerical or discrete scales, graphic scales, standard scales, checklists, forced choice scales, ranking methods, and q-sort or card sorting--only discrete scales, graphic scales, and ranking methods will be reviewed in detail, due to their present application in mental workload research (Dyer, Matthews, Wright, and Yudowitch, 1976; Helmstadter, 1964). The attitude measures discussed will be Thurstonian scaling, Likert scaling, and Guttman scaling. Other psychometric measures presented are semantic differential scaling, Behaviorally Anchored Rating Scales

(BARS), multidimensional scaling, and magnitude scaling. Conjoint measurement, a psychometric technique currently being used in mental workload, will be reviewed in detail in the section on mental workload scales. It should be noted that little evidence of rating scale or attitude scale reliability and validity has been documented. But, rating scales and attitude scales are still used widely to obtain subjective opinions, regardless of their statistical deficiencies. Requirements needed for the scales to be reliable and valid are presented in Shaw and Wright (1976).

Rating Scales

In general, rating scales introduce a systematic approach for obtaining subjective opinions and provide numerical values for analysis. They are easy to implement, inexpensive, not intrusive, and have proven to be valuable instruments giving satisfactory research results. The effectiveness of a scale lies in the scale construction and the behavior being rated. To be meaningful, the scale and its descriptors must contain a means of driving the subject to a summation of his or her behavioral attributes. If designed properly, rating scales have been found to be more reliable than the paired comparisons and the magnitude estimation methods (Dyer et al., 1976). For example,

support for the use of graphic scales as opposed to the psychometric technique of magnitude estimation can be found in Weiss (1980).

However, rating scales also have their disadvantages. They are more vulnerable to biases and errors than are other psychometric measures. In using rating scales, two types of errors can arise, errors resulting from rater characteristics and errors resulting from the behavior being rated. The most common of these errors are halo error, leniency error, and central tendency error (Korman, 1977). Adequate treatment of these problems through proper selection of behaviors, proper scale building, selection and training of the raters, proper administration of scales, and proper analysis and use of the results can reduce some of the possible inherent error of the rating scale (Dyer et al., 1976; Helmstadter, 1964).

There are four general levels of rating scale measurement: nominal, ordinal, interval, and ratio. All of these scales provide a direct means of obtaining subjective opinion measurements. Their names reflect the type of data each of the scales obtains. When designing rating scales extreme care must be exercised in constructing the interval and ratio relationships and the general scale descriptors. Rating scales are only as effective as their initial design

characteristics and relationships (Kuznicki and Nagle, 1980; Shaw and Wright, 1967; Spector, 1980). Considerations for scale design are presented in this section. They include the choice of anchors or descriptors, scale graphics, and the number of levels or absolute judgements.

To classify scales further, they may be grouped according to their anchors or category notation. The classifications are (1) non-adjectival (numeric-alphabetic), (2) descriptive adjective, and (3) behavior-sample (descriptions of behavior) (Stone and Helm, 1981). Each of the classifications describes the type of anchors which are used to define the points on the rating scale. These descriptors need to be short, concise, and unambiguous. The cues which are associated with the individual scale points should avoid terminology which may connote a different meaning to different raters. Since rating scales have a variety of anchors, they may be made and used for almost anything--people, groups, objects, perceptions, and systems. Also, using the various types of anchors, discrete and continuous rating scales may be constructed as "numerical" scales--the descriptive cues are presented as discrete categories; and graphic rating scales--the descriptors are associated with points on a line or graph (Dyer et al., 1976).

When constructing graphical scales, the line or graph can be presented either vertically or horizontally and the associated numbers of the descriptors on the scale may run from low to high or from high to low, without causing statistically significant differences in the responses of the rater (Edvardsson, 1980). Graphic scales offer one advantage to rating that the numerical scales do not, namely, that the rater is not restricted to a categorical judgement. The rater has flexibility in arriving at a rating. A major drawback to the graphic rating scale is that the continuous nature of the scale makes it more difficult to score than other scaling methods. The graphic scale tends to imply a degree of precision or accuracy which is unwarranted (Dyer et al., 1976).

Another important consideration in the development of rating scales is the number of discrete levels or categories of the scale. The main objective here is to design a scale to achieve the desired level of discrimination among factors or stimuli. If a scale is designed with too few levels, the scale will be coarse and a great deal of the rater's discriminative power will be lost. Conversely, a scale with a large number of levels can be too fine, that is, the levels of discrimination are beyond the rater's capabilities (Guilford, 1954; Kuznicki and Nagle, 1980). It should be

kept in mind that if the raters are cooperative and willing, finer discrimination is appropriate (Guilford, 1954). In previous literature regarding the number of scale levels to use, there are discrepancies (Guilford, 1954; Kuznicki and Nagle, 1980; Norwich, 1981). The optimal number of levels ranges from 7 to 25 (Guilford, 1954).

To understand better the range of scale levels suggested in the literature, Norwich offers a solution to the dilemma. Norwich, using Miller's (1956) "seven, plus or minus two" theory, illustrates that the number of absolute judgements possible is directly related to "an entropy function, H_s , which is a measure of the uncertainty of a subject (or a sensory receptor) as he or she perceives the magnitude of an applied stimulus" (Norwich, 1981, p. 409). Norwich found through his function that the range of values in making absolute judgements is between 2-4 bits per stimulus, while the order of magnitude per stimulus is about 3 bits. These values were found to be constant across paired stimuli, absolute judgements, and psychophysical laws--Weber-Fechner Law, Steven's Law, and the Bunsen-Roscoe Law (Norwich, 1981). Since using seven categories in making absolute judgements contains approximately 2.5 bits, it seems reasonable to accept the concept of "seven, plus or minus two", as originally hypothesized by Miller (1956).

Additional support for the use of five to nine categories is evident in Hallsten and Borg (1975). Hallsten and Borg concluded in using 5, 7, 9, and 11 levels on numerical and graphical scales, when rating perceived difficulty, that 9 levels resulted in the greatest statistical sensitivity. They also reported that the rater's preference was to use scales with five or seven categories, which is consistent with a rater's ability to discriminate.

An important aspect to keep in mind when choosing the number of category levels is the type of rating scale design. For example, in the mental workload section the Cooper-Harper rating scale will be introduced. This rating scale has a decision tree structure. The number of categories to use in a decision tree rating scale may not agree with the number of categories which have been successful in a numerical or graphic rating scale. Additional research in this area could reveal important findings.

Ranking Methods

Because ranking has been used in some of the mental workload estimation techniques, a brief review of the advantages and disadvantages of ranking is in order. As an

overall technique, ranking is generally less precise than are rating scales, but is a comparable method (Dyer et al., 1976). Ranking takes little time to administer and is a familiar technique to the raters. The rater merely needs to order objects; no judgement on the size of the intervals is required. The responses obtained from ranking are typically easier to score and code than classical paired comparisons. Disadvantages in the rank order method include the lack of relative differences between objects, the questionable nature of results unless the basis of ranking is established clearly, the difficulty of rank ordering a large number of alternatives, and slight ambiguity in ranking which can only be eliminated by further questioning (Dyer et al., 1976).

Attitude Scales and Scaling Techniques

In many instances in research, it is important to obtain an understanding of the individual's attitude or feeling about a situation or subject. Attitude scales are used to reveal the individual's attitude and are composed of a number of questions or statements which together access the rater's attitude. Typically the questions or statements used in an attitude scale are from a larger pool of questions and statements.

In attitude scaling techniques, there are two general methods for constructing attitude scales (Dyer et al., 1976). The first method uses a psychological scaling method and a judging group. The resulting scale contains a set of statements which are assigned values on a psychological continuum. This method can be used to scale statements of objects as well as attitudes. The second method is based on direct responses of agreement or disagreement to specific attitude statements. The results are discrete and are not assigned scale values on a psychological continuum.

Three scaling methods for use in obtaining attitude scales were developed by Thurstone (Dyer et al., 1976). These methods are the method of equal-appearing intervals, the method of paired comparisons, and the method of successive intervals. In developing the method of equal-appearing intervals, Thurstone introduced the first major method of attitude scaling. Using a large number of statements of opinion, judges sort the statements into 11 separate piles. It is assumed that the statements can be ordered on a continuum of "favorableness" to "unfavorableness", and that the ordering reflects equal distances between adjacent statements. After ordering, the statements are reduced to 20 or 25 statements in each pile and scale values are assigned. The statements which remain

appear to have approximately equal or relatively spaced values on the psychological continuum.

The method of paired comparisons, on the other hand, develops a scale using the law of comparative judgements. This technique creates an interval scale by rank ordering all possible object pairs. In the applied sciences this method is not widely used because of the large number of comparisons needed if there are a moderate number of objects (e.g., if there are 20 objects, 190 comparisons would be required).

Similiar to the method of paired comparisons is the method of successive intervals. The method of successive intervals rank orders the objects as in paired comparisons, but no assumption is made concerning the psychological equality of the resulting categories or levels. Order is the only construct that is established. Scale values are obtained by statistically estimating the category widths between adjacently ranked categories. Research has shown that there is a linear relationship between attitude scales developed by paired comparisons and successive intervals (Dyer et al., 1976).

The last two attitude scaling methods, Likert scaling and Guttman scaling, arrive at discrete attitude scales and use the second scaling method described above. The Likert

scale is often called the method of summated ratings. The assumption in this technique is that all statements reflect the same attitude dimension and so they are all related. The Guttman scale approach is called a scalogram. A set of statements is used to reflect a unidimensional scale with a cumulative property. When a series of statements is properly scored, the scale is a deterministic model and can be used as a measure of attitude. Thus, if two people obtain equivalent scores, they have responded the same to all previous statements.

Semantic Differential Scale

Originally the semantic differential scale was developed as an attitude scaling technique. It is now primarily used to measure attitudes and values. Osgood and Suci (1955) presented the semantic differential scale as a seven-point, bipolar scale. The scale form is typically a descriptive-adjective scale, with adjectives or adverbs defining the scale end points. The scale can be expanded to include any number of descriptors, and 5, 7, 9, 11, or 13 adjectival anchors (Stone and Helm, 1981).

By using adjectives as anchors, sometimes the adjectival description may not seem relevant, but it may reflect a feeling which the subject could not put into

words. An advantage such as this is what makes the semantic differential scale so popular in research. There is evidence that when properly designed the semantic differential scale has "validity and reliability", and it produces a stastically sensitive attitude measure. (Dyer et al., 1976; Shaw and Wright, 1967). Disadvantages of the scale are directly related to the scale's construction. The descriptors need to be chosen with extreme care. Any ambiguity will result in response error. Further details concerning the semantic differential scale are reviewed in the section dealing with mental workload scales.

Other Scales

Some of the new scaling methods which may add to the psychological rating scale realm are multidimensional scaling, magnitude scaling, conjoint measurement, and Behaviorally Anchored Rating Scales (BARS). The multidimensional and magnitude scaling methods are presented in summary below, while the conjoint measurement technique are presented in detail in the next section. As for BARS, this method was developed to offer superior and more accurate measures by adding resistance to rating errors (Smith and Kendall, 1963). As yet, the scale has not produced satisfactory results and there is a great deal of

controversy in the literature regarding the successful use of BARS (Hom, DeNisi, Kinicki, and Bannister, 1982). Thus, to include BARS in this research would be premature. The scale needs to be developed further to reduce its deficiencies.

Multidimensional scaling. A technique which has provided a means of evaluating perceptual data is the multidimensional scaling method (MDS) (Kruskal and Wish, 1978). A major advantage of this scaling technique is that the subject's rating task is unstructured. After assembling a list of objects, the rater rates the objects in a series of pairwise comparisons determining the differences within pairs until all possible pairs have been exhausted. With MDS the subjects are not told what characteristics of the objects to attend to in making their difference judgements. Instead, the underlying characteristics are inferred from the MDS results (Billings and Cornelius, 1980; Derrick, 1981).

Once the objects have been compared, the perceptual judgements given to the set of paired objects are analyzed by computer. The software program (e.g., INDSCAL or ALSCAL) contains a set of analytic models (Kruskal and Wish, 1978). The analytic models uncover a latent spatial structure that is inherent in the data set. The spatial structure is

typically a low dimension Euclidean space (Kruskal and Wish, 1978). From the analysis, either an order representation or an actual numerical evaluation can result. When a numerical evaluation results, it is a set of distances--"psychological distances"--between objects in the Euclidean space. Thus, the objects are all relative to one another by "psychological distances" in the spatial structure. As mentioned previously, an attractive advantage of MDS is its ability to create its own dimensions and attributes of the objects being compared. That is, the resulting Euclidean dimensions are a direct result of the rater's perceptual judgements (Billings and Cornelius, 1980; Derrick, 1981; Kruskal and Wish, 1978).

Magnitude scaling. Magnitude scaling is an advanced and improved method of category scaling, which is especially used in the social and behavioral sciences (Lodge, 1981). It is used to determine the direction and strength of people's beliefs and preferences. It is believed by Lodge that social scales should behave as lawfully as do scales of physical sensations; thus, a magnitude scale is aimed at being a social psychophysical scale. In magnitude scaling, as in MDS, objects in a set are compared. But the objects in magnitude scaling are all compared to a standard object. Thus, a set of perceptual judgements is obtained and analyzed.

After all objects have been compared to the standard by the group of subjects, each of the comparison object's values are averaged over the group of raters. By averaging over a number of subjects, the data are smoothed and a basis for analyzing the results is provided. Ideally, by presenting the stimulus several times and in different orders the test-retest reliability of the test may be increased. Next, "numeric estimates" are calculated from the group (see Lodge, 1981, for "numeric estimates"). By plotting the "numeric estimates", it can be determined if a psychophysical relationship is present in the data. The magnitude of the judgements is proportional to the magnitude of the objects. Both magnitude scaling and multidimensional scaling have promise, but as the number of objects increases so do the time and cost of the experiment. This may limit the use of these techniques in mental workload research due to the large number of dimensional attributes of mental workload.

RATING SCALES IN MENTAL WORKLOAD ESTIMATION

Mental Workload

Depending on the discipline of study of the workload investigator, the definition of mental workload varies. The various methodologies with which workload is investigated and their respective definitions of workload are presented in Moray (1979). In general, workload is considered to be a multidimensional theoretical construct (Williges and Wierwille, 1979).

Of the main workload measurement categories (physiological, primary task, spare mental capacity, and subjective opinion), subjective opinion measures have displayed consistent sensitivity to various workload conditions. It should be noted that sensitivity as related to workload refers to the scale's ability to statistically discriminate among various workload levels. Subjective opinion measure sensitivity is evident in papers by Bird (1981), Casali and Wierwille (1982; 1983), Connor and Wierwille (1981), Hart, Childress, and Bortolussi (1981), Helm (1981), Eggemeier, Crabtree, Zingg, Reid, and Shingledecker (1982), and Moray (1982), as well as in others. As a result of the subjective opinion scale's

sensitivity in evaluating mental workload, Ruggiero and Fadden (1982) comment that the use of subjective ratings should be central to any investigation of workload. By using subjective opinion scales along with other workload measures, better interpretation of all the workload data can be obtained by universally examining the results. A review of some current subjective measurement techniques and their relationship to physical, cognitive, manual-control, and time-stress dimensions as well as operator subjective load experience can be found in the paper by Moray (1982). Moray also discusses the identification of factors or variables which contribute to the subjective experience of mental workload for use in subjective rating scales.

Several methods of subjective opinion rating have been employed in measuring mental workload. Among these techniques are rating scales, questionnaires, and interviews. In this report, only rating scales will be discussed in detail due to their attractiveness in research and their abundant use in mental workload studies. Advantages which make rating scales attractive in mental workload research include wide-spread acceptability, ease of administration, low cost, lack of intrusion on the primary task, bolstering of the subject's ego, and little, if any, delay before administration. These advantages will become

more apparent through the following discussion of the rating scale techniques.

Cooper-Harper Rating Scale

Subjective rating scales for the estimation of mental workload evolved from a series of rating scales developed for the assessment of aircraft handling qualities beginning in the early 1930s (VanCott and Kinkade, 1972). The first scales were designed for specific purposes and a standard scaling method was not in existence. Through the efforts of Cooper and Harper (1969), a reasonably sensitive scale for the evaluation of aircraft handling qualities was developed (Figure 1). The Cooper-Harper rating scale is the most widely used of all the subjective rating scales (Helm and Heimstra, 1981) because of its sensitivity in discriminating among loading levels and because it is easily modified (Stone and Helm, 1981).

Characteristics which were originally incorporated into the Cooper-Harper rating scale design include (1) easily understood terminology for all individuals using the scale, (2) a decision tree logic of the rating categories to lead the rater in his or her decision process, (3) concise category definitions which emphasized the rater's decisions in selecting a final category rating, and (4) elimination of

HANDLING QUALITIES RATING SCALE

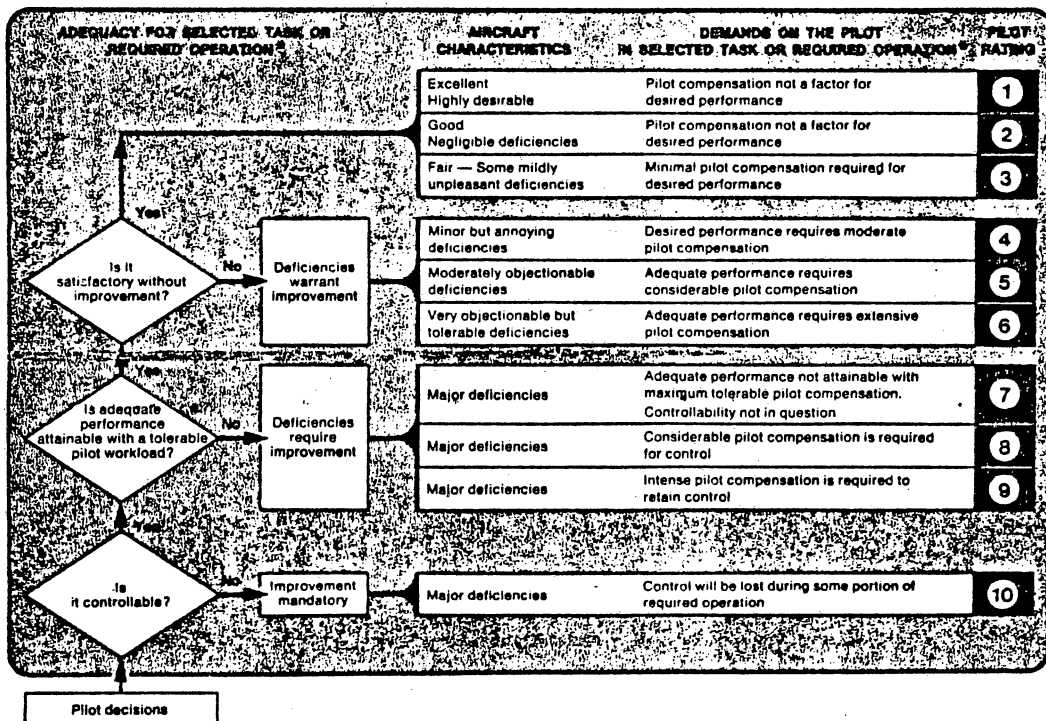


Figure 1: Cooper-Harper (1969) Rating Scale

repetitive descriptive terms (Cooper and Harper, 1969). The rating scale, in general, is an ordinal, 10 category, linear scale with numerical anchors which descend with progression through the scale logic. Each of the numerical ratings is directly dependent upon the category definitions associated with it. The definitions of each category are rated on the Cooper-Harper scale, not the numbers. As a result, it should be emphasized that no comparisons should be drawn between the rating numbers--the scale is an absolute ordinal scale, no mathematical relationships or conclusions can be derived. Another feature of the scale is that it has major subdivisions at the 9.5, 6.5, and 3.5 scale values.

Before administering the Cooper-Harper rating scale, a set of detailed instructions on making ratings is presented to the subject. It is extremely important that the subject thoroughly understand the terminology and logical rating procedures of the scale. The scale is typically administered immediately after the task is completed. This gives the rater the opportunity to evaluate the induced workload of the task while the task is in memory.

There are some inherent disadvantages in the scale design. The scale is limited because of its ordinal structure. After a rater completes several tasks, only relative conclusions can be drawn. Also, it is difficult to

evaluate the means by which the rater arrived at the rating. Did the rater truly follow the decision tree logic or was he or she searching for key words or numbers? Finally, the rating scale does not consider any of the components of mental workload, only the overall picture of mental workload.

As mentioned above, one of the great advantages of the Cooper-Harper rating scale is its ability to be modified. Once the rating scale is removed from aircraft assessment, it must be modified to reflect the qualities of the specific task at hand. Three modified rating scales for different applications are shown in Figures 2 and 3 (Wierwille and Casali, 1983; Wolf, 1977). The Wierwille and Casali scale (Figure 2) experimentally resulted in high sensitivity, but was unable to distinguish differences among all levels of loading. This modified scale is equivalent in format to the Cooper-Harper scale. The modifications lie in the primary decisions of the decision tree and the category definitions and descriptors. As long as major care is used in modifying the Cooper-Harper scale, it can be used for a variety of tasks. The appeal of the Cooper-Harper scale has been greatly enhanced as a result of its versatility.

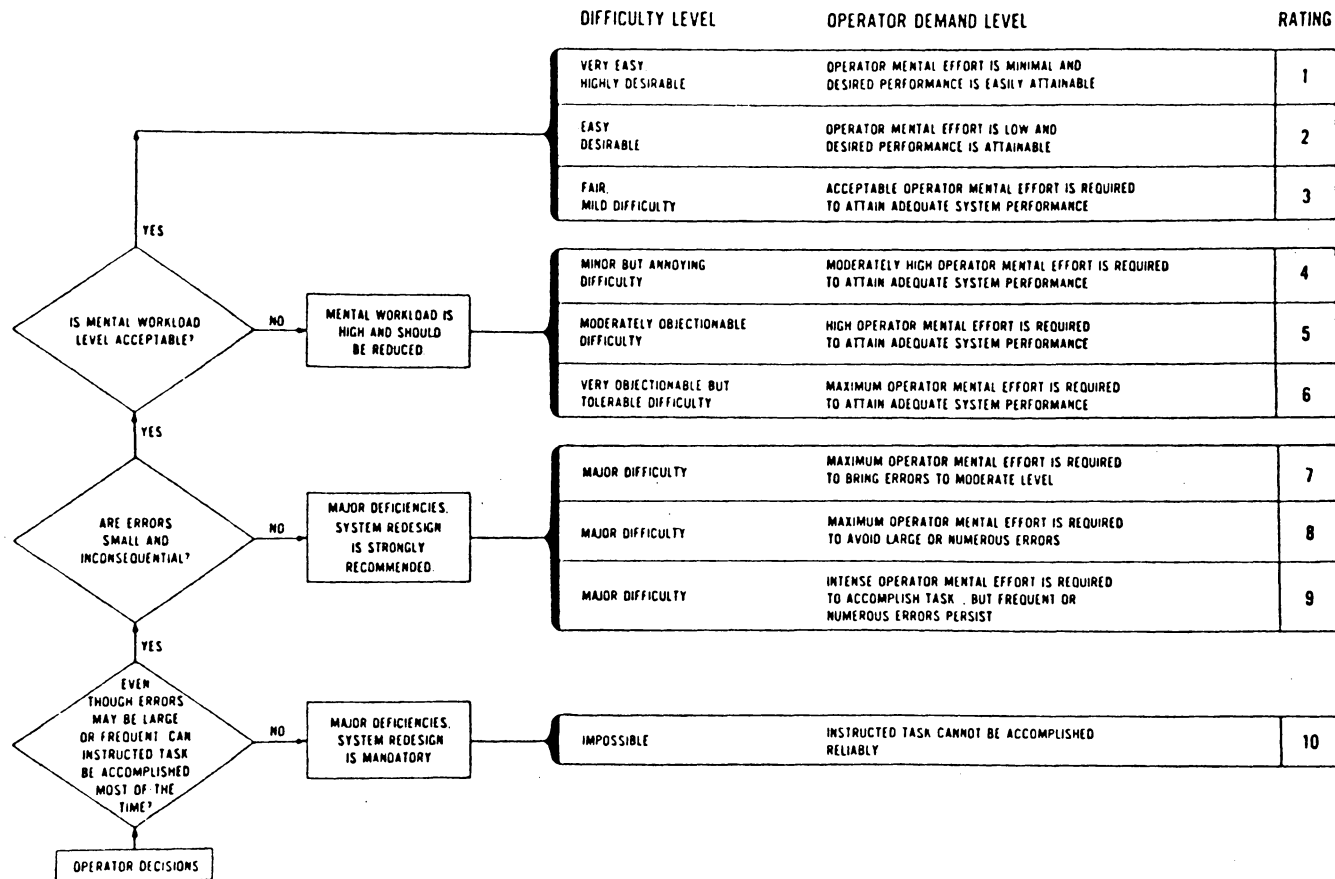


Figure 2: Wierwille and Casali (1983) Modified Cooper-Harper Rating Scale (56% reduction)

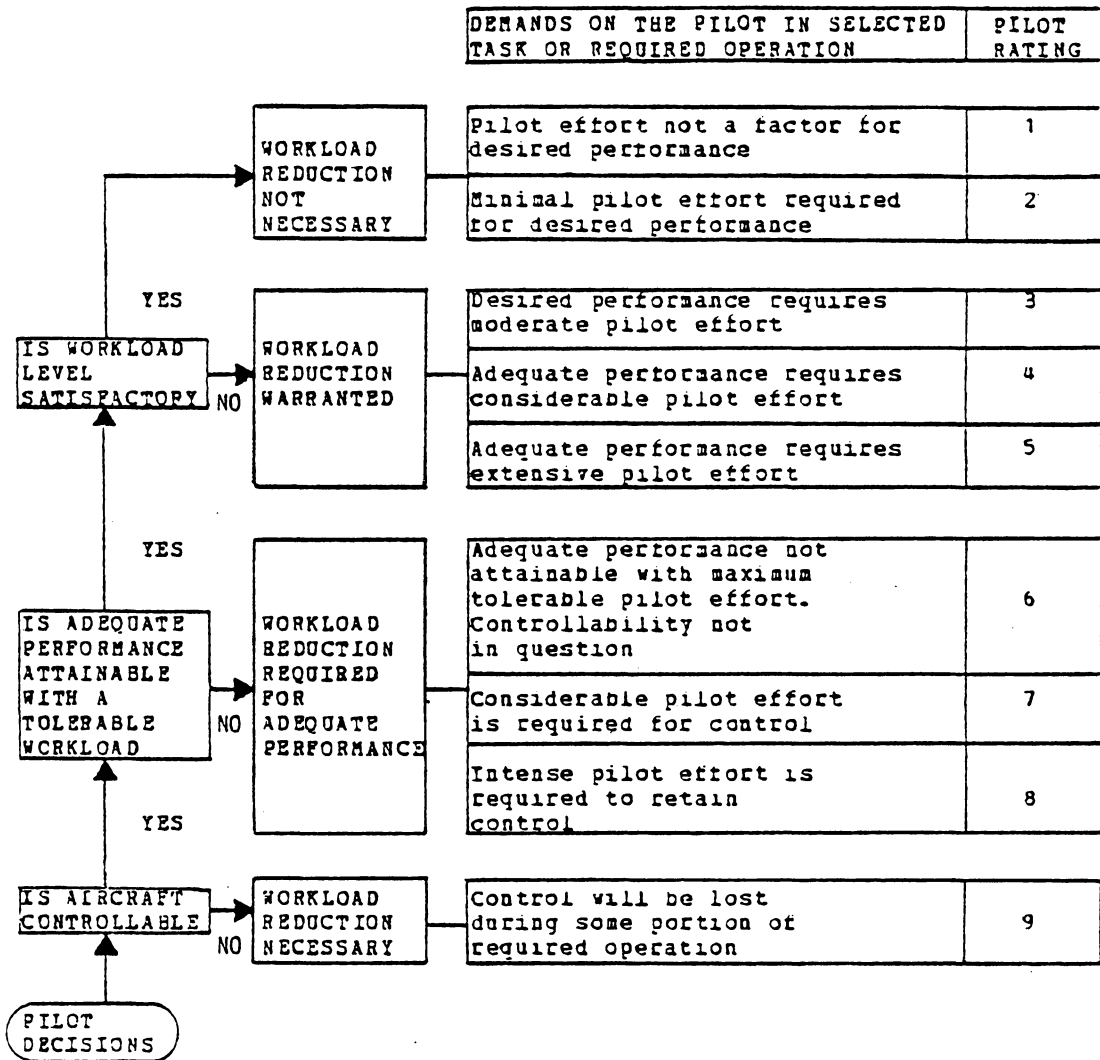


Figure 3: Wolf's (1977) Modified Cooper-Harper Rating Scale

Decision Trees and Their Logic

Some additional information which aids in the support of the Cooper-Harper scale in measuring the multidimensional concept of mental workload has to do with the constructional aspects of the scale. These include the decision tree aspect with a sequential decision process and the flow diagram incorporated in the design of the rating scale.

First, the decision tree aspect of the scale parallels the use of decision trees in many other disciplines which specifically use decision trees to simplify complex, multidimensional decisions. Decision trees have been resourceful in fault tree analysis (Brown, 1976), management science decision theory (Anderson, Sweeney, and Williams, 1979), Vroom's leadership model (Gilmer and Deci, 1977), and capital investment decision analysis (Canada and White, 1980). As stated by Canada and White (1980, p. 370), "Decision trees have great usefulness in practice because they make it possible to look at a large complicated problem in terms of a series of smaller simple problems and they enable objective analysis and decision making which includes explicit consideration...."

In the Cooper-Harper scale the decision tree and flow diagram allow the evaluation of workload to be broken into stages. This aids the rater in making a decision and also

breaks down the number of levels used in making an absolute judgement. The overall scale has 10 levels, but at each decision point the number of levels is 2 or 3. Thus, it seems possible that by adding additional levels at each stage of the decision tree a finer discriminability could be obtained because the number of absolute judgements is raised towards seven, plus or minus two.

In the support of the flow diagram, Mallamad, Levine, and Fleishman (1980) revealed that the use of decision flow diagrams employed in conjunction with rating scales improved the overall subjective evaluation of personnel job requirements. Thus, the use of a flow diagram along with a rating scale has merit in other multidimensional rating areas.

Conjoint Scaling Methods

Some of the more recent subjective rating scales which are used in the assessment of handling qualities in aircraft use the concept of additive conjoint measurement (Krantz, Luce, Suppes, and Tversky, 1971; Krantz and Tversky, 1971; Luce and Tukey, 1964). What makes conjoint measurement a promising and appealing technique for the estimation of mental workload is that the method converts ordinal data to interval data. If it is possible to order the dependent

variables and define properties which that order satisfies, then the dependent variables can be expressed numerically according to the proposed composite rules. Composite rules describe the functional relation among the dependent variables that are true for all the variables. In the case of mental workload, the composite rules are additivity based on the qualitative relationships of the attributes of mental workload (Reid, Shingledecker, and Eggemeier, 1981). If properly used, the conjoint measurement technique has promise in mental workload estimation.

Three conjoint measurement techniques which have been developed for the assessment of aircraft are Mission Operability Assessment Technique (MOAT)(Helm and Donnell, 1978), workload/compensation/interference and technical effectiveness (WCI/TE) (Donnell, 1979), and the Subjective Workload Assessment Technique (SWAT) (Reid et al., 1981 Reid, Shingledecker, Nygren, and Eggemeier, 1981; and Reid, Eggemeier, and Nygren, 1982). None of these scales have been experimentally validated as yet, but they have been successful in discriminating among workload levels in assessing aircraft workload.

Each of these scales represents a stage in the development of conjoint measurement, with MOAT (Figure 4) being the predecessor of the WCI/TE, and SWAT being a more

advanced stage of WCI/TE. The WCI/TE scale (Figure 5) is characterized by a matrix. One dimension is defined as technical effectiveness, while the second dimension is defined as pilot workload and combines workload, compensation, and interference descriptors. The SWAT scale, being a more advanced conjoint scale, adds a third dimension to its matrix. The three dimensions are (1) time load, (2) mental effort load, and (3) stress load. The object of the conjoint scales is to produce a joint effect called workload using the scales' dimensions.

To arrive at a joint effect, two steps or procedures need to be performed in conjoint measurement. First, a scale development phase is required, followed by the actual event scoring phase. In the scale development phase, raters rank order the cells of the matrix--the WCI/TE has 16 cells, while the SWAT has 27 cells. After the cells are ranked, the ratings between pilots are checked for correlation. If high correlation exists between the pilot ratings, one rating matrix is used for all the pilots' workload ratings. If low correlation is obtained, a number of rating matrices is used, one matrix for each of the respective correlation groups from the initial rank ordering procedure of the cells. Next, the ordinal data obtained by ranking are converted to interval data using the delta method (Coombs,

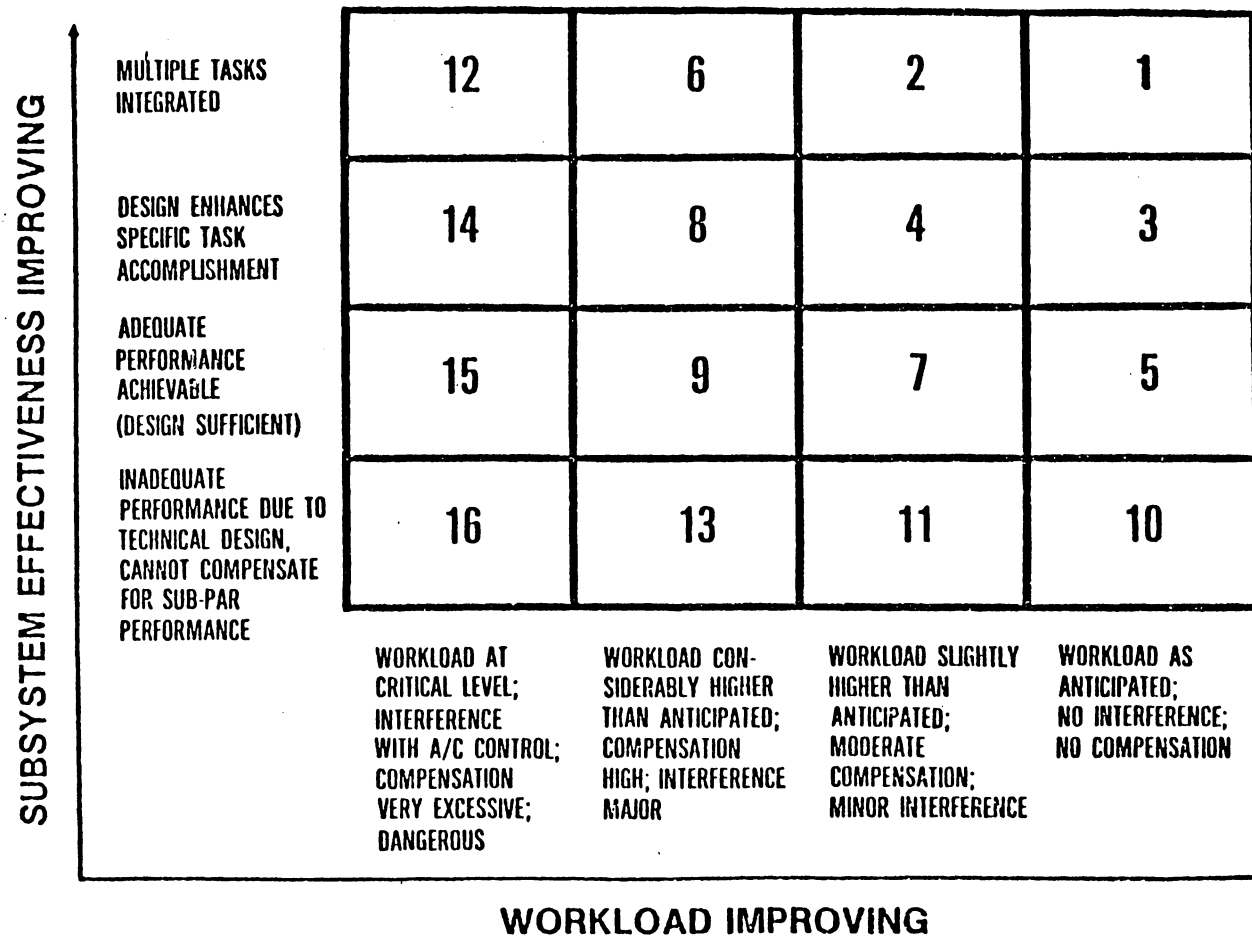


Figure 4: MOAT: Helm and Donnell (1978) Conjoint Measurement Rating Scale

		1	2	3	4	Column	Row
T E C H N I C A L E F F E C T I V E N E S S	Multiple Tasks Integrated						4
	Design Enhances Specific Task Accomplishment						3
	Adequate Performance Achievable; Design Sufficient to Specific Task						2
	Inadequate Performance Due to Technical Design						1
		Workload Extreme; Compensation Extreme; Interference Extreme	Workload High; Compensation High; Interference High	Workload Moderate; Compensation Moderate; Interference Moderate	Workload Low; Compensation Low; Interference Low		
WORKLOAD/COMPENSATION/INTERFERENCE (Mental and Physical)							

Figure 5: WCI/TE: Donnell (1979) Conjoint Measurement Rating Scale

1964). Having converted the matrix data to interval data, the interval data are rescaled from 1 to 100. The scale is now complete and ready for the event scoring phase to begin. Each pilot rates his or her subjective workload using the scale constructed above.

Reid et al., (1982) added an extension to the SWAT scaling procedure which accounts for individual differences in ratings. Instead of scaling the pilot's ratings all together, the pilot's ratings are divided into homogeneous subgroups based on the pilot's ranking of the major scale dimensions--time load, mental effort load, and stress load. The groups are then transformed into an overall final SWAT scale. This procedure was applied to the subjective workload assessment in a short-term memory task and was found to be highly sensitive (Eggemeier et al., 1982). The new approach added precision to the SWAT scaling procedure.

Some drawbacks regarding the conjoint measurement procedure have been pointed out by Boyd (1982). The concerns regard the rating procedure; the intersubject variability and dispersion of data; and the use of the Kendall's W statistic, a coarse statistic, in determining correlations of rankings in data with a general trend. Perhaps the objections by Boyd should be investigated further before the application of the conjoint measurement

technique is expanded to more dimensions as suggested by Reid et al., 1981.

Semantic Differential

Another rating scale technique which has successfully distinguished among various levels of workload is the semantic differential scales from psychology. A series of semantic differential, bipolar scales, each having a single specified dimension of mental workload, are presented to the rater. By presenting the rater with a number of dimensions (see scales in Appendix A, Figure 13), a multidimensional type measure called a multi-descriptor is obtained (Casali and Wierwille, 1982). The scale values can be used in two ways: (1) to average an overall rating (Casali and Wierwille, 1982; Rahimi and Wierwille, 1982) or (2) to use each rating value individually in the statistical analysis (Bird, 1981). Of these studies all three resulted in a monotonic increasing trend with increased loading levels and two (Bird, 1981; Casali and Wierwille, 1982) were sensitive to load.

One final subjective technique used by Smit and Wewerinke (1978) uses a combination of three semantic differential scales (Figure 6). The three scale dimensions used are effort, controllability and decision, and pilot

demands. The effort scale uses non-adjectival anchors, while the controllability and decision scale and the pilot demands scale use both numerical and adjectival anchors. The subjective ratings obtained discriminated well among the tasks (Smit and Wewerinke, 1978).

Other Scales

An additional group of studies in the subjective estimation of mental workload involve the comparison of ratings between two tasks. Derrick (1981) used the method of multidimensional scaling (MDS) to draw comparisons between tasks. The main effort in Derrick's study was to evaluate the rater's judgements and produce attributes which underlie the perception of workload. The dimensions used to describe Derrick's 18 tasks were input complexity, response complexity, effort demands, feedback adequacy, time demands, and task difficulty.

Major drawbacks in such comparison studies are due to the intersubject variability. Subjects have different perceptions and strategies in completing a task. To arrive at conclusive workload results using task comparisons, subjects would need to react and think the same way under the same conditions. This is not the case. One additional drawback to this technique is that as the number of objects

Name:

Task:

Using the scale below, indicate the degree of effort you spend on performing the task

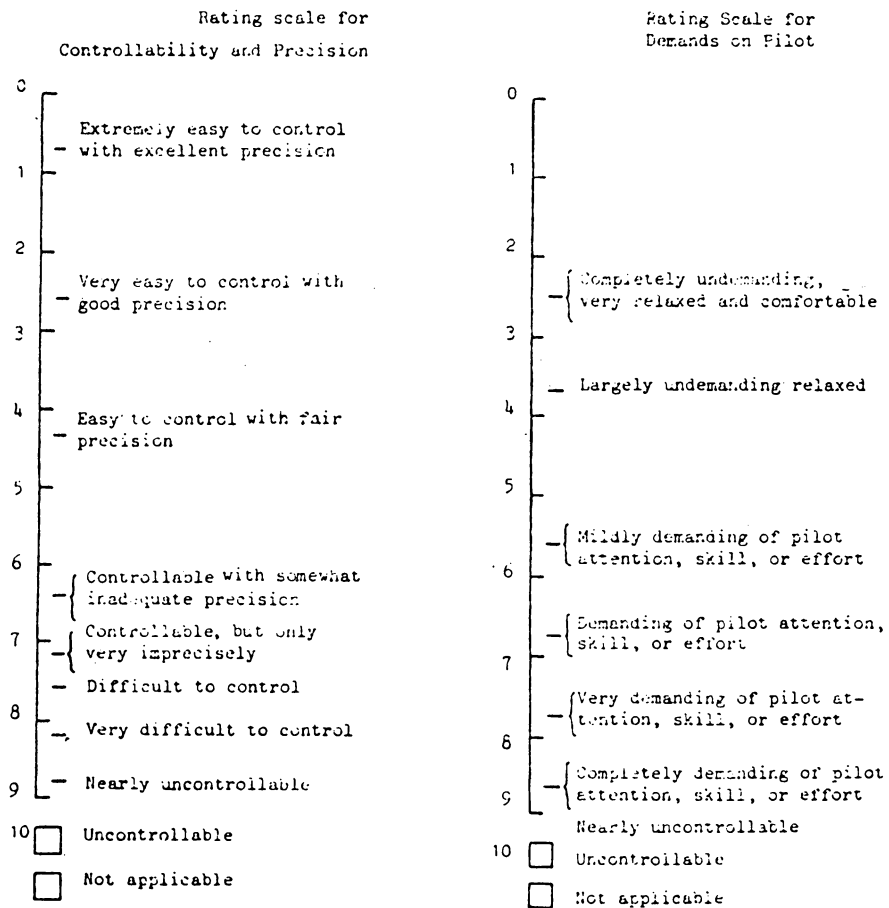
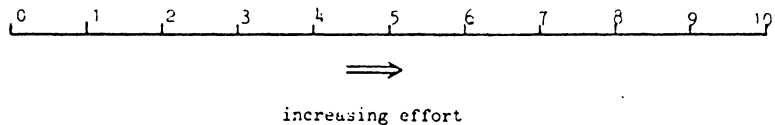


Figure 6: Smit and Wewerinke (1978) Semantic Differential Rating Scale

for comparison increases, the time and cost of the study increase proportionally.

Other scales which have been employed in mental workload measurement, but have little documentation, include Thurstonian paired comparisons (Eggemeier, 1981), psychophysical techniques of magnitude estimation and categorical scaling (Eggemeier, 1981; Stone and Helm, 1981), and non-ordinal rating scales (Stone and Helm, 1981).

RESEARCH OBJECTIVES

Objectives

Subjective rating scales have proven to be highly sensitive in the measurement of mental workload. The most popular subjective rating scales utilized in workload research are the Cooper-Harper rating scale and various modifications of the Cooper-Harper rating scale. Even though these decision tree scales have widespread appeal and success, no one knows what characteristics cause decision making with these scales to attain their apparent sensitivity. Because workload researchers are continually searching for more sensitive scaling techniques for workload estimation, the emphasis of this study was to examine an existing sensitive workload estimation technique--the Modified Cooper-Harper (MCH) rating scale developed by Wierwille and Casali (1983). By examining six modifications of the MCH decision tree rating scale, more could be learned regarding the use of a decision tree to simplify complex subjective rating decisions. The major objective of revising the MCH scale was to possibly attain a more sensitive measure of mental workload.

Some justifications for attempting to improve the MCH rating scale lie in the scale's initial development: its design is not based on well-established psychological scale development concepts; the scale is unidimensional, while mental workload is believed to be multidimensional; the scale is based solely on engineering judgements; and the number of categories incorporated in the scale and the scales divisions (3-3-3-1) were acquired arbitrarily. Also, the scale logic of the MCH is based on a decision tree. The use of decision trees is virtually non-existent in psychological scaling, but decision trees have been used successfully in leadership models by Vroom (Gilmer and Deci, 1977) and in job descriptions and evaluations by Mallamad et al. (1980).

Through closer examination of the MCH scale's characteristics, a better understanding of the scale's attributes and deficiencies could be obtained. Objectives which were to be explored in this study were:

1. Does the use of a decision tree contribute to the scale's sensitivity, and how will the scale's sensitivity change with the elimination of the decision tree structure?

2. Does the decision tree force more uniform logic upon the subjects?
3. Are ten categories of rating optimal when using a decision tree structure?
4. What are the effects of the scale divisions (3-3-3-1)?
5. Does progressing through the primary decision tree decisions step-by-step with the computer affect the ability of the rating scale to discriminate among loading levels compared to the paper-and-pencil scales?

METHOD

Experimental Design

The overall experimental design of this study was 6 experience levels by 6 rating scales by 3 loading levels (6 x 6 x 3), mixed factor, complete factorial design. One pilot was randomly assigned by experience level to each of the 36 cells in the experimental design, see Figure 7. The rating scale factor was a fixed-effect, between-subject variable. The six rating scales used were the Modified Cooper-Harper scale (MCH), a computerized version of the Modified Cooper-Harper scale (COMPMCH), a 15 point modified Cooper-Harper scale (MCH+), a computerized 15 point modified Cooper-Harper scale (COMPMCH+), a performance based modified Cooper-Harper scale (PBMCH), and a tabular modified Cooper-Harper scale (NDT). The load presentation order was a fixed-effect, between-subjects variable. The six permutations for the load presentation order consisted of low (L), medium (M), and high (H) communications loading levels (HLM, MHL, LMH, MLH, HML, and LHM).

Load presentation order was totally confounded with the experience level of the pilots. Each experience level

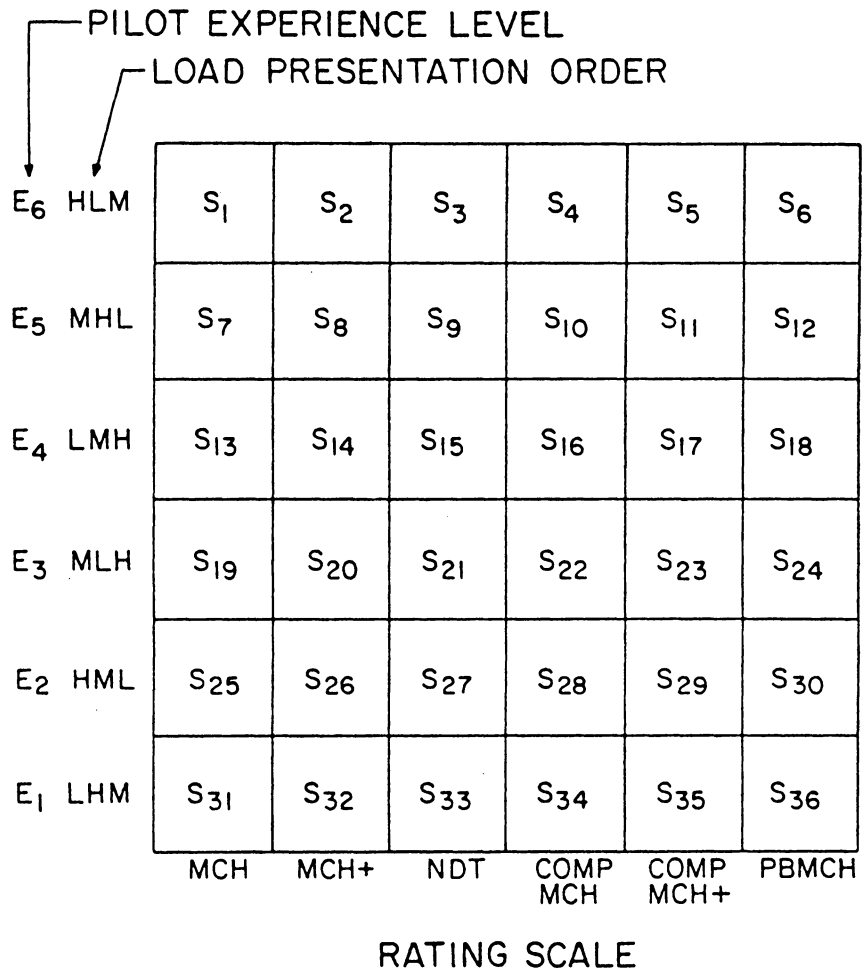


Figure 7: Experimental Design Layout Model

grouping received the same load presentation order in each of the six rating scales (e.g., experience level 6 always received the order--high, low, medium--regardless of the rating scale they used). However, each pilot regardless of his/her grouping received all three loading levels. Thus, load was a fixed-effect, within-subject variable. The load presentation orders were completely counterbalanced across rating scales to eliminate differential transfer effects from one loading level to the next.

Apparatus

Flight simulator. The flight simulator used in this study was a Singer-Link GAT 1B flight simulator for general aviation training located in Virginia Tech's Vehicle Simulation Laboratory. The simulator is a moving-base simulator with three degrees of freedom (pitch, roll, and yaw). The flight characteristics of the simulator were completely controllable by means of a control panel on the outside of the simulator cabin. The simulator cockpit was fully instrumented for IFR flight. A photograph of the simulator cockpit may be seen in Appendix B, Figure 14.

In this study, the pilots operated the simulator according to IFR procedures under controlled fluorescent room lighting. To prevent distractions and maintain a

constant illumination level in the simulator cockpit, translucent blinders were placed over each cockpit window. Also, the simulator was equipped with simulated engine sound and touch-down tire squeal to provide the pilot with typical aircraft flight cues.

In addition, a cockpit speaker system was installed in the simulator and the pilot was equipped with a lapel microphone with which to communicate with the experimenter when in flight. Tape recorded messages with communications stimuli and flight related instructions, developed by Casali and Wierwille (1983), were played to the pilots via the cockpit speaker system. The communications tape messages, which had a male voice, were played from a BIC Model T-1 dual-channel cassette recorder.

For the pilot to communicate with the experimenter ("tower"), the pilot was provided with a push-to-talk snap-action switch (Micro-switch) on the yoke of the simulator. The communication responses of the pilot were recorded on separate channels of a Sony model TC-270 reel-to-reel tape recorder. The recorded tapes were used in the data analysis of the pilot's responses to the communication task stimuli.

A photograph of the experimental station for the communications flights may be seen in Appendix B, Figure 15.

Microcomputer. A TRS-80 Model III microcomputer was used to computerize the MCH and the MCH+ rating scales. The rating scales were programmed in BASIC. Each of the pilots assigned to the computerized scales performed their ratings on the TRS-80 microcomputer. A photograph of a pilot making a typical rating on the TRS-80 may be seen in Appendix B, Figure 17.

Rating Scale Designs

To examine the inherent sensitivity of the MCH scale and possibly to find a more sensitive rating scale using the main characteristics of the MCH scale, six rating scale designs were used. The first rating scale (Appendix C, Figure 18) was the Modified Cooper-Harper rating scale with a standard decision tree developed by Wierwille and Casali (1983). The scale's basic design has a 3-3-3-1 category layout, for a total of 10 rating categories.

The second rating scale used was a computerized MCH scale (COMPMCH). The computerized scale was used to investigate the decision-by-decision process of the MCH decision tree logic and to find out if the pilots were utilizing the decision tree logic as instructed. The scale was programmed so that only a single rating decision would appear on the microcomputer CRT screen at a time (see

Appendix C, Figure 19, and Appendix B, Figure 16 for a step-by-step rating and a photograph of a typical screen view on the COMPMCH, respectively). The computer scale led the pilot into the MCH scale at the bottom and progressed step-by-step through the rating scale based on the pilot's decisions. The scale progression was exactly the same as that of the MCH except for the fact that the pilot could not see where decisions would ultimately lead numerically on the scale. The pilot could not reverse rating decisions, but was permitted to rate again if the final rating was unsatisfactory. Rerating was permitted to eliminate ratings which were improperly performed. (The pilots also had this opportunity with the paper-and-pencil ratings prior to handing the experimenter their final ratings.)

The third rating scale used was an expanded modified Cooper-Harper rating scale (MCH+) (Appendix C, Figure 20). This rating scale incorporated the standard decision tree design on the left side of the MCH scale requiring the same binary decisions at each decision point, but the right side of the scale had additional rating categories bringing the total number of categories to 15. The MCH scale was expanded from a 3-3-3-1 design to a 4-4-4-3 design. The purpose of adding the additional levels was to test whether the MCH scale's sensitivity could be increased. The number

of discrete rating categories used in the MCH scale had never been examined, especially with regard to its decision tree structure. The MCH might have added sensitivity with additional categories because of increased discrimination and precision. This third scale design supported the decision tree approach of the MCH while attempting to extend its ability to discriminate among loading levels.

The fourth scale used was a computerized version of the MCH+ rating scale (COMPMCH+). The scale was computerized in the same manner as the COMPMCH scale discussed earlier (see Appendix C, Figure 21 for an example rating on the COMPMCH+).

In an attempt to study the effects of a different type of decision tree structure, the fifth rating scale used was designed sequentially from left to right. The first decision in the scale involved the errors the pilot perceived while performing the instructed flight task. Thus, the scale was generally a performance-based rating scale, since the performance of the pilot determined what half of the rating scale the pilot entered. This PBMCH scale can be seen in Appendix C, Figure 23).

To examine the effects of the visual decision tree structure on the MCH, the sixth rating scale was the MCH rating scale with the decision tree broken down into a

tabular format, NDT (no visual decision tree) rating scale (Appendix C, Figure 22). The scale maintained the standard adjectival category descriptors, on the right side of the MCH, but was collapsed into a vertical tabular format on the left side. This type of scale was used to determine the positive or negative effects of the MCH's visual decision tree.

Communications Task

Because the primary interest of this study was to examine the ability of various subjective rating scales to discriminate among different levels of mental workload, the designed flight task needed to incorporate adjustable levels of mental activity. A task which proved to be extremely useful in providing transferable workload measures in a previous investigation of mental workload metrics was a communications task designed by Casali and Wierwille (1983). The task provided the subjects with routine flight procedures that required a large degree of mental activity. The subjective workload rating of the pilot on the communications task was largely composed of mental loading, not of extraneous psychomotor and/or perceptual loadings.

The communications task developed by Casali and Wierwille (1983) consisted of two major components: an

aircraft control aspect and a communications aspect. For the aircraft control aspect, the subject was instructed to maintain predesignated performance standards. The flight standards required the subject to fly straight and level while maintaining a flight altitude of 2000 feet ($\pm$ 100 feet), a heading of 360 degrees ($\pm$ 5 degrees), and an indicated airspeed of 100 mph ($\pm$ 10 mph). However, when the "tower" commanded the pilot to change altitude, heading, or airspeed, the pilot was instructed to carry out the command. These flight standards remained constant throughout the experimental runs.

The communications aspect of the flight task required the pilot to listen to an 8-minute tape recorded message played over the simulator loudspeaker. The taped scenario, which had a male voice, included commands to change heading, change altitude, report present altitude and heading, report aircraft model (Cessna 150) and call sign (Cessna-One-Four-India-Echo), change altimeter setting, change radio frequency, and report airspeed. In addition to these commands, randomly constructed abbreviated aircraft call signs consisting of two single-digit numbers (0-9) and two international phonetic letters (e.g., Six-Alpha-Niner-Foxtrot and One-Seven-Bravo-Zulu) were presented in series in the taped scenario. The pilots were

instructed to respond only to their call sign (target), One-Four-India-Echo or certain permutations of that call sign (One-India-Four-Echo, One-India-Echo-Four, One-Echo-India-Four, One-Echo-Four-India, and One-Four-Echo-India) by responding "now" into a microphone attached to the pilot's lapel. Note that the pilot's target began with the numeral "one" in the first position of the call sign sequence. A placard was mounted on the top of the simulator instrument panel to remind the pilot of the specified target call signs. Also, to indicate the pilot's specific call sign at all times during the flight, a typed label with the call sign was mounted on the simulator's instrument panel.

To change the loading levels of the various experimental runs, the call sign recognition frequency was manipulated between flights. Low, medium, and high loading levels were used. The average call sign presentation rates for the respective levels were: low--one call sign every 12 seconds, with no target sign permutations for non-targets call signs; medium--one call sign every 5 seconds with 30 percent of the non-target call signs as permutations; and, high--one call sign every 2 seconds with 40 percent of the non-target call signs as permutations.

The subjects were instructed that adequate performance consisted of carrying out all instructed commands and

identifying all target call signs correctly and as quickly as possible.

Primary Task Measures

During each of the flights, primary task data were recorded on magnetic tape as well as by hand. The primary task data included errors of omission, errors of commission, and reaction time.

Errors of omission. An error of omission occurred when a specified target call sign--One-Four-India-Echo, One-India-Four-Echo, One-India-Echo-Four, One-Echo-Four-India, One-Echo-India-Four, and One-Four-Echo-India--appeared in the communications scenario but was not recognized by the pilot with the response of "now". The total count of missed target call signs was recorded for each experimental flight.

Errors of commission. An error of commission was counted when the pilot responded "now" to a call sign which was not one of the six target call signs discussed earlier. A total number of errors of commission was recorded for each experimental flight.

Reaction time. During the communications flight task, all the responses of the pilot were recorded on magnetic tape. The tape was later played back and the pilot's reaction times to the target call signs were measured in

seconds. Response time was measured from the end of a the target call sign presentation to the pilot's response of "now". Only correct target responses were timed. Response times were recorded for each experimental flight.

Subjects

In this study, a total of 36 pilots, both private pilots and student pilots from the Virginia Tech area, participated. The pilots were required to have 20/25 (far) visual acuity or better, and to have adequate speech frequency hearing to respond properly to the appropriate call signs to qualify for participation. The experience level of the pilots ranged from 25.5 hours to 2109.9 hours of logged flight time, with an average of 193.6 hours. There were four female pilots and 32 male pilots. Four of the pilots were rated instrument pilots. The pilots were sorted and equally distributed according to experience level (number of flight hours) and assigned to six groups of six pilots. Each of the six pilots in a given group rated exclusively on one rating scale (e.g., MCH, COMPMCH, MCH+, COMPMCH+, PBMCH, or NDT). Experience level six represented the more experienced pilots, while experience level one represented the least experienced pilots. By using six subjects in each rating scale group, one from each of the

six experience levels, it was possible to completely counterbalance rating method across load presentation order.

Procedure

First, each pilot was presented with a general description of the experiment and an informed consent form (Appendix D). The general description of the experiment familiarized the pilot with the purpose and procedures of the experiment. The informed consent form outlined the subject's rights. These documents provided the pilot with information upon which he or she might decide whether or not to participate in the communications experiment. After completion of the experimental runs, the pilot was debriefed and paid for participation in the study and then dismissed.

Pre-testing. Before performing the communications task in this study, the pilots completed the following preliminary tests:

1. A personal data sheet for the pilot was filled out via telephone interview. The data sheet included personal data, the pilot's total flight hours, and the types of aircraft flown. The pilot's total flight hours were confirmed before the experimental session by the presentation of the pilot's logbook.

2. The pilot was then asked to present his or her FAA physical card (minimal FAA Class 3) for verification of 20/30 visual acuity and adequate hearing for recognition of flight communications. If the physical card had expired (must be dated within two years of the present date to be valid), the pilot was asked to read a Landolt ring eye chart set up for 20/25 far visual acuity and complete a voice hearing test (McCormick, 1978). For the hearing test, the pilot sat in a chair 8 feet from the experimenter. The experimenter turned her back to the pilot and the pilot was asked to repeat random sentences of text which the experimenter read aloud. If the pilot repeated the text correctly the pilot had adequate hearing to complete the communications task.

After successfully completing the experimental pretests, the pilots were presented with their respective rating scales and rating scale instructions. The rating scales can be seen in Appendix C, while the rating scale instructions appear in Appendix E.

Communications task. Before entering the simulator, the pilot was given primary task (communications task)

instructions. To maintain consistent objectives among all the pilot's flights, the instructions included the objectives, procedures, and adequate performance parameters for the pilot to follow, see Appendix F. After entering the simulator, the pilot was asked to become familiar with the simulator instruments. The pilot was then cleared for takeoff. Each pilot was asked to take off and climb at 500 feet per minute to a straight and level course at an altitude of 2000 feet ($\pm$ 100 feet), an indicated airspeed of 100 mph ($\pm$ 10 mph), and a heading of due North ($\pm$ 10 degrees). The pilot was given time to adjust to the simulator and its flight characteristics before the taped communications task scenarios were introduced. After the pilot stabilized and controlled the simulator according to the flight performance standards, the communications task taped scenario was played over the cockpit loudspeaker system. The first scenario was an 8-minute practice flight. As the taped communications message was played, the pilot responded to pre-designated call signs (targets) and reported flight parameters to the experimenter using the lapel microphone and press-to-talk snap-action switch. The practice flight allowed the pilot to become familiar with the instructed communications task. After the practice flight was completed the simulator was placed in autopilot

mode and the pilot was asked to leave the simulator to make a rating on the rating scale and answer a questionnaire regarding the rating (see Appendix G for the questionnaire). If any misunderstandings were detected during the practice, the task was clarified before the experiment commenced.

Next, each pilot performed three experimental data flights. Between flights the simulator was placed in autopilot mode so that the pilot could exit the simulator to make a rating on each run and complete the questionnaire. After completion of the third experimental run, the pilot landed the simulator.

To assure that the pilot continually provided control inputs and scanned the flight instruments during the flight, mild wind gust turbulence was introduced into the simulator's flight path using the GAT-1B's wind gust generator. The wind gusts had peak amplitudes of about six miles per hour. Each of the experimental runs lasted approximately 10 minutes.

RESULTS

The data analyses for this experiment include a sensitivity analysis on standardized z-scored rating scale scores, a multivariate analysis on the primary task measures, post-hoc analyses including pilot experience level and load presentation order effects, and a comparison analysis between the Casali and Wierwille MCH data and this study's MCH data. In addition, response frequencies from the questionnaire presented to the pilots appear in a tabular format to provide additional information concerning the experiences of the pilots in performing their instructed tasks at the three communications loading levels. All of the statistical data analyses were performed using the Statistical Analysis System (SAS) computer package (SAS Institute, 1982).

Sensitivity Analyses

The goals of the sensitivity analyses were to determine which rating scales were the most sensitive in differentiating among the three communications loading conditions and to suggest the use of a rating scale which

displayed the highest degree of load discrimination. Transformed scores of each of the rating scales were analyzed.

Conversion to z-Scores

Since there were both 10-point and 15-point rating scales, the raw scores from each of the six rating scale techniques were converted to z-scores. The raw scores were standardized for each rating scale across loading levels. Thus, 18 raw score values were converted for each scale. Transformation equations for the mean z-score values to mean raw scores for each loading level can be found in Appendix H, Table 12. By using this standardization method, the rating scales could be reviewed and discussed on a relative basis.

Overall Sensitivity ANOVA

An overall sensitivity analysis was conducted on the 3 x 6, load by rating scale complete factorial design. Load was a within-subject, fixed-effect variable, while rating scale was a between-subjects, fixed-effect variable.

The results of the overall two-way ANOVA indicate no statistically significant main effect of rating scale. This main effect was precluded due to the standardization of the

z-scores. There was a statistically significant main effect of load $F(2,60) = 37.08$, $p = 0.0001$. The analysis showed no interaction between load and rating scale $F(10,60) = 0.93$, $p = 0.5131$. This result was anticipated, because all the rating scales used in the study had the potential ability to discriminate among loading levels. (The results of an unbalanced sensitivity ANOVA, which includes the Casali and Wierwille MCH data with this study's MCH data, appear in Appendix I, Table 13. These results were not substantially different from the results presented in the balanced ANOVA analysis).

Since the interaction of Load x RS was not statistically significant in the overall sensitivity analysis, the conclusion which may be drawn is that the six rating scales discriminate among the three communications loading levels equally well. However, to analyze the rating scales further so as to not dismiss or overlook information of value regarding the rating scales, a new method of analysis was adopted--individual sensitivity ANOVAs.

Individual Sensitivity ANOVAs

A one-way individual sensitivity ANOVA was performed on each of the six standardized rating scale data sets at $\alpha = 0.01$. (Since the data obtained from the six rating scale

TABLE 1

ANOVA Summary Table for the Overall Experimental Design

<u>Source</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Between-Subjects				
Rating Scale (RS)	5	0.0000	0.00	1.0000
Subjects (S/RS)	30	1.4433		
Within-Subjects				
Load (L)	2	15.1727	37.08	0.0001
L x RS	10	0.3803	0.93	0.5131
L x S/RS	<u>60</u>	0.4092		
Total	107			

experiments were ordinal ratings, the results of the parallel non-parametric analysis (NP), Friedman's Rank Sums (two-way) analysis, are also presented in Table 2.) The loading levels were treated as within-subject, fixed-effect variables. The summary tables for the individual ANOVAs appear in Table 2. As can be seen from the results, the MCH, COMPMCH, and PBMCH rating scales were all significant at $p = 0.01$.

Duncan's Multiple Comparison Tests

For the three rating scales, MCH, COMPMCH, and PBMCH, which were statistically significant in the individual ANOVAs above, Duncan's test was used for multiple comparisons among load means. Results of the Duncan's tests appear in Table 3, along with the mean z-score values for low, medium, and high loads. Plots of the rating scale mean z-score values and significant comparisons can be seen in Figures 8, 9, and 10. The MCH rating scale showed the greatest sensitivity, in that the MCH discriminated not only between low and high loading as did the COMPMCH and PBMCH scales, but also between low and medium loading levels.

TABLE 2

Summary Tables for the Individual Rating Scales

<u>Source</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>	NP <u>p</u>
<u>Rating Scale: Modified Cooper-Harper (MCH)</u>					
Load	2	4.6742	10.52	0.0035	<0.01
Subject x Load	10	0.4444			
<u>Rating Scale: 15 Point Modified Cooper-Harper (MCH+)</u>					
Load	2	2.1844	2.47	0.1347	N.S.
Subject x Load	10	0.8856			
<u>Rating Scale: Computerized Modified Cooper-Harper (COMPMCH)</u>					
Load	2	1.5632	13.33	0.0015	<0.01
Subject x Load	10	0.1172			
<u>Rating Scale: Computerized 15 Point Modified Cooper-Harper (COMPMCH+)</u>					
Load	2	1.1264	4.37	0.0433	N.S.
Subject x Load	10	0.2578			
<u>Rating Scale: Tabular Modified Cooper-Harper Scale (NDT)</u>					
Load	2	1.9881	5.57	0.0237	<0.01
Subject x Load	10	0.3570			
<u>Rating Scale: Performance Based Modified Cooper-Harper (PBMCH)</u>					
Load	2	5.5382	14.09	0.0012	<0.01
Subject x Load	10	0.3930			

TABLE 3

Duncan's Tests for the Statistically Significant Rating Scales*

<u>Rating Scale</u>	<u>L</u>	Mean Values <u>M</u>	<u>H</u>
<u>Modified Cooper-Harper (MCH)</u>			
	-1.0111	0.3946	0.6165
<u>Computerized Modified Cooper-Harper (COMPMCH)</u>			
	-0.5104	0.0000	0.5104
<u>Performance Based Modified Cooper-Harper (PBMCH)</u>			
	-0.9828	0.0457	0.9371

*Means with common underlines do not differ significantly at $p < 0.01$.

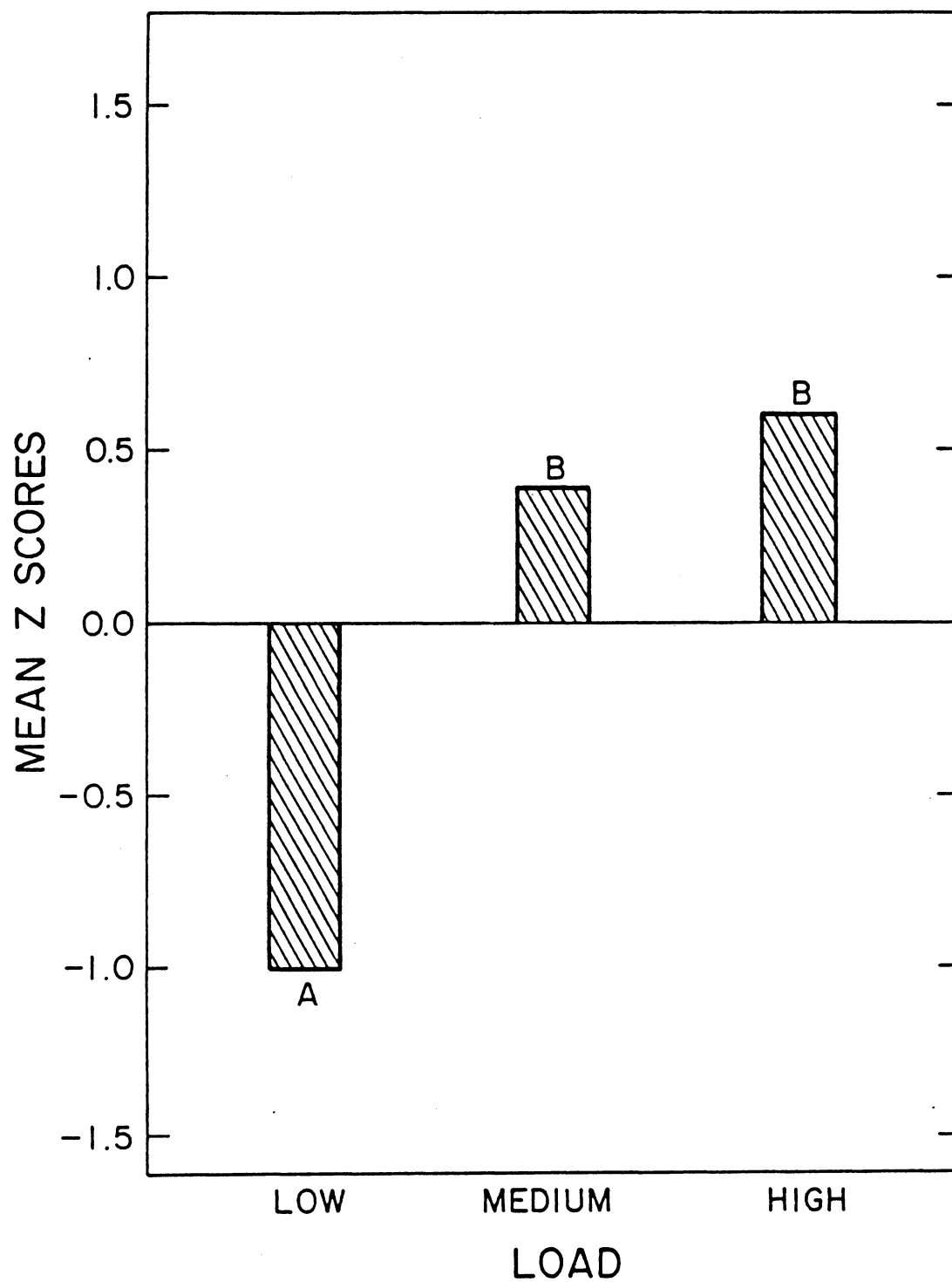


Figure 8: Mean Z-scores and Duncan's Test for the MCH Scale

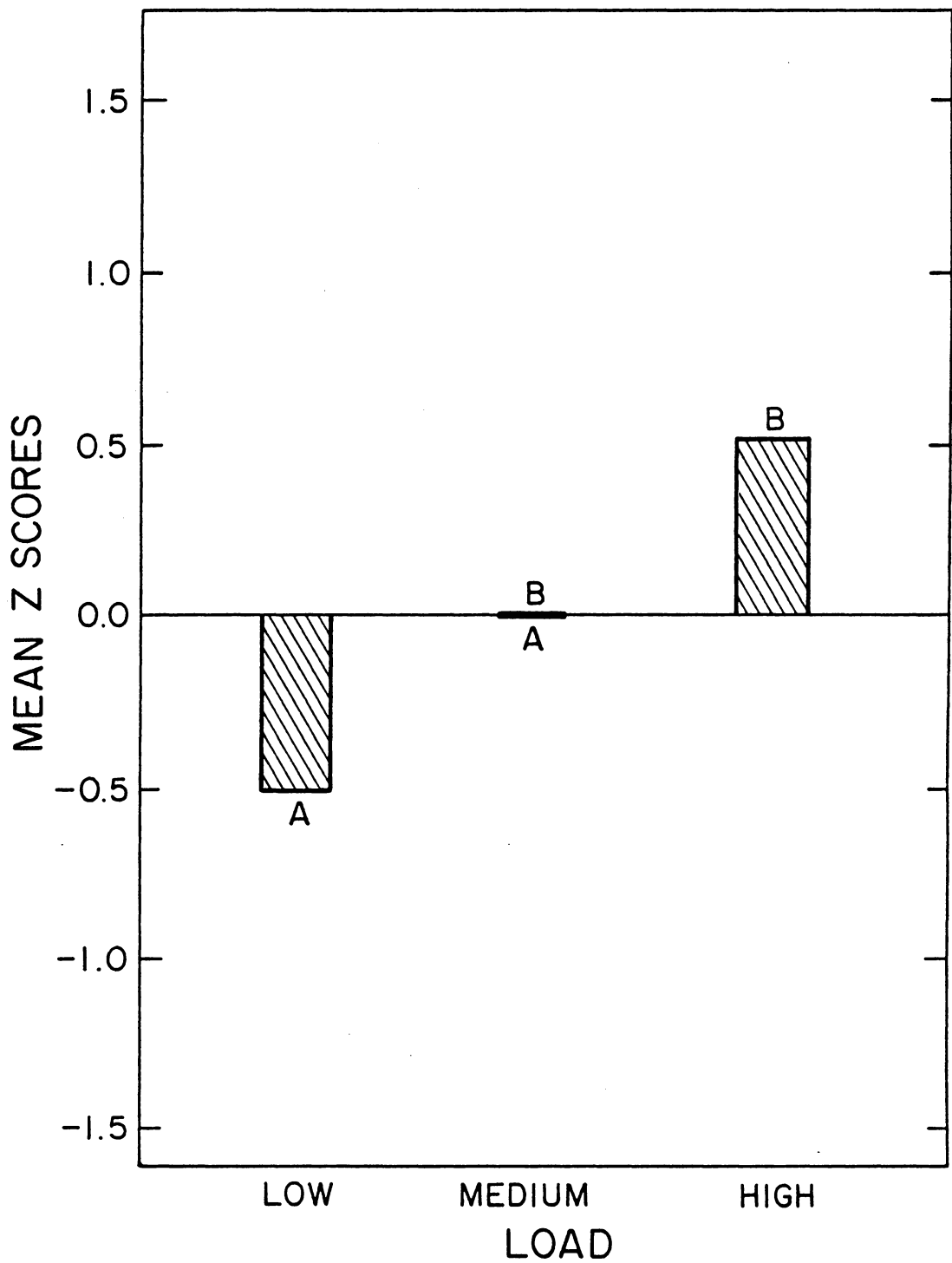


Figure 9: Mean Z-scores and Duncan's Test for the COMPMCH Scale

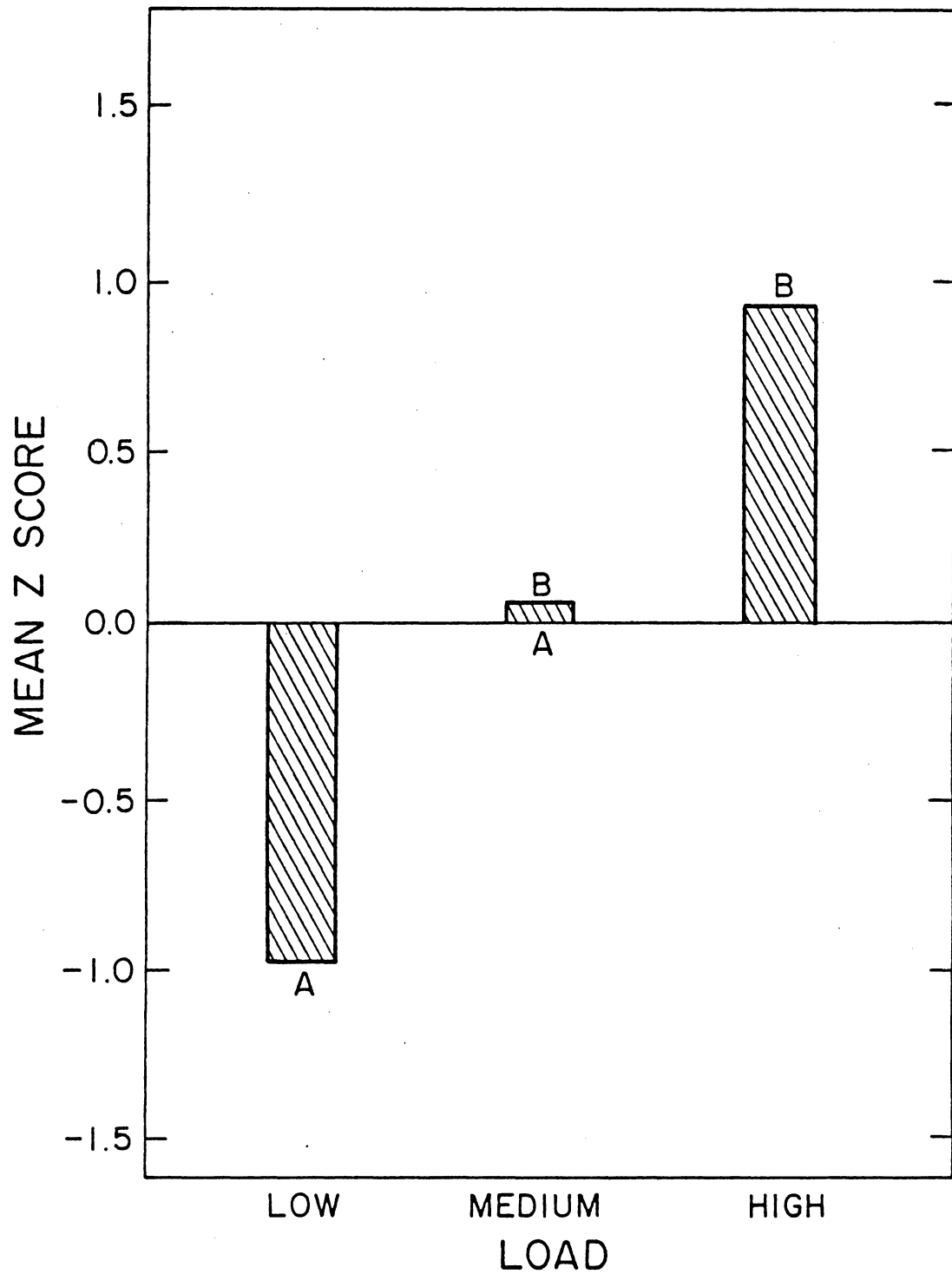


Figure 10: Mean Z-scores and Duncan's Test for the PBMCH Scale

Additional Information

The following information is purely speculative, but is provided so that additional information and insight into the nature of the rating scales used in this study may be obtained. An important result which should not be withheld is that all six of the rating scales used in this study showed an increasing monotonic trend from low to high communications loading. All the scales except one, the MCH+, met the less conservative testing criterion of $\alpha = 0.05$ (Table 2). In fairness to the remaining scales, the Duncan's test was performed at the $\alpha = 0.05$ level. The results of the Duncan's test appear in Table 4. As can be seen, the discrimination among loading levels changes for the COMPMCH and the PBMCH rating scales at $\alpha = 0.05$. Reasons for the shift will be discussed later.

Primary Task Measures

The primary task performance dependent measures which were collected during the experimental trials included errors of omission, errors of commission, and reaction times. These dependent measures were obtained to test for any significant performance variations between rating scale

TABLE 4

Duncan's Tests with a Less Conservative Testing Criterion

<u>Rating Scale</u>	<u>L</u>	<u>Mean Values</u> <u>M</u>	<u>H</u>
<u>Modified Cooper-Harper (MCH)</u>			
	-1.0111	0.3946	0.6165
<u>Computerized Modified Cooper-Harper (COMPMCH)</u>			
	-0.5104	0.0000	0.5104
<u>Performance Based Modified Cooper-Harper (PBMCH)</u>			
	-0.9828	0.0457	0.9371
<u>Computerized 15 Point Modified Cooper-Harper (COMPMCH+)</u>			
	-0.4839	0.1320	0.3519
<u>Tabular Modified Cooper-Harper (NDT)</u>			
	-0.5222	-0.0950	0.6172

*Means with a common underline do not differ significantly at $p < 0.05$.

groups and between pilot experience levels. The raw score data for each primary task measure were analyzed in a multivariate analysis of variance (MANOVA) procedure. The F-approximation of the Wilk's U criterion statistic was used. The MANOVA summary table for variations between rating scale groups appears in Table 5. The MANOVA summary table for experience level performance effects appears in the next section.

Neither rating scale groups (Table 5) nor experience level (Table 6) was statistically significant. The F approximation values for the U-statistic and p-values were $\underline{F}(15,77) = 1.40$, $p = 0.1663$ and $\underline{F}(15,77) = 1.61$, $p = 0.0895$, respectively. That is, the pilot groupings had no statistically significant performance variations. Thus, it can be assumed that there were no differential effects of performance within the experience level or rating scale groupings to cause any deterioration or enhancement of a single rating scale. All groups performed the instructed communications task at approximately the same performance levels.

TABLE 5
MANOVA Summary Table for RS Performance Effects

<u>Source</u>	<u>dv</u>	<u>df</u> _H	<u>df</u> _E	<u>F</u>	<u>p</u>
Between-Subjects					
Rating Scale (RS)	3	5	30	1.40	0.1663
Subjects (S/RS)	(Error Term for RS)				
Within-Subject					
Load (L)	3	2	60	30.37	0.0001
L x RS	3	10	60	1.20	0.2343
L x S/RS	(Error Term for L and L x RS)				

where: dv = number of dependent measures
 df_H = degrees of freedom for treatment effect
 df_E = degrees of freedom for error effect
 F = F approximation for Wilk's U
 likelihood ratio statistic

Pilot Experience Level Analyses

The procedure for the MANOVA analysis for differential effects due to experience level groupings was discussed above. This analysis tested the dependent performance measures for variations in performance among experience level groups. The results are presented in Table 6.

A second analysis was conducted to examine differential experience level effects on the independent rating scale raw scores. Since the experience level scores are totally confounded within the presentation order of each of the six experience levels, the rating scale raw scores were summed across load to arrive at a single value for each pilot subject. Although there is no interval or ratio relationship between the 15-point and the 10-point rating scales, the 15-point scales (COMPMCH+, MCH+, and PBMCH) were divided by 1.5 to arrive at a 10-point value to use in this post hoc analysis. In general, a step like this should not be used to compare these rating scales. But, due to the fact that the emphasis of this study is not on experience level variations and that the post hoc analysis is being used to obtain general information about the subject population, a step such as this should not create major biases in the results. The results will still show differential effects of the pilot experience levels if they

TABLE 6

MANOVA Summary Table for EL Performance Effects

<u>Source</u>	<u>dv</u>	<u>df_H</u>	<u>df_E</u>	<u>F</u>	<u>p</u>
Between-Subjects					
Exper. level (EL)	3	5	30	1.61	0.0895
Subjects (S/EL)			(Error Term for EL)		
Within-Subject					
Load (L)	3	2	60	31.00	0.0001
L x EL	3	10	60	1.47	0.0681
L x S/EL			(Error Term for L and L x EL)		

where: dv = number of dependent measures
df_H = degrees of freedom for treatment effect
df_E = degrees of freedom for error effect
F = F approximation for Wilk's U
likelihood ratio statistic

are indeed present in the data. A one-way sensitivity ANOVA was performed with experience level as a between-subjectss, fixed-effect variable. The ANOVA summary appears in Table 7. Neither the analysis for the dependent (MANOVA) (Table 6) nor the independent (ANOVA) measures was statistically significant. However, it should be noted that the ANOVA main effect of experience level was close to acceptable significance ($p = 0.0579$). In plotting the mean values of the experience level ratings, seen in Figure 11, there is no apparent trend or order to the results. Even if experience level was significant, interpretation of this result would be difficult because of the cell-to-cell variation and the absence of a monotonic trend.

Nevertheless, to test for a possible decrease in score values with increasing experience, a linear regression analysis was performed on the collapsed pilot scores (not the mean values in Figure 11). The pilot's logged flight hours were used as the independent variable, while the pilot rating scores collapsed across load were the response variables in the analysis. The pilot's actual logged flight hours were used in the regression analysis to represent each of the pilot's hours of flight experience. The analysis indicated that there was no significant correlation between the logged flight hours and the pilot mean scores, $r =$

TABLE 7

ANOVA Summary Table for Experience Level Effects

<u>Source</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Experience Level (EL)	5	49.6643	2.43	0.0579
Subjects (S/EL)	<u>30</u>	20.4321		
Total	35			

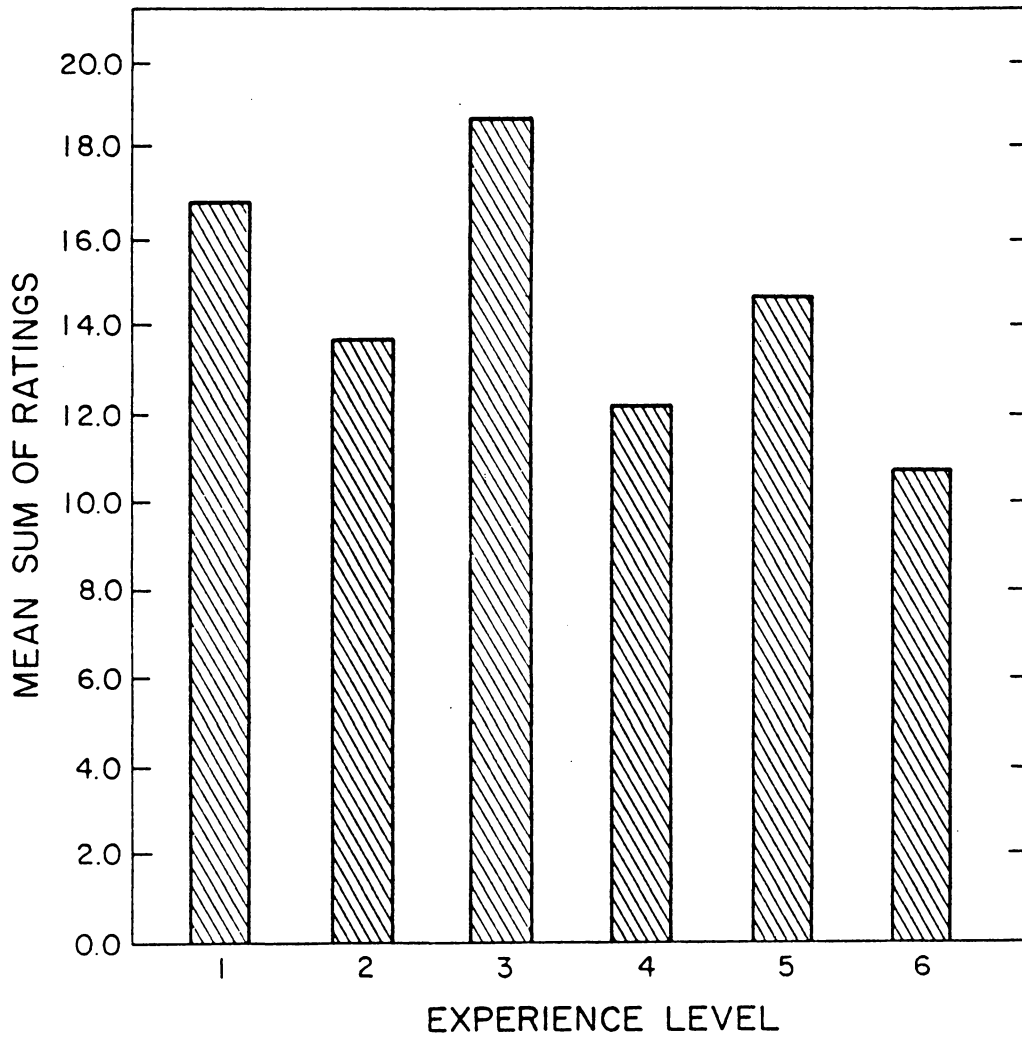


Figure 11: Plot of Mean Values for Experience Levels

-0.1115 (linear fit of the regression line). Thus, there was no statistically reliable decreasing trend in ratings due to the pilot's experience level. The regression equation for the analysis was $y = -0.00154x + 14.809$, with flight hours as the independent variable in the analysis. The regression ANOVA results were $F(1,35) = 0.43$, $p = 0.5173$. The regression equation was not a good linear model for mean rating score based on the pilot's logged hours of experience.

Load Presentation Order Analysis

To conduct an analysis on the presentation order of the communications loading level, the 15-point scale raw scores, converted to 10-point scale values (as discussed earlier) were used. A one-way ANOVA with loading level as a within-subject, fixed-effect variable was then performed. The ANOVA summary table appears in Table 8. The results support the null hypothesis that there was no differential variation due to the presentation order of the loads on the independent rating scale values.

TABLE 8

ANOVA Summary Table for Load Presentation Order Effects

<u>Source</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Between-Subjects				
Subject (S)	35	8.4125		
Within-Subject				
Order (O)	2	1.0649	0.30	0.7425
O x S	<u>70</u>	3.5609		
Total	107			

Comparison of MCH Ratings with Previous Experiment's Ratings

The goal of this sensitivity analyses was to establish the reliability of this experiment's rating scale data. Although other rating scale comparisons could have been made here, only the MCH rating scale data were used. The Modified Cooper-Harper (MCH) rating scale raw score data of the communications study by Casali and Wierwille (1983) and this experiment's MCH data were analyzed in a two-way analysis of variance (ANOVA). Both studies utilized the MCH rating scale (Appendix C), but in the sensitivity analysis the two MCH scales were treated as different scales (experiments) for the purpose of detecting statistical differences in the ratings of the two studies. The two experiments' rating scales were treated as a between-subjects variable, while the loading levels (L, M, and H) were treated as a within-subject variable.

The summary table for the ANOVA appears in Table 9. The ANOVA resulted in a significant main effect of load $F(2,20) = 21.87$, $p = 0.0001$. This result verifies that the manipulation of load was effective in changing the rating scale responses of the pilots. The main effect of experiment was not statistically significant $F(1,10) = 4.60$, $p = 0.0576$. The fact that the main effect of load was near statistical significance does not effect the statistical

reliability of this study's results. The two studies were conducted nearly two years apart from one another and the simulator turbulence setting for the instructed communications flight task in the Casali and Wierwille study was not precisely known. Thus, the two studies may not have been identical replicates of one another. There was no statistically significant Load x Exp interaction, $F(2,20) = 0.86$, $p = 0.4396$. Thus, the rating scales' raw scores did not differ significantly across the three loading levels. Each of the scales resulted in increasing monotonic trends from the low to the high loading levels. The similarity of the two experiments' data sets is illustrated in Figure 12. (The mean values in the figure are standardized mean z-scores for each experiment).

Questionnaire Data Presentation

In an attempt to gain additional insight into the individual ratings of the pilots, a questionnaire (see Appendix G) was presented to the pilots after each experimental trial rating. The pilots were asked to explain their rating based upon their experiences or perceptions while performing the instructed communications task. The answers to the questionnaires were easily definable as positive and negative comments (Tables 10 and 11). As the

TABLE 9

ANOVA Summary Table for Comparison of the MCH Data Sets

<u>Source</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Between-Subjects				
Experiment (EXP)	1	26.6944	4.60	0.0576
Subjects (S/EXP)	10	5.8056		
Within-Subjects				
Load (L)	2	34.7500	21.87	0.0001
L x EXP	2	1.3611	0.86	0.4396
L x S/EXP	<u>20</u>	1.5889		
Total	35			

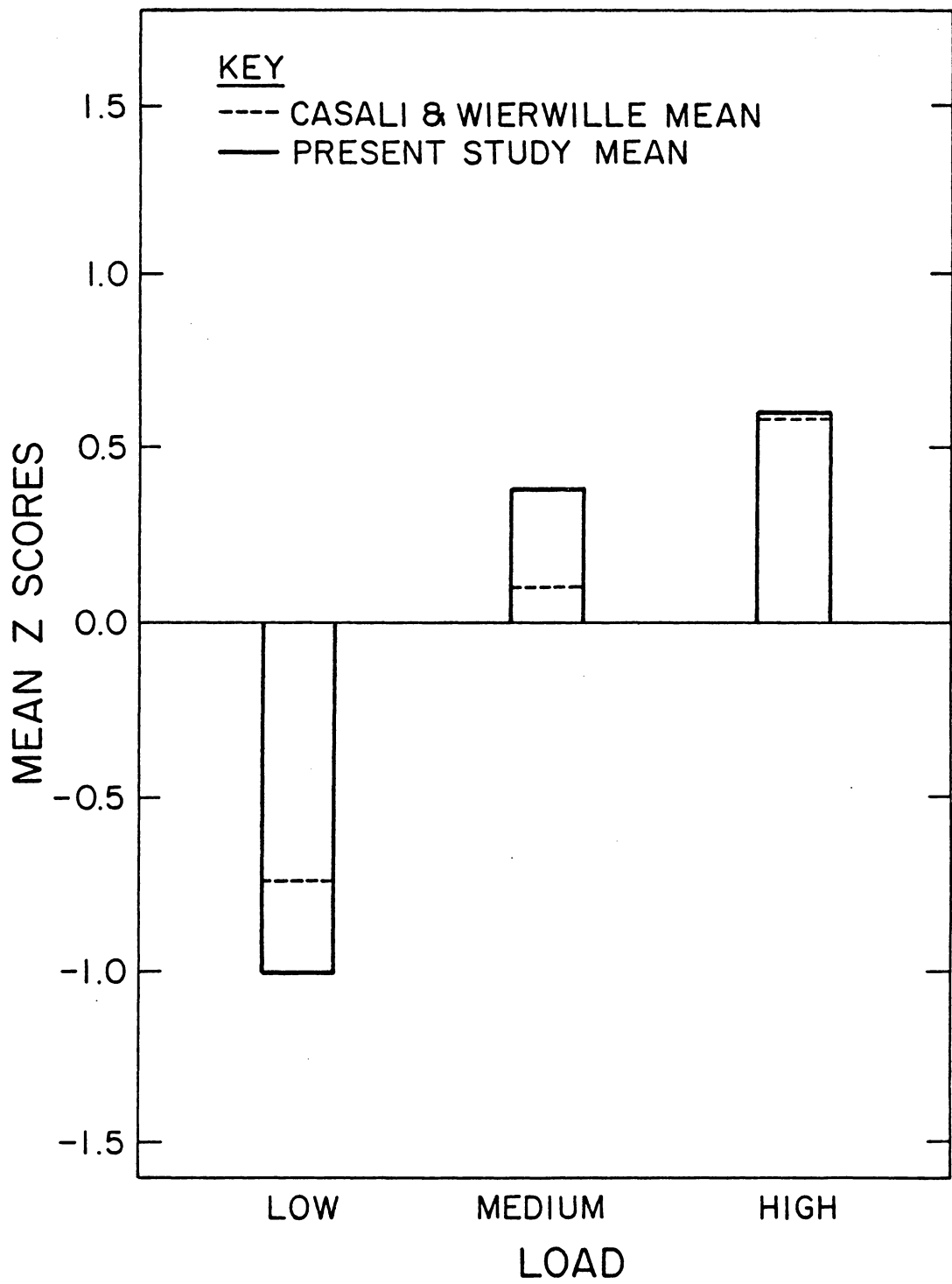


Figure 12: Plotted Means for MCH Data Sets

communications load changes from low to medium to high, the comments from the pilots typically shifted from positive to negative feelings for their performance.

A 2 x 3 (response tone x load) Chi-square contingency table analysis resulted in a significant effect of response tone versus loading level χ^2 , 2 df = 68.326, $p < 0.0001$. This fact supports the effectiveness of the manipulation of the communications loading levels.

TABLE 10
Questionnaire Negative Responses

<u>Response</u>	<u>L</u>	<u>M</u>	<u>H</u>
Recognition of Target Call Sign Difficult	Total = 10	22	23
.Too fast, could not separate call signs	--	10	14
.High concentration required	1	7	2
.Call signs immediately after commands pose difficulties	2	2	4
.Inadequate time to respond	1	2	2
.Too many call sign strings	--	1	1
.Errors inevitable despite full concentration	--	--	4
.Moderate concentration required	6	--	--
Time-Sharing Difficult, Diverts Attention	Total = 6	15	19
.Emphasis not specified	2	5	6
.Missed or forgot commands while attending to the call signs	2	2	5
.Emphasis on communications	2	1	4
.Communications hampered control	--	5	2
.Emphasis on aircraft control	--	2	2
.Tended to ignore communications	--	--	4
Aircraft Control Difficult	Total = 3	3	2
.Parameters difficult to maintain	1	3	1
.Turbulence high	1	--	1
.Simulator sensitive	1	--	--
Slight Stress	Total = 0	0	1

TABLE 11
Questionnaire Positive Responses

<u>Response</u>	<u>L</u>	<u>M</u>	<u>H</u>
Recognition of Target Call			
Sign Adequately Performed	Total = 20	1	0
.Sufficient time to respond	7	1	--
.Call signs reasonably spaced	6	--	--
.More time to fly	3	--	--
.Could time-share attention	2	--	--
.Easy to understand	2	--	--
Time-Sharing Attainable	Total = 14	0	0
.More time to fly	3	--	--
.Allocated attention	2	--	--
.Little mental effort	2	--	--
.Communications did not interfere			
with control	2	--	--
.Alleviated stress	1	--	--
.Could listen for call signs	1	--	--
.Less concentration required	1	--	--
Aircraft Control Easy	Total = 4	1	0
.Minor errors	2	1	--
.Good control performance	2	--	--
.No perceived problems	1	--	--
.Low turbulence	1	--	--
Able to Block Out Unimportant			
Call Signs	Total = 0	0	1

DISCUSSION

Sensitivity of the Individual Rating Scales

The three rating scales which were statistically sensitive (able to discriminate between or among loading levels) were the MCH ($p = 0.0035$), COMPMCH ($p = 0.0015$), and PBMCH ($p = 0.0012$). The MCH scale, as was mentioned earlier, discriminated between low and medium loading levels and between low and high loading levels, while the COMPMCH and the PBMCH only discriminated between the low and high loading levels. Thus, the MCH rating scale was the most sensitive rating scale, since it was able to discriminate among more loading levels. Possible reasons for the sensitivity of the MCH scale will be discussed later.

The presentation of the pilot performance errors as the first item in the rating scale appeared to increase the PBMCH's ability to discriminate. The PBMCH was the only 15-point scale which attained statistical significance. Since the other 15-point scales were not as effective, the design of the scale might have caused the important difference. In the PBMCH scale, pilots were first queried regarding errors. Subsequently, they were queried regarding

either workload level or task accomplishability. They might therefore have progressed in a more orderly manner through the rating scale. However, the Duncan's test resulted in a discrimination only between the low and high loading levels, which indicated less discriminability than did the MCH.

The three rating scales which were not statistically sensitive to loading levels at $\alpha = 0.01$ were the MCH+ ($p = 0.13$), the COMPMCH+ ($p = 0.04$), and the NDT ($p = 0.02$). The MCH+ was the poorest of these scales with $p = 0.13$, since it did not meet the less conservative criterion of $\alpha = 0.05$. Thus, the 15-point rating scale with additional rating categories did not increase the MCH rating scale's ability to discriminate. The additional levels reduced the MCH sensitivity. This was also evident in the COMPMCH+.

In the results presented so far, the computer aided rating appeared to influence the ability of the MCH scale to discriminate. The COMPMCH resulted in a decreased ability to discriminate when compared to the MCH. That is, the COMPMCH did not discriminate between the low and medium loading levels, while the MCH did discriminate between these levels.

Additional Sensitivity Information

By examining the additional analyses made at the $\alpha = 0.05$, a few aspects of the rating scales become more evident. At $\alpha = 0.05$, the NDT rating scale was only able to discriminate between the low and high loading levels. The only other scale that displayed such a low ability to discriminate was the COMPMCH+ scale. Since the NDT discriminates at such a low level, it appears that the visual decision tree was an aid to the pilot when rating the workload of the communications task.

The Duncan's test at $\alpha = 0.05$ for the COMPMCH and the PBMCH rating scales resulted in a higher level of discrimination. These two scales discriminated among all three loading levels. The MCH might be a more reliable rating scale, since the MCH was consistent in discriminating regardless of the criterion used, conservative (0.01) or less conservative (0.05). The MCH appeared to have consistency, while the other scales did not.

The computer implemented COMPMCH+ appeared to enhance the MCH+ by causing the rater to progress through the rating scale on a decision-by-decision basis. The MCH+ lacked statistical significance at $\alpha = 0.05$, while the COMPMCH+ was significant and discriminated between the low and high loading levels. There was also evidence that utilizing the

computer enhanced the MCH rating scale. The computer aided ratings appear to have helped the pilot arrive at a rating by introducing one major decision at a time. This suggested that the raters may not be using the decision tree properly when making their ratings on the paper-and-pencil rating scales.

Primary Task Measures

The primary task measures of errors of omission, errors of commission, and reaction time were included to investigate any differential effects due to pilot groupings. The two pilot groupings which were examined were grouping of pilots by rating scale and grouping by experience level. Neither of the primary task MANOVAs resulted in statistically significant performance variations among the pilot groupings. The p-values were 0.234 and 0.068, respectively, for any Load x Rs and Load x El interactions. Pilot performance, therefore, did not vary significantly among groupings.

Experience Level and Load Presentation Order

As the statistical analyses indicate, no significant variations in the ratings resulted from pilot experience level or from load presentation order. The pilot's

experience level (number of logged flight hours) did not enhance or degrade the subjective ratings ($p = 0.057$). While the load presentation order might have severely affected the results of the study due to the effects of training or differential transfer between loading levels, the sensitivity analysis indicated no significant variation between the load presentation order groupings ($p = 0.742$).

Comparison of the Two MCH Data Sets

The fact that there was no statistically significant Load x Exp interaction ($p = 0.439$) between the Casali and Wierwille (1983) MCH data and this study's MCH data supports the hypothesis that the two communications task experiments were the same. The two MCH data sets were statistically quite similar. Another similarity which should be noted is that the levels of loading which the MCH scales were able to discriminate were identical. The Casali and Wierwille (1983) use of the MCH scale resulted in discrimination between low and medium loading levels and between low and high loading levels using the Duncan's test at $\alpha = 0.01$. The same results were obtained for the MCH scale in this study.

Questionnaire Responses

In general the responses to the questionnaire regarding the pilots' individual ratings reflect an increase in relative difficulty from the low to medium to high communications loading levels. At the low loading levels, the pilots typically responded with positive comments including "recognition of target call signs was adequately performed", "time-sharing was attainable", and "aircraft control was easy". As the communications task required more activity from the pilots at the medium and high communications levels the responses shifted to negative comments--"recognition of target call signs was difficult", "time-sharing was difficult, diverted attention", "aircraft control was difficult", and "slight stress". The most prominent of these comments reflected the inability of the pilots to separate the target call signs, to time-share between the aircraft control and the communications task scenario, and to remember the "tower" commands while attending to the instructed task. The concentration level required to perform the instructed task was high, and at times above the capability of the pilot.

The questionnaire was given to the pilots after rating to provide information which might be valuable in the redesign of subjective rating scales. The pilots' responses

reflect their problems in performing the task well, but the information obtained does not provide insight into the redesign of subjective rating scales as was hoped. A possible keyword which might be applicable to the communications task was "time-sharing". However, the instructions given to the pilot prior to performing the flight task did not ask the pilot to separate the two task aspects by time-sharing, or to favor one task aspect over another aspect.

Comparison of the Communications Results with a Mediation Study

Additional information which aids in the interpretation of the results of this study can be obtained from a companion study involving a mediational flight task by Rieger (1983). Rieger's study involved an investigation of the same six rating scales using a mediational task which required high mediational and perceptual activity from the pilot. The mediational task was conducted after the communications task, and the same six pilots were used for each of the six rating scales. Thus, the six ratings for the communications and for the mediational tasks for each rating scale were performed by the same six pilots.

The MCH rating scale on the mediational task was sensitive ($p = 0.0039$). But, the scale discriminated only

between the low and high loading levels for the mediational task loading.

Several discrepancies between the Reiger (1983) results and the communications study results occurred. Three of these differences which appeared to be rather extreme were the statistical sensitivity of the MCH+ ($p < 0.0001$), COMPMCH+ ($p < 0.01$), and NDT ($p < 0.0001$). None of these scales reached statistical significance at $\alpha = 0.01$ in the communications task, but with the mediational task these scales resulted in a high ability to discriminate. Using the Duncan's test at $\alpha = 0.01$, the MCH+ was able to discriminate among all loading levels, the NDT was able to discriminate between low and high loading levels and between medium and high loading levels, and the COMPMCH+ discriminated only between low and high loading levels.

An explanation for these results might be based on the task designs. While the pilots were performing the communications task, typical comments from the pilots were that they did not know how many errors they had made or if the errors were inconsequential or not. This problem of identifying errors was due to the continuous nature of the communications task scenarios. Task errors tended to be masked, making it difficult for the pilots to quantify the number of errors they actually made during the flight

scenarios. The pilots for the most part were relying entirely on the feeling for how well they had performed, not knowing if they had errors of omission or commission, when rating.

On the other hand, the mediational task was designed so that each mediational problem was presented at 25-second intervals on a slide screen. The mediational problems were wind triangle problems which required the pilot to perform triangle matching, addition or subtraction, and multiplication or division. Since the mediational problems involved the use of mathematical skills for which the pilots had a general idea of how well they performed beforehand, the pilots were able to quantify their errors more readily than they were able to in the communications task. Also, since the slides were automatically advanced after 25 seconds, if the pilot had not given an answer before the next slide appeared it was an evident error. In general, the mediational task was more quantifiable than the communications task. This might have caused the mediational rating scales to be more sensitive in discriminating among loading levels. Thus, there was task dependency of the rating scales.

Due to the quantifiable nature of the mediatioanl task, the pilots appeared to use the MCH+ 15-point rating scale

with an increased effectiveness ($p < 0.0001$). The additional rating scale categories increased the scales' ability to discriminate. The MCH+ discriminated among all three loading levels. Yet, the other 15-point scales, the COMPMCH+ and the PBMCH, were not as effective in increasing the discrimination among loading levels. The COMPMCH+ only discriminated between the low and high loading levels, while the PBMCH was not statistically significant at $p = 0.01$.

A result which contradicts the idea of the mediational task being quantifiable is the lack of statistical sensitivity in the PBMCH ($p = 0.38$). Since the PBMCH is a performance based scale and the number of errors are considered first in rating, it would seem that use of the scale should result in statistical significance; but, it did not. This result is difficult to interpret since the same six pilots performed the ratings for the communications and the mediational experiments, and there were no differential effects in the ratings due to load presentation order or experience level for either experiment (mediational, $p = 0.28$).

The aid of the computer in rating the mediational task did not reflect the same increase in discrimination as did the communications task. The MCH discriminated between low and high loading levels, while the COMPMCH discriminated between low and high loadings and between medium and high

loadings. This result verifies the contributions of the computer in rating. However, the COMPMCH+ results were the opposite. The MCH+ discriminated among all loading levels, while the COMPMCH+ discriminated only between the low and high mediational loading levels.

One final similarity in the results of the two studies was the statistical significance of the COMPMCH ($p = 0.0030$), for the mediational task. However, the COMPMCH for the mediational task discriminated between the low and high loading levels as well as between the medium and high loading levels, whereas the communications COMPMCH only discriminated between low and high loading levels.

SUMMARY

In general, it appears as though rating scales are task dependent. A rating scale can appear to be highly appropriate for one application and be incompatible in another application, as with the PBMCH rating scale discussed earlier. Further study needs to be conducted concerning a rating scale's ability to discriminate on specific tasks.

It is apparent that a computer aided rating was effective on the communications task, whereas it was less effective with the mediational task. The fact that the utilization of the computer enhanced ratings at all is supportive of the step-by-step decision process inherent in a decision tree.

The decision tree rating scales in the communications task aided the pilot in simplifying the decision process. The fact that the communications task was difficult for the pilots to quantify supports the use of a decision tree in an ambiguous situation, since the pilots were able to use the rating scales effectively and the scales discriminated among loading levels. When the task was difficult to quantify the

10-point rating scales or the performance based rating scale resulted in a higher degree of discrimination. However, for a quantifiable task such as the mediational task the 15-point MCH+ scale appeared to be more appropriate.

An aspect which should be noted when addressing subjective rating scales is that they are able to discriminate among various levels of mental workload. Only three of the scales in the communications experiment and one of the scales for the mediational experiment were not significant at $p = 0.01$. Subjective rating scales are effective. Due to the apparent task dependency of the scales, one scale may be appropriate for one task, but not for another. The selection of a rating scale appears to be extremely critical. If this is the case, a set of guidelines for the selection of a rating scale needs to be established based exclusively on the type of task being performed and rated.

In conclusion, for a task which has a high communications loading, the MCH scale is recommended. But, if the task has high perceptual and mediational loading, as in Rieger's (1983) mediational study, a scale with a larger number of rating categories such as the MCH+ is preferred. However, when the loading characteristics are divided and one task aspect does not outweigh any other task aspect, it

is advisable to implement the MCH scale. The MCH was the most sensitive scale for the communications task, and was fairly sensitive for the mediational task. When compared with other techniques of mental workload measurement, the MCH scale has provided the most consistent results (Wierwille and Casali, 1983).

REFERENCES

- Anderson, D. R., Sweeney, D. J., and Williams, T. A. An introduction to management science quantitative approaches to decision making. (Second edition) St. Paul, MI: West Publishing Co., 1976.
- Billings, R. S. and Cornelius, E. T. Dimensions of work outcomes: A multidimensional scaling approach. Personnel Psychology, 1980, 33, 151-162.
- Bird, L. K. Subjective rating scales as a workload assessment technique. Proceedings, 17th Annual Conference on Manual Control, Los Angeles, CA, June 16-18, 1981, 33-39.
- Boyd, S. P. The use of conjoint analysis for interval subjective scaling of mental workload. Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, April, 1982.
- Brown, D. B. Systems analysis and design for safety. New Jersey: Prentice-Hall, 1976.

Canada, J. R. and White, J. A., Jr. Capital investment decision analysis for management and engineering. New Jersey: Prentice-Hall, 1980.

Casali, J. G. A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activities. Unpublished Ph. D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA, August, 1982.

Casali, J. G. and Wierwille, W. W. A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activities. Proceedings, International Conference on Cybernetics and Society, 1982, 598-602.

Casali, J. G. and Wierwille, W. W. Communication imposed pilot workload: A comparison of sixteen estimation techniques. Proceedings of the 2nd Annual Symposium on Aviation Psychology, Columbus, OH, April 26-27, 1983, in press.

Childress, M. E., Hart, S. G., and Bortolussi, M. R. The reliability and validity of flight task workload ratings. Proceedings, 26th Annual Meeting of the Human Factors Society, Seattle, Washington, October, 1982, 319-323.

Connor, S. A. and Wierwille, W. W. A comparison of pilot workload assessment techniques using a psychomotor task in a moving base aircraft simulator. Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, August, 1981.

Coombs, C. H. A theory of data. New York: John Wiley and Sons, 1964.

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.

Derrick, W. L. The relationship between processing resource and subjective dimensions of operator workload. Proceedings, 25th Annual Meeting of the Human Factors Society, Rochester, NY, October, 1981, 532-536.

Donnell, M. L. The application of decision-analytic techniques to the test and evaluation phase of acquisition of a major air system: Phase III. McLean, Virginia: Decisions and Designs Technical Report, PR-79-6-91, May, 1979.

Dyer, R. F., Matthews, J. J., Wright, C. E., and Yudowitch, K. L. Questionnaire construction manual. Fort Hood, TX: U.S. Army Research Institute for the Behavioral and Social Sciences, Field Unit, TR-P-77-1, July, 1976.

Edvardsson, B. Effect of reversal of response scales in a questionnaire. Perceptual and Motor Skills, 1980, 50, 1125-1126.

Eggemeier, F. T. Some current issues in workload assessment. Proceedings, 24th Annual Meeting of Human Factors Society, 1980, 669-673.

Eggemeier, F. T., Crabtree, S. M., Zingg, J. J., Reid, G. B., and Shingledecker, C. A. Subjective workload assessment in a memory task. Proceedings, 26th Annual Meeting of the Human Factors Society, Seattle, Washington, October, 1982, 643-647.

Gartner, W. B. and Murphy, M. R. Pilot workload and fatigue: A critical survey of concepts and assessment techniques. Washington D.C.: National Aeronautics and Space Administration, Report No., NASA-TN-D-8365, November, 1976.

Gilmer, B. H. and Deci, E. L. Industrial and organizational psychology (4th edition). New York: McGraw-Hill, 1977.

- Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.
- Hallsten, L., and Borg G. Six rating scales for perceived difficulty. Report from the Institute of Applied Psychology, the University of Stockholm, No. 58, 1975.
- Helm, W. R. The reliability of psychometric measures of task difficulty under different levels of information load. University of South Dakota - Human Factors Laboratory, Report No. HFL-81-3, January, 1981.
- Helm, W. R. and Donnell, M. L. Mission operability assessment technique: A system methodology. Fort Mugu, CA: PMTC Report TP-79-31, October, 1979.
- Helm, W. R. and Heimstra, N. W. A preliminary investigation of psychometric measures of task difficulty and task performance. University of South Dakota - Human Factors Laboratory, Report No. HFL-81-2, January, 1981.
- Helmstadter, G. C. Principles of psychological measurement. New Jersey: Prentice-Hall, 1964.
- Hom, P. W., DeNisi, A. S., Kinicki, A. J., and Bannister, B. D. Effectiveness of performance feedback from behaviorally anchored rating scales. Journal of Applied Psychology, 1982, 67(5), 568-576.

- Korman, A. K. Organizational behavior. New Jersey: Prentice-Hall, 1977.
- Keppel, G. Design and analysis: A researcher's handbook. New Jersey: Prentice-Hall, 1973.
- Krantz, D. H., Luce, R. D., Suppes, P. R., and Tversky, A. Foundations of measurement: Volume I. New York: Academic Press, 1971.
- Krantz, D. H. and Tversky, A. Conjoint measurement analysis of composition rules in psychology. Psychological Review, 1971, 78(2), 151-169.
- Kruskal, J. and Wish, M. Multidimensional scaling. Beverly Hills, CA: Sage Publications, 1978.
- Kuznicki, J. T. and Nagle, D. Magnitude estimates of distances along category rating scales. Perceptual and Motor Skills, 1980, 50, 147-150.
- Lodge, M. Magnitude scaling: Quantitative measurement of opinions. Beverly Hills, CA: Sage Publications, 1981.
- Luce, R. D. and Tukey, J. W. Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology, 1964, 1, 1-27.

- Mallamad, S. M., Levine, J. M., and Fleishman, E. A.
Identifying ability requirements by decision flow
diagrams. Human Factors, 1980, 22(1), 57-68.
- McCormick, E. J. Human factors in engineering and design,
Fourth edition. New York: McGraw-Hill, 1976.
- Miller, G. A. The magical number seven, plus or minus two:
Some limits on our capacity for processing information.
Psychological Review, 1956, 63, 81-97.
- Moray, N. (Ed.) Mental workload: Its theory and
measurement. New York: Plenum Press, 1979.
- Moray, N. Subjective mental workload. Human Factors, 1982,
24(1), 25-40.
- Norwich, K. H. The magical number seven: Making a "bit" of
"sense". Perception and Psychophysics, 1981, 29(5),
409-422.
- Osgood, C. E. and Suci, G. J. Factor analysis of meaning.
Journal of Experimental Psychology, 1955, 50(5), 325-338.
- Rahimi, M. Evaluation of the sensitivity and intrusion of
workload estimation techniques in piloting tasks
emphasizing mediational activity. Unpublished Ph. D.
dissertation, Virginia Polytechnic Institute and State
University, Blacksburg, VA, July, 1982.

Rahimi, M. and Wierwille, W. W. Evaluation of the sensitivity and intrusion of workload estimation techniques in piloting tasks emphasizing mediational activity. Proceedings, International Conference on Cybernetics and Society, 1982, 593-597.

Rieger, C. A. Decision tree rating techniques for the assessment of pilot mental workload in a simulated flight task emphasizing mediational behavior. Unpublished master's thesis. Virginia Polytechnic Institute and State University, Blacksburg, Virginia, July, 1983.

Reid, G. B., Eggemeier, F. T., and Nygren, T. E. An individual difference approach to SWAT scale development. Proceedings, 26th Annual Meeting of the Human Factors Society, Seattle, Washington, October, 1982, 639-642.

Reid, G. B., Shingledecker, C. A., and Eggemeier, F. T. Application of conjoint measurement to workload scale development. Proceedings, 25th Annual Meeting of the Human Factors Society, Rochester, New York, October, 1981, 522-526.

Reid, G. B., Shingledecker, C. A., Nygren, T. E., and Eggemeier, F. T. Development of multidimensional

- subjective measures of workload. Proceedings, International Conference on Cybernetics and Society, October, 1981, 403-406.
- Ruggerio, F. T. and Fadden, D. M. Pilot subjective evaluation: A practical supplement to traditional assessments of workload. Proceedings, International Conference on Cybernetics and Society, 1982, 603-607.
- SAS Institute, Inc. SAS user's guide: 1979 edition. Cary, North Carolina, 1982.
- Shaw, M. E. and Wright, J. M. Scales for the measurement of attitudes. New York: McGraw-Hill, 1967.
- Smit, J. and Wewerinke, P. H. An analysis of helicopter pilot control behavior and workload during instrument flying tasks. AGARD-CP-255, May, 1978, 30-1 thru 30-11.
- Smith, P. C. and Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.
- Spector, P. E. Ratings of equal and unequal response choice intervals. Journal of Social Psychology, 1980, 112, 115-119.

Stone, J. D. and Helm, W. R. An investigation of the scaling of mental workload. Phase I: Review of mental workload scaling literature. University of South Dakota - Human Factors Laboratory, Report No. HFL-81-4, January, 1981.

Thurstone, L. L. The measurement of values. Chicago, IL: University of Chicago Press, 1959.

U.S. Government. Code of federal regulations. Title 14: parts 1 to 59, January, 1982, FAR 25.1523, FAA, 367-386.

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

Weiss, D. J. Note on choosing a response scale. Perceptual and Motor Skills, 1979, 50, 472-474.

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. (To appear.)

Williges, R. C. and Wierwille, W. W. Behavioral measures of aircrew mental workload. Human Factors, 1979, 21(5), 549-574.

Wolf, J. D. Workload evaluation of control-display configurations for approach to landing. Minneapolis, Minnesota: Honeywell, SCR Report F0548-IR, August, 1977.

Appendix A

MULTI-DESCRIPTOR RATING SCALE

ATTENTIONAL DEMAND refers to the portion of your total time required (or the amount of attention required) to perform the instructed task.

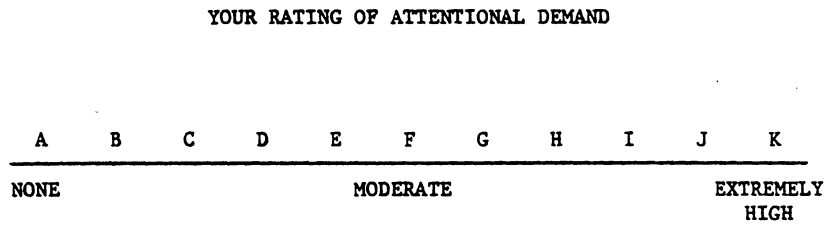


Figure 13: Casali and Wierwille Multi-Descriptor Rating Scale

STRESS LEVEL refers to your emotional reaction while you performed the instructed task. Stress may be considered as your feeling of anxiety, concern, uneasiness and uncertainty brought on as a direct result of performing the instructed task.

YOUR STRESS LEVEL RATING										
A	B	C	D	E	F	G	H	I	J	K
ABSOLUTE CALM				MODERATE STRESS				EXTREME STRESS		

Figure 13: (continued)

OVERLOAD refers to the magnitude and frequency of times you were overloaded by the instructed task. When portions of the task could not be completed because you were too busy, you were overloaded.

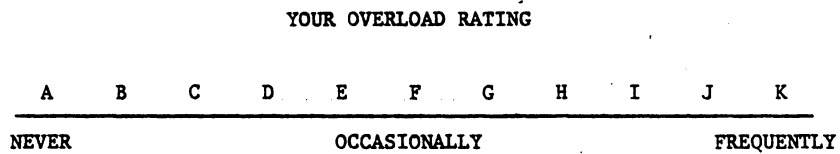


Figure 13: (continued)

DIFFICULTY refers to how hard or difficult you found the instructed task.

YOUR DIFFICULTY RATING										
A	B	C	D	E	F	G	H	I	J	K
EXTREMELY EASY				MODERATE				EXTREMELY DIFFICULT		

Figure 13: (continued)

ERROR LEVEL refers to the magnitude and frequency of mistakes, omissions, incorrect procedures, and incompletions you made in performing the instructed task.

YOUR ERROR LEVEL RATING

A	B	C	D	E	F	G	H	I	J	K
NO ERRORS				MODERATE LEVEL				NUMEROUS AND/OR LARGE ERRORS		

Figure 13: (continued)

TASK COMPLEXITY refers to how complicated or complex you found the instructed task.

YOUR TASK COMPLEXITY RATING										
A	B	C	D	E	F	G	H	I	J	K
EXTREMELY EASY			MODERATE					EXTREMELY COMPLEX		

Figure 13: (continued)

MENTAL WORKLOAD refers to the integrated mental effort required to perform the instructed task. It includes such factors as depth of thinking and level of concentration required by the instructed task.

YOUR MENTAL WORKLOAD RATING										
A	B	C	D	E	F	G	H	I	J	K
EXTREMELY LOW			MODERATE					EXTREMELY HIGH		

Figure 13: (continued)

Appendix B

PHOTOGRAPHS OF EXPERIMENTAL EQUIPMENT AND
STATIONS



Figure 14: Simulator Cockpit--Instrumentation and Labels

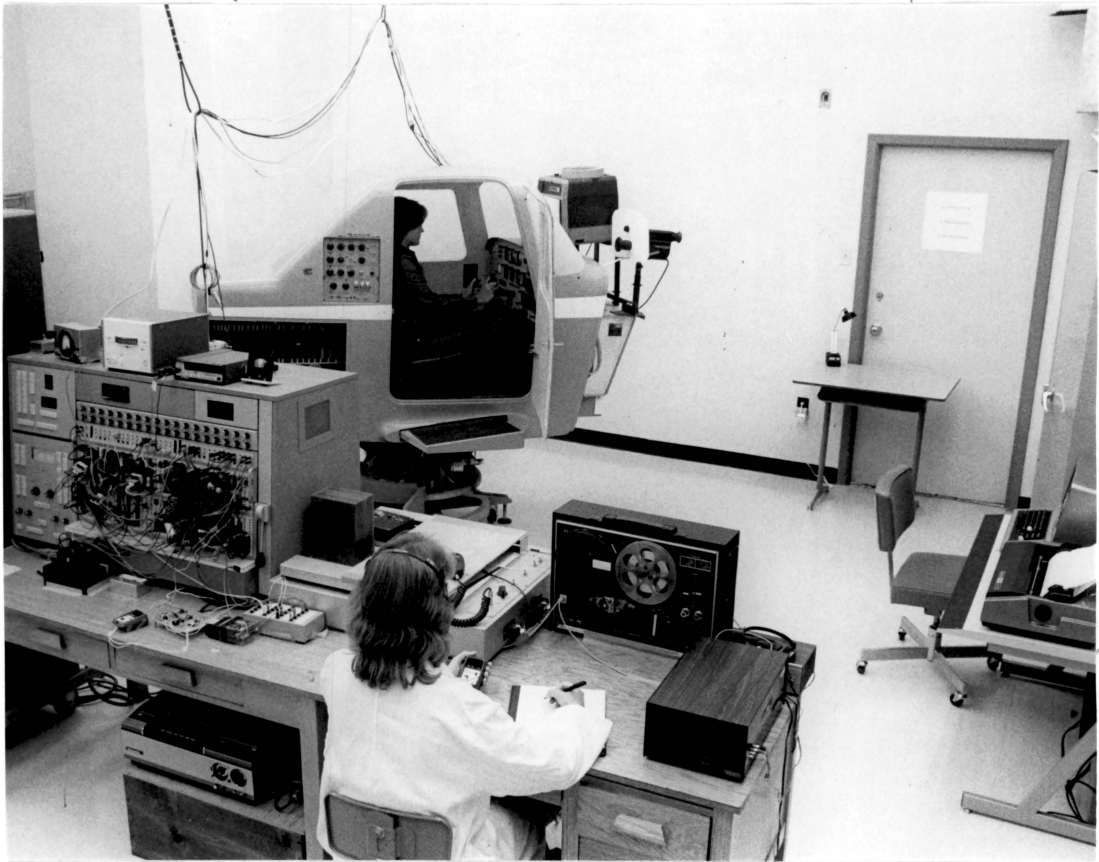


Figure 15: Experimental Station for the Communications Flights

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 16: Photograph of a CRT Screen for the COMPMCH Rating



Figure 17: Photograph of a Pilot Making a Typical Computer Rating

Appendix C
RATING SCALE DESIGNS

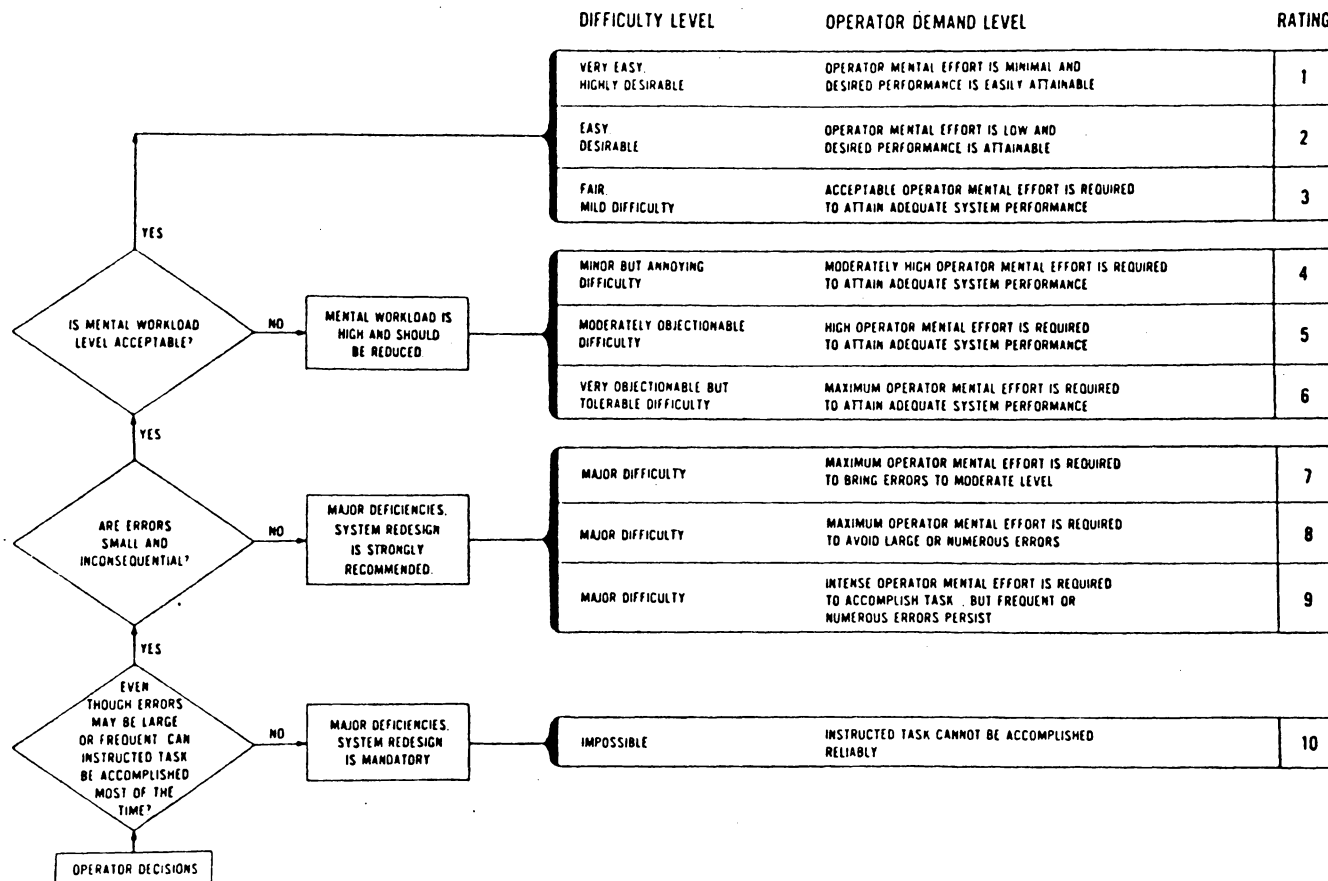


Figure 18: Modified Cooper-Harper (MCH) Rating Scale (56% reduction)

Computerized Modified Cooper-Harper (COMPMCH) Rating Scale

Following is the step-by-step decision process of the COMPMCH rating scale. Each of the divisions are different screens which appear on the TRS-80 CRT screen.

OPERATOR DECISIONS

PRESS ENTER TO CONTINUE

?

EVEN THOUGH ERRORS MAY BE LARGE OR FREQUENT, CAN INSTRUCTED TASK BE ACCOMPLISHED MOST OF THE TIME?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?Y

ARE ERRORS SMALL AND INCONSEQUENTIAL?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?Y

Figure 19: Computerized Modified Cooper-Harper Rating Scale

IS MENTAL WORKLOAD LEVEL ACCEPTABLE?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?N

MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.

PRESS ENTER TO CONTINUE

?

CHOOSE ONE OF THE FOLLOWING ALTERNATIVES:

PRESS ENTER TO CONTINUE

?

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.?

4

Figure 19: (continued)

LET'S REVIEW YOUR DECISIONS

PRESS ENTER TO CONTINUE

?

EVEN THOUGH ERRORS MAY BE LARGE OR FREQUENT, CAN INSTRUCTED
TASK BE ACCOMPLISHED MOST TO THE TIME?
YOUR ANSWER TO THIS DECISION WAS YES.

ARE ERRORS SMALL AND INCONSEQUENTIAL?
YOUR ANSWER TO THIS DECISION WAS YES.

IS MENTAL WORKLOAD LEVEL ACCEPTABLE?
YOUR ANSWER TO THIS DECISION WAS NO.

MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.

PRESS ENTER TO CONTINUE

?

YOUR RATING WAS AS FOLLOWS.

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

-4-

ARE YOU SATISFIED WITH THIS RATING?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?Y

THE RATING WAS 4.

THE RATING SCALE WAS COMPLETED 1 TIME(S).

>READY

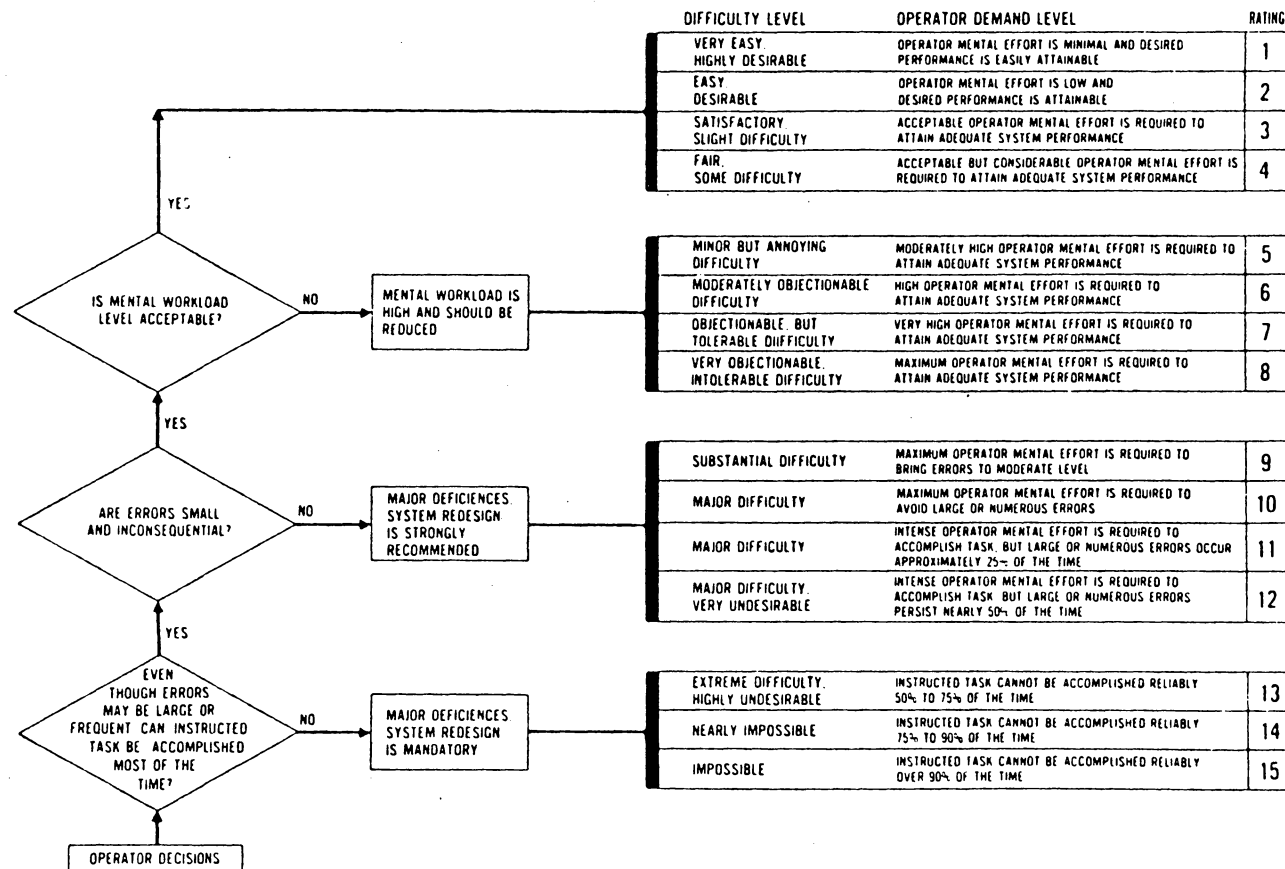


Figure 20: 15 Point Modified Cooper-Harper (MCH+) Rating Scale (54% reduction)

Computerized 15 Point Modified Cooper-Harper (COMPMCH+)
Rating Scale

Following is the step-by-step decision process of the COMPMCH+ rating scale. Each of the divisions are different screens which appear on the TRS-80 CRT screen.

OPERATOR DECISIONS

PRESS ENTER TO CONTINUE

?

EVEN THOUGH ERRORS MAY BE LARGE OR FREQUENT, CAN INSTRUCTED
TASK BE ACCOMPLISHED MOST OF THE TIME?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?Y

ARE ERRORS SMALL AND INCONSEQUENTIAL?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?Y

Figure 21: Computerized 15 Point Modified Cooper-Harper
(COMPMCH+) Rating Scale

IS MENTAL WORKLOAD LEVEL ACCEPTABLE?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?N

MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.

PRESS ENTER TO CONTINUE

?

CHOOSE ONE OF THE FOLLOWING ALTERNATIVES:

PRESS ENTER TO CONTINUE

?

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

MODERATELY OBJECTIONABLE DIFFICULTY: HIGH
OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN
ADEQUATE SYSTEM PERFORMANCE. -6-

OBJECTIONABLE, BUT TOLERABLE DIFFICULTY: VERY
HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO
ATTAIN ADEQUATE SYSTEM PERFORMANCE. -7-

VERY OBJECTIONABLE, INTOLERABLE DIFFICULTY:
MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO
ATTAIN ADEQUATE SYSTEM PERFORMANCE. -8-

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

5

Figure 21: (continued)

LET'S REVIEW YOUR DECISIONS

PRESS ENTER TO CONTINUE

?

EVEN THOUGH ERRORS MAY BE LARGE OR FREQUENT, CAN INSTRUCTED
TASK BE ACCOMPLISHED MOST TO THE TIME?
YOUR ANSWER TO THIS DECISION WAS YES.

ARE ERRORS SMALL AND INCONSEQUENTIAL?
YOUR ANSWER TO THIS DECISION WAS YES.

IS MENTAL WORKLOAD LEVEL ACCEPTABLE?
YOUR ANSWER TO THIS DECISION WAS NO.

MENTAL WORKLOAD IS HIGH AND SHOULD BE REDUCED.

PRESS ENTER TO CONTINUE

?

YOUR RATING WAS AS FOLLOWS.

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

-5-

ARE YOU SATISFIED WITH THIS RATING?

TYPE Y FOR YES OR N FOR NO.
THEN PRESS ENTER.

?Y

THE RATING WAS 5.

THE RATING SCALE WAS COMPLETED 1 TIME(S).

>READY

Figure 21: (continued)

OPERATOR DECISIONS						
	TASK IS ACCOMPLISHABLE MOST OF THE TIME	ERRORS ARE SMALL AND INCONSEQUENTIAL	MENTAL WORKLOAD IS ACCEPTABLE	DIFFICULTY LEVEL	OPERATOR DEMAND LEVEL	RATING
	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 MILD 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 22: Tabular Modified Cooper-Harper (NDT) Rating Scale (54% reduction)

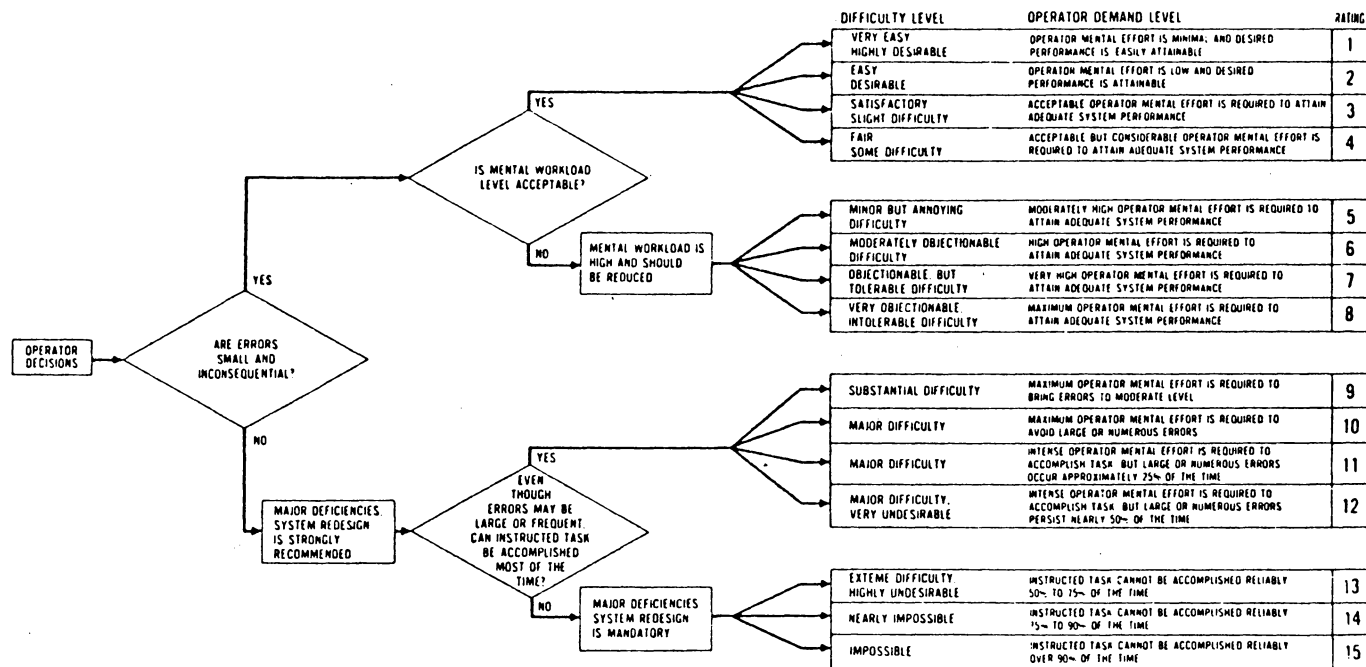


Figure 23: Performance Based Modified Cooper-Harper (PBMCH) Rating Scale (56% reduction)

Appendix D

INTRODUCTION TO THE WORKLOAD EXPERIMENT

INTRODUCTION TO THE WORKLOAD EXPERIMENT

The present study investigates the responses of pilots under varying levels of task difficulty. The research is being conducted at the Vehicle Simulation Laboratory, Department of Industrial Engineering and Operations Research (IEOR), Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 24061, telephone number 961-5358. This research is being sponsored by NASA Ames Research Center, Moffett Field, California. The research team consists of two graduate students, Christine Rieger and Julie Skipper, under the direction of Dr. Walter W. Wierwille, professor in the Department of IEOB.

Your task as a participant in this study is to perform flight tasks in a GAT-1B aircraft simulator, which is often used for pilot training. For the purpose of this study, the simulator will be equipped with some peripheral research equipment to aid in the presentation of the tasks. The added equipment will in no way interfere with the task operation. In the flight task, you will be required to perform some navigational problems involving crosswinds and to perform air traffic control communications. You should fly the simulator as if you were flying a real aircraft.

Participation in this study is entirely voluntary and you have the right to discontinue the experiment at any time. If you choose to participate you will need to attend two experimental sessions of approximately two and one-half hours each. A total of seven simulated flights will be performed in two sessions. Each flight session will consist of a standard takeoff, about 30 minutes of straight, level flying, and a simple approach and landing. In the first session, you will fly one simulated flight for practice, followed by three experimental data flights. You will then be given a short break and the second session of three additional experimental flights will follow. Data will be collected during the six experimental flights. You will receive \$5.00 per hour for the time that you are present, including the practice session and breaks.

We hope that this experiment will be an interesting and enjoyable experience for you. It is possible that at times you may feel frustrated or stressed, and the task may seem difficult. Your performance on the task reflects the difficulty of the task, not your personal abilities and talents. The experiments are not designed to test your skill. After you have completed the experimental sessions, your data will be treated anonymously.

Your participation is greatly appreciated. If you have any questions about the experiment or your rights as a participant, please do not hesitate to ask. We will answer your questions as openly and honestly as possible. Answers to some of your questions may be delayed until after completion of the experimental runs to avoid affecting the outcome of the experiment. Please do not discuss the experiment with other persons, especially pilots who may participate in this study. Since it is expected that all data will be collected by July 1, 1983, you may feel free to discuss the experiment with anyone after that time.

PARTICIPANT'S INFORMED CONSENT

The purpose of this document is to obtain your consent to participate in this experiment and to inform you of your rights as a participant.

(1) You have the right to stop participating in the experiment at any time. If you choose to terminate the experiment, you will receive pay only for the time you participated in the experiment.

(2) You have the right to see your data and to withdraw them from the experiment. If you decide to withdraw your data, please notify the experimenter immediately. Otherwise, identification of your particular data will not be possible, because the data will be treated anonymously.

(3) You have the right to be informed of the overall results of the experiment. If, after participation, you wish to receive information regarding this study, please include your address (four months hence) with your signature below. If more detailed information is desired after receiving the results, please contact the Vehicle Simulation Laboratory, and a full report will be made available to you.

The faculty and graduate students involved in this study will greatly appreciate your help as a participant.

Your signature below indicates you have read the above stated rights and you consent to participate in the study described. If you include your printed name and address below, a summary of the experimental results will be sent to you.

Signature

Printed name and address

Appendix E

RATING SCALE INSTRUCTIONAL SETS

MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on a Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given to you. Before you make any flights, we will review:

1. The definition of the terms used in the scale,
2. The steps you should follow in making your rating on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points please ask me.

Important Definitions

To understand and use the Modified Cooper-Harper Scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "instructed task" is the flight task you have been assigned to perform in this experiment. It includes

flying the aircraft within specified levels of accuracy and performing all duties that are requested of you during the time interval designated by the experimenter.

Second, the "operator" in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the "system" is the complete group of equipment you will be using in performing the instructed task. Together you and the system make up the "operator/system". (For the present experiment, the system is composed of the aircraft simulator, its instruments and controls, and the communications system.)

Fourth, "errors" include any of the following: mistakes, incorrect actions or responses, blunders, and incompletions. In other words, errors are any appreciable deviation from desired "operator/system" performance.

Finally, "mental workload" is the integrated mental effort required to perform the instructed task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the instructed task.

Rating Scale Steps

On the Modified Cooper-Harper Scale you will notice that there is a series of decisions which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. Thus, you should follow the logic sequence on the scale for each of your ratings in the experiment.

The steps which you will follow in using the rating scale logic are as follows:

1. First you will decide if the instructed task can be accomplished most of the time; if not, then your rating is a 10 and you should circle the 10 on the rating scale.
2. Second, you will decide if adequate performance is attainable. Adequate performance means that the errors are small and inconsequential in performing the instructed task. Adequate performance will be defined in the set of instructions you will be given later for the flight task. If they are not, then there are major deficiencies in the system and you should proceed to the right. By reading the descriptions associated with the numbers 7, 8, and 9, you should be able to select the one that best describes the situation you have experienced. You would then circle the most appropriate number.

3. If adequate performance is attainable your next decision is whether or not your mental workload for the instructed task is acceptable. If it is not acceptable, you should select a rating of 4, 5, or 6. One of these three ratings should describe the situation you have experienced, and you would circle the most appropriate number.
4. If mental workload is acceptable, you should then move to one of the top three descriptions on the scale. You would read and carefully select the rating 1, 2, or 3 based on the corresponding description that best describes the situation you have experienced. You would circle the most appropriate number.

Remember you are to circle only one number, and the number should be arrived at by following the logic of the scale. You should always begin at the lower left and follow the logic path until you have decided on a rating. In particular, do not skip any steps in the logic. Otherwise, your rating may not be valid and reliable.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized. First, be sure to try to perform the instructed task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

If you have any questions, please ask the experimenter at this time.

COMPUTERIZED MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on a Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given you. Before you make any flights, we will review:

1. The definitions of the terms used in the scale,
2. The steps you should follow in making your rating on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points, please ask me.

Important Definitions

To understand and use the Modified Cooper-Harper scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "instructed task" is the task you have been assigned in this experiment. It includes flying the aircraft within specified levels of accuracy and performing the duties you have been requested to perform over the time interval designated by the experimenter.

Second, the "operator" in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the "system" is the complete group of equipment you will be using in performing the instructed task. Together you and the system make up the "operator/system". (For the present experiment, the system is composed of the aircraft simulator, its instruments and controls, and the communications system.)

Fourth, "errors" include any of the following: mistakes, incorrect actions or responses, blunders, omissions, and incompletions. In other words, errors are any appreciable deviation from desired operator/system performance.

Finally, "mental workload" is the integrated mental effort required to perform the instructed task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the instructed task.

Rating Scale Steps

A micro-computer will be used to present the Modified Cooper-Harper scale when you make your ratings. The computer program will provide you with instructions for proceeding at every step. You will only need to read what is presented on the screen, type in a response and/or press the enter key to continue.

You will need to become familiar with the following keys: Y, N, numbers 0 - 9, ENTER and the (LEFT ARROW) which is located above the ENTER key. You will use the ENTER key to advance through the program from frame to frame. The LEFT ARROW key will backspace and delete one character each time it is depressed. Use this key if you have made a typing error or a mistake. Try to avoid making errors by carefully typing your responses.

In some of the frames, you will be asked a question and you must type in an option, e.g., Y or N or a number 1 - 10. You must answer every question with an appropriate option and then press ENTER, otherwise the program will not allow you to advance. In other frames, some information will be presented and those frames will instruct you to press ENTER to continue. After you have read those frames, press ENTER.

On the written Modified Cooper-Harper scale (hardcopy version) you will notice that there is a series of decisions

which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. The computerized version of the scale will allow you to follow the logic sequence for each of your ratings in this experiment.

The steps which you will follow in using the rating scale logic are as follows:

1. First you will decide if the instructed task can be accomplished most of the time; if not, then your rating is a 10 and you would choose the 10 on the rating scale.
2. Second, you will decide if adequate performance is attainable. Adequate performance means that the errors are small and inconsequential in performing the instructed task. If they are not, then there are major deficiencies in the system and you should proceed. By reading the descriptions associated with the numbers 7, 8, and 9, you should be able to select the one that best describes the situation you have experienced. You would then choose the most appropriate number.
3. If adequate performance is attainable, your next decision is whether or not your mental workload for the instructed task is acceptable. If it is not

acceptable, you should select a rating of 4, 5, or 6. One of these three ratings should describe the situation you experienced, and you would choose the most appropriate number.

4. If mental workload is acceptable, you should then proceed. You would read and carefully select the rating 1, 2, or 3 based on the corresponding description that best describes the situation you have experienced. You would choose the most appropriate number.

Remember, you will choose only one number, and the number will be arrived at by following the instructions on the computer. After selecting one of the rating alternatives you will have the opportunity to review your decisions. You will decide whether or not you are satisfied with your rating. If for any reason you would like to change your rating, then you would type N and press ENTER. The computer will then return to the beginning of the program and you will repeat the entire process.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized. First, be sure to try to perform the instructed task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

Please feel free to ask any questions about the use of the rating scale.

15 POINT MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on a Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given to you. Before you make any flights, we will review:

1. The definition of the terms used in the scale,
2. The steps you should follow in making your rating on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points please ask me.

Important Definitions

To understand and use the Modified Cooper-Harper Scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "instructed task" is the flight task you have been assigned to perform in this experiment. It includes flying the aircraft within specified levels of accuracy and performing all duties that are requested of you during the time interval designated by the experimenter.

Second, the "operator" in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the "system" is the complete group of equipment you will be using in performing the instructed task. Together you and the system make up the "operator/system". (For the present experiment, the system is composed of the aircraft simulator, its instruments and controls, and the communications system.)

Fourth, "errors" include any of the following: mistakes, incorrect actions or responses, blunders, and incompletions. In other words, errors are any appreciable deviation from desired "operator/system" performance.

Finally, "mental workload" is the integrated mental effort required to perform the instructed task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the instructed task.

Rating Scale Steps

On the Modified Cooper-Harper Scale you will notice that there is a series of decisions which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. Thus, you should follow the logic sequence on the scale for each of your ratings in the experiment.

The steps which you will follow in using the rating scale logic are as follows:

1. First, you will decide if the instructed task can be accomplished most of the time (50% of the time or more); if not, then there are major system deficiencies and system redesign is mandatory. You should proceed to the right. By reading the descriptions associated with the numbers 13, 14, and 15, you should be able to select the one that best describes the situation you experienced. You would then circle the most appropriate number of the three.
2. Second, you will decide if adequate performance is attainable. Adequate performance means that the errors are small and inconsequential in performing the instructed task. Adequate performance will be defined in the set of instructions you will be given later for the flight task. If they are not, then

there are major deficiencies in the system and you should proceed to the right. By reading the descriptions associated with the numbers 9, 10, 11, and 12, you should be able to select the one that best describes the situation you have experienced. You would then circle the most appropriate number.

3. If adequate performance is attainable next decision is whether or not your mental workload for the instructed task is acceptable. If it is not acceptable, you should select a rating of 5, 6, 7, or 8. One of these four ratings should describe the situation you have experienced, and you would circle the most appropriate number.
4. If mental workload is acceptable, you should then move to one of the top four descriptions on the scale. You would read and carefully select the rating 1, 2, 3, or 4 based on the corresponding description that best describes the situation you have experienced. You would circle the most appropriate number.

Remember you are to circle only one number, and the number should be arrived at by following the logic of the scale. You should always begin at the lower left and follow the logic path until you have decided on a rating. In

particular, do not skip any steps in the logic. Otherwise, your rating may not be valid and reliable.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized. First, be sure to try to perform the instructed task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves

are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

If you have any questions, please ask the experimenter at this time.

COMPUTERIZED 15 POINT MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on a Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given you. Before you make any flights, we will review:

1. The definitions of the terms used in the scale,
2. The steps you should follow in making your rating on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points, please ask me.

Important Definitions

To understand and use the Modified Cooper-Harper scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "instructed task" is the task you have been assigned in this experiment. It includes flying the aircraft within specified levels of accuracy and performing the duties you have been requested to perform over the time interval designated by the experimenter.

Second, the "operator" in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the "system" is the complete group of equipment you will be using in performing the instructed task. Together you and the system make up the "operator/system". (For the present experiment, the system is composed of the aircraft simulator, its instruments and controls, and the communications system.)

Fourth, "errors" include any of the following: mistakes, incorrect actions or responses, blunders, omissions, and incompletions. In other words, errors are any appreciable deviation from desired operator/system performance.

Finally, "mental workload" is the integrated mental effort required to perform the instructed task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the instructed task.

Rating Scale Steps

A micro-computer will be used to present the Modified Cooper-Harper scale when you make your ratings. The computer program will provide you with instructions for proceeding at every step. You will only need to read what is presented on the screen, type in a response and/or press the enter key to continue.

You will need to become familiar with the following keys: Y, N, numbers 0 - 9, ENTER and the (LEFT ARROW) which is located above the ENTER key. You will use the ENTER key to advance through the program from frame to frame. The LEFT ARROW key will backspace and delete one character each time it is depressed. Use this key if you have made a typing error or a mistake. Try to avoid making errors by carefully typing your responses.

In some of the frames, you will be asked a question and you must type in an option, e.g., Y or N or a number 1 - 15. You must answer every question with an appropriate option and then press ENTER, otherwise the program will not allow you to advance. In other frames, some information will be presented and those frames will instruct you to press ENTER to continue. After you have read those frames, press ENTER.

On the written Modified Cooper-Harper scale (hardcopy version) you will notice that there is a series of decisions

which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. The computerized version of the scale will allow you to follow the logic sequence for each of your ratings in this experiment.

The steps which you will follow in using the rating scale logic are as follows:

1. First you will decide if the instructed task can be accomplished most of the time (50% of the time or more); if not, then there are major system deficiencies and system redesign is mandatory. By reading the descriptions associated with the numbers 13, 14, and 15, you should be able to select the one that best describes the situation you experienced. You would choose the most appropriate number of the three.
2. Second, you will decide if adequate performance is attainable. Adequate performance means that the errors are small and inconsequential in performing the instructed task. Adequate performance will be defined in the set of instructions you will be given later for the flight task. If they are not, then there are major deficiencies in the system and you should proceed. By reading the descriptions

associated with the numbers 9, 10, 11, and 12, you should be able to select the one that best describes the situation you have experienced. You would then choose the most appropriate number.

3. If adequate performance is attainable, your next decision is whether or not mental workload for the instructed task is acceptable. If it is not acceptable, you should select a rating of 5, 6, 7, or 8. One of these four ratings should describe the situation you have experienced, and you would choose the most appropriate number.
4. If mental workload is acceptable, you should then proceed. You would read and carefully select the rating 1, 2, 3, or 4 based on the corresponding description that best describes the situation you have experienced. You would choose the most appropriate number.

Remember, you will choose only one number, and the number will be arrived at by following the instructions on the computer. After selecting one of the rating alternatives you will have the opportunity to review your decisions. You will decide whether or not you are satisfied with your rating. If for any reason you would like to change your rating, then you would type N and press ENTER.

The computer will then return to the beginning of the program and you will repeat the entire process.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized. First, be sure to try to perform the instructed task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings

which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

Please feel free to ask any questions about the use of the rating scale.

TABULAR MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on a Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given to you. Before you make any flights, we will review:

1. The definition of the terms used in the scale,
2. The steps you should follow in making your rating on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points please ask me.

Important Definitions

To understand and use the Modified Cooper-Harper Scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "instructed task" is the flight task you have been assigned to perform in this experiment. It includes flying the aircraft within specified levels of accuracy and performing all duties that are requested of you during the time interval designated by the experimenter.

Second, the "operator" in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the "system" is the complete group of equipment you will be using in performing the instructed task. Together you and the system make up the "operator/system". (For the present experiment, the system is composed of the aircraft simulator, its instruments and controls, and the communications system.)

Fourth, "errors" include any of the following: mistakes, incorrect actions or responses, blunders, and incompletions. In other words, errors are any appreciable deviation from desired "operator/system" performance.

Finally, "mental workload" is the integrated mental effort required to perform the instructed task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the instructed task.

Rating Scale Steps

On the Modified Cooper-Harper Scale you will notice that the scale contains six columns and ten rows. Each of the rows represents a specific rating category. Each of the ten rating categories is described by a set of five descriptors. For example: rating number 4's descriptors are:

			minor but	moderately high operator
yes	yes	no	annoying	mental effort is required
			difficulty	to attain adequate system
				performance

These descriptors correspond to their respective column headings. The column headings help to identify factors which you should have experienced while performing the instructed task, and which need to be addressed when rating. If a dash appears in any of the columns, as in rating number 10, this means that this column factor is not applicable to the rating. Thus, it should not be considered in the row selection of that rating.

To rate on the Modified Cooper-Harper Scale, you must agree on a single row of descriptors. To aid you in your

row selection process additional descriptive information is provided on the left side of the rating scale for groups of rating categories. The steps which you will follow in using the rating scale are as follows:

1. First, enter the scale and agree on a row of descriptors which best describes the situation you experienced.
2. After selecting the proper row move to the right of the scale and circle the number associated with that row.

Remember you are to circle only one number.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized. First, be sure to try to perform the instructed task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by

the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

If you have any questions, please ask the experimenter at this time.

PERFORMANCE BASED MODIFIED COOPER-HARPER RATING SCALE INSTRUCTIONS

Overview

After each of the following flights, you will be asked to give a rating on a Modified Cooper-Harper Scale for workload. This rating scale and important definitions for using the scale are shown on the sample which I have given you. Before you make any flights, we will review:

1. The definitions of the terms used in the scale,
2. The steps you should follow in making your rating on the scale, and
3. How you should think of the ratings.

If you have any questions as we review these points, please ask me.

Important Definitions

To understand and use the Modified Cooper-Harper scale properly, it is important that you understand the terms used on the scale and how they apply in the context of this experiment.

First, "instructed task" is the task you have been assigned in this experiment. It includes flying the aircraft within specified levels of accuracy and performing the duties you have been requested to perform over the time interval designated by the experimenter.

Second, the "operator" in this situation is you. Because the scale can be used in different situations, the person performing the ratings is called an operator. You will be operating the system and then using the rating scale to quantify your experience.

Third, the "system" is the complete group of equipment you will be using in performing the instructed task. Together you and the system make up the "operator/system". (For the present experiment, the system is composed of the aircraft simulator, its instruments and controls, and the communications system.)

Fourth, "errors" include any of the following: mistakes, incorrect actions or responses, blunders, omissions, and incompletions. In other words, errors are any appreciable deviation from desired operator/system performance.

Finally, "mental workload" is the integrated mental effort required to perform the instructed task. It includes such factors as level of attention, depth of thinking, and level of concentration required by the instructed task.

Rating Scale Steps

On the Modified Cooper-Harper scale you will notice that there is a series of decisions which follow a predetermined logical sequence. This logic sequence is designed to help you make more consistent and accurate ratings. Thus, you should follow the logic sequence on the scale for each of your ratings in this experiment.

The steps which you will follow in using the rating scale logic are as follows:

1. First, you will decide if adequate performance is attainable. Adequate performance means that the errors are small and inconsequential in performing the instructed task. If they are not, then there are major deficiencies in the system and you would then proceed to Step 2.
2. Second, you would then decide if the instructed task can be accomplished most of the time. If not, then you will choose among the three alternatives which will be presented (numbers 13, 14, or 15), and circle the most appropriate one. If the task is accomplishable most of the time, then you will circle one of the four alternatives presented (numbers 9, 10, 11, or 12).

3. If you answered that adequate performance was attainable in Step 1, then your next decision is whether or not your mental workload for the instructed task is acceptable. If it is not acceptable, you should circle one of the four alternatives presented (numbers 5, 6, 7, or 8). One of these ratings should describe the situation you have experienced, and you would choose the most appropriate number.
4. If mental workload is acceptable, you should then choose one of the four alternatives presented (numbers 1, 2, 3, or 4). You would read and carefully select a rating based on the corresponding description that best describes the situation you have experienced. You would circle the most appropriate number.

Remember, you are to circle only one number, and the number should be arrived at by following the logic of the scale. You should always begin at the left and follow the logic path until you have decided on a rating. In particular, do not skip any steps in the logic. Otherwise your rating may not be valid and reliable.

How You Should Think of the Rating

Before you begin making ratings there are several points that need to be emphasized. First, be sure to try to perform the instructed task as instructed and make all your evaluations within the context of the instructed task. Try to maintain adequate performance as specified for your task.

Second, the rating scale is not a test of your personal skill. On all of your ratings, you will be evaluating the system for a general user population, not yourself. You may assume you are an experienced member of that population. You should make the assumption that problems you encounter are not problems you created. They are problems created by the system and the instructed task. In other words, don't blame yourself if the system is deficient, blame the system.

Third, try to avoid the problem of nit picking an especially good system, and of saying that a system which is difficult to use is not difficult to use at all. These problems can result in similar ratings for systems with quite different characteristics. Also, try not to overreact to small changes in the system. This can result in ratings which are extremely different when the systems themselves are quite similar. Thus, to avoid any problems, just always try to "tell it like it is" in making your ratings.

Please feel free to ask any questions about the use of the rating scale.

Appendix F
COMMUNICATIONS TASK INSTRUCTIONS

COMMUNICATIONS

(FOR PRACTICE AND DATA FLIGHTS)

Flight procedures and objectives are listed below. For the practice flight (occurring first) you are to follow instructions 1 through 13. Between each subsequent data flight, the simulator will be put in autopilot and you will not have to land or takeoff again. Therefore, for each data flight you are to follow only instructions 11 through 13.

1. Assume that the airplane is situated on a typical runway. The field elevation is 180 feet and the barometric pressure is 29.78 inches Hg. You will be positioned at a heading of 0 degrees (due North) for takeoff purposes and to follow during flight, unless a new specified heading is presented over the cockpit speaker.
2. The experimenter will inform you over the cockpit speaker when you are cleared for takeoff. Takeoff speed in the flight simulator is 75 mph. Remember to

release the parking brake and check that flaps are up prior to takeoff.

3. After takeoff, climb straight out at approximately 75 mph, the speed at which the maximum rate of climb is realized. Continue climbing to 2000 feet, level off, and fly at a constant speed of 100 mph. After leveling off at 2000 feet and attaining an airspeed of 100 mph, you should trim the aircraft to a straight and level attitude using the thumbwheel on the lower right portion of the dash panel. You will begin to encounter mild turbulence at this point in your flight.
4. Once you feel that you have the aircraft trimmed at 100 mph, at a steady altitude of 2000 feet, and at a steady heading of 0 degrees (due North), inform the experimenter. When you wish to transmit, depress the black microswitch on the yoke at your left thumb and speak. When the black microswitch is depressed, your radio is in the transmit-only mode and you are unable to receive. Press the black microswitch only when you wish to transmit. Your voice is picked up by the lapel microphone you are wearing. The experimenter's transmission will be heard over the cockpit speaker.

5. Your objective as a pilot in this experiment is to perform adequately on all aspects of a instructed task. There are two major aspects of this instructed task: the communications aspect and the aircraft control aspect.
6. For the aircraft control aspect, you are to fly the aircraft at a constant altitude of 2000 feet, constant heading of 0 degrees, and constant airspeed of 100 mph, unless instructed otherwise. During your flight you may receive instruction commands to change altitude or headings. If so, follow the instructions. You should try to maintain adequate performance on the aircraft control aspect throughout the total duration of the flight (approximately 10 minutes). Adequate performance is defined as:
 - 1) Maintaining instructed altitude within +100 feet,
 - 2) Maintaining instructed heading within +5 degrees,
 - 3) Maintaining instructed airspeed within +10 mph.
7. You should continue to perform all aspects of the instructed task until you hear the words, "END TASK." At this point the experimenter will put the aircraft

in autopilot mode. You will then be asked to make a rating of the flight. Subsequently, you will have a chance to relax for a few moments prior to the next flight.

8. There will be one practice flight and three data flights. At the end of the third data flight, the experimenter will tell you to descend. Reduce your airspeed to 75 mph and lower one half flaps. Descend at a rate of 500 feet per minute at the same heading until you make a landing. Field elevation is 180 feet, and barometric pressure at the field is 29.78 inches Hg.
9. Upon landing, you should pull the parking brakes, leave the engine on. Then, you will be instructed to exit the simulator.
10. If you have any questions, please ask the experimenter at this time.

COMMUNICATIONS TASK INSTRUCTIONS

11. Soon after you have trimmed the aircraft, you will hear the words "THE COMMUNICATIONS TASK WILL NOW BEGIN." This signals the start of the communications task. From this point on you will hear a tape-recorded voice over the cockpit speaker and you are to transmit only when responding to this tape-recorded voice. For the communications task you will hear tape-recorded commands over the cockpit speaker. Do exactly what the commands specify and nothing more. For instance, do not transmit to acknowledge receipt of a command. Each command will be preceded by the complete, unabbreviated call sign of your aircraft, which is Cessna One Four India Echo. Note that the word "Cessna" will always precede "One Four India Echo" just prior to a command. The call sign appears on the instrument panel for easy reference. The commands themselves contain instruction to change heading, change altitude, report present altitude and heading, report aircraft model (Cessna 150) and call sign (Cessna One

Four India Echo), change altimeter setting, change radio frequency, and report airspeed. For commands which request you to make a transmission, always depress the black microswitch and speak while holding the switch down. Remember that while the black switch is depressed, you are unable to receive transmission; therefore, only depress the black switch when speaking.

12. Between tape-recorded messages you will hear a series of abbreviated aircraft call signs over the cockpit speaker. Each abbreviated call sign will consist of two numbers and two phonetic letters, in any order, such as Six-Alpha-Niner-Foxtrot, Bravo-Zulu-One-Seven, or Two-Five-Xray-Tango. Your objective is to depress the black microswitch and simply say the word "now" whenever you hear one of the following abbreviated call signs:

One-Four-India-Echo

One-Four-Echo-India

One-Echo-Four-India

One-Echo-India-Four

One-India-Echo-Four

One-India-Four-Echo.

Note that there must be a "One" in the first position, followed by any combination of Four-India-Echo in the second, third, and fourth positions. The card on the simulator dash panel is to remind you of the proper call signs to which you should respond. Respond only to these call signs and no others. Be sure to depress the black microswitch and respond as quickly as possible with the word "now." Please strive for maximum accuracy and speed. At some time you may feel that the task is too difficult for your ability. Do not be concerned about this because the task is designed as such. Simply do your best at all times.

In some instances, a series of call signs will be strung together, back-to-back, with pause between adjacent call signs. Remember that all call signs consist of four alphanumerics: two numbers and two phonetic letters. When several call signs are strung together, it is your responsibility to know where one call sign ends and where the next call sign begins. For instance, given the following three strings of call signs, you should respond as indicated by the arrow:

- 1) "Eight Two Alpha Echo One Four India Bravo One
India Echo Four"
"Now"
- 2) "Mike Victor Niner One Four India Echo One India
Four One Echo"
(no reponse because correct sequence is out of place)
- 3) "Lima Three Charlie One Echo India Four One One
Four India Echo Three Foxtrot Yankee Niner One
Hotel Sierra Five"
"Now"

13. While attending to the abbreviated call sign task, you are also to carry out the commands given over the cockpit speaker (if any) and furthermore to fly the aircraft within the specified levels of adequate performance. Remember that you will always be addressed by the call sign "Cessna One Four India Echo" prior to a command. Conversely, the abbreviated call sign sequences you are to listen and respond to will always begin with a number or phonetic letter and will never use the word "Cessna."

You should try to perform the communications task as accurately as possible while also accurately flying the aircraft and carrying out the commands.

Remember, your task is as follows:

1. Respond correctly to all target call signs,
and,
2. Do not respond to all other call signs.

(While you carry out commands from the "tower".)

If you have any questions, please ask the
experimenter at this time.

Appendix G
QUESTIONNAIRE

QUESTIONNAIRE

Please describe in a sentence or two the factors that influenced your rating selection. (Example: There was inadequate time to perform the instructed task.)

Appendix H

Z-SCORE TRANSFORMATION EQUATIONS FOR MEAN SCORES

TABLE 12

Z-score Transformation Equations for Mean Scores

Transformation Equations for transforming the mean z-scores for each of the rating scales to mean raw scores for the low, medium, and high loading levels.

<u>Rating Scale</u>	<u>Equation</u> <u>x</u>	<u>s.d.</u>
MCH	$X = 5.6111 + 2.2528(Z)$	
COMPMCH	$X = 4.0000 + 2.6122(Z)$	
MCH+	$X = 7.2222 + 2.9014(Z)$	
COMPMCH+	$X = 8.0000 + 3.7885(Z)$	
NDT	$X = 4.2222 + 2.3403(Z)$	
PBMCH	$X = 7.5556 + 2.4307(Z)$	

Appendix I

UNEQUAL-N ANOVA SUMMARY TABLE FOR SENSITIVITY
ANALYSIS

TABLE 13

Unequal-N ANOVA Summary Table for Sensitivity Analysis

<u>Source</u>	<u>df</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Between-Subjects				
Rating Scale (RS)	5	0.0000	0.00	1.0000
Subjects (S/RS)	36	1.5502		
Within-Subjects				
Load (L)	2	15.7379	42.75	0.0001
L x RS	10	0.3166	0.86	0.5738
L x S/RS	<u>72</u>	0.3681		
Total	125			

**The three page vita has been
removed from the scanned
document. Page 1 of 3**

**The three page vita has been
removed from the scanned
document. Page 2 of 3**

**The three page vita has been
removed from the scanned
document. Page 3 of 3**