Improved Qualitative Flight Data Rating Scales

C. VANCE SHUFELDT* AND DONALD M. LAYTON†
Naval Postgraduate School, Monterey, Calif.

Various rating scales have been derived to assist in the quantification flight test data. Several of these scales currently in use are examined and compared with emphasis on transference from scale to scale. The problem of nonlinearity of scales that depend upon adjective descriptions is considered, and a hypothesis is advanced that a rater may transpose his impressions of performance directly to any scale without recourse to adjective descriptions and thereby relate his mentally derived, divisionless scale directly to a numerical index. Experimental data, though limited, tend to support this hypothesis.

Introduction

WITH the advent of flight vehicles with operating envelopes ranging from terra firma to the threshold of space and beyond, the environmental and dynamic spectrums encountered on a single flight are all-encompassing. Man is the low-frequency response component[1] in the over-all closed-loop man-machine system, therefore, control systems must be designed within manageable limits. In short, the effort expended in vehicle control must be minimized so that the pilot may be free to complete other duties in the cockpit.

Consequently, the suitability of a machine system to serve its intended mission is ultimately determined by a series of evaluations. The most difficult of these assessments occurs at the man-machine system interface.

Pilot evaluation of handling qualities determines the suitability of the machine system, yet there remains to be found a set of universally acceptable parameters for this evaluation. The complete nature of a pilot's task, work load, mental stress and acuity have not been described in any form of analytically determined transfer function or performance index.[2] It is assumed, however, that there exists a relationship between pilot comment and performance and/or vehicle handling qualities.

Efforts to standardize the qualitative aspects of language into a quantitative handling quality rating have been made. It is the purpose of this study to examine and compare the rating scales presently in use and to investigate the possibilities of a linear rating scale with its inherent advantages.

History

Early Developments

During the early 1930's when aviation was maturing, the need to delineate acceptable aircraft parameters was recognized. Consequently, a "check list" for this purpose was proposed by Edward P. Warner.[3] Subsequent work by Soulé and by R. R. Gilruth at the Langley Laboratory of NACA condensed these requirements and a set of specifications for military aircraft acceptance eventually resulted.[4]

After this initial break-through in establishing aircraft specifications, emphasis was placed on devising pilot opinion ratings aimed at specific problem areas. The concept of a general pilot rating received little attention.

Received December 13, 1971; revision received March 23, 1972.
Index category: Aircraft Testing.

* Graduate Student, Department of Aeronautics; also Lieutenant Commander, U.S. Navy.
† Associate Professor, Department of Aeronautics. Associate Fellow AIAA.

Cooper Scale

In 1957 at the annual meeting of the Flight Testing Session, Institute of Aeronautical Sciences, NACA Ames Chief Research Pilot G. E. Cooper introduced a generalized pilot rating scale which enjoyed immediate and almost total acceptance.[5] This epoch scale (Fig. 1) synthesized the previous work of NACA Langley and thereby provided an authenticated scale which could be applied to any aircraft handling qualities evaluation. It was the first rating scale to associate the qualitative nature of pilot opinion with a quantitative index.

In applying this scale, it was recommended that the evaluator pay particular attention to question formulation. The question had to be sufficiently specific so as to minimize interpretation and ambiguity.

The pilot, in answering the question, was required to channel his exposure, sensations and reactions into the scale vocabulary by first considering four handling qualities categories: Satisfactory, Unsatisfactory, Unacceptable, and Unprintable. As may be noted from Fig. 1, these categories were separated, for description purposes, at the approximate values 3.5, 6.5, and 9.5, respectively. Within each category, the pilot was required to further define his opinion in terms of the scale vocabulary and a secondary mission (landing).

Once the pilot had formulated his opinion with respect to the scale, his evaluation had to be weighted in consideration of his viewpoint, experience and adaptability. For example, a patrol pilot might evaluate the stall-associated buffet and departure in a fighter as "Unacceptable-Dangerous" (numerical rating 8); whereas, a fighter pilot might evaluate the same characteristics as "Satisfactory, but with some unpleasant characteristics" (numerical rating 3). Then, with some exposure, the same two pilots might reevaluate the characteristics at 4 and 2, respectively. The rating scale was, therefore,

ADJECTIVE RATING	NUMERICAL RATING	DESCRIPTION
SATISFACTORY	1	EXCELLENT, INCLUDES OPTIMUM
	2	GOOD, PLEASANT TO FLY
	3	SATISFACTORY, SOME MILDLY UNPLEASANT CHARACTERISTICS
UNSATISFACTORY	4	ACCEPTABLE, BUT WITH UNPLEASANT CHARACTERISTICS
	5	UNACCEPTABLE FOR NORMAL OPERATION
	6	ACCEPTABLE FOR EMERGENCY CONDITION ONLY
UNACCEPTABLE	7	UNACCEPTABLE EVEN FOR EMERGENCY CONDITIONS
	8	UNACCEPTABLE - DANGEROUS
	9	UNACCEPTABLE - UNCONTROLLABLE
UNPRINTABLE	10	MOTIONS POSSIBLY VIOLENT ENOUGH TO PREVENT PILOT ESCAPE

Fig. 1 Cooper scale.

very subject to experience and adaptability. To eliminate this deficiency and to provide some measure of consistency, it was suggested that the scale be used only by test pilots.

Though the Cooper Scale was widely used by virtue of being first in the field, it was somewhat ambiguous in its definitions and was complicated in that it placed stipulations on pilot opinion. It would appear that the scale was designed to evaluate a broad spectrum of handling qualities and not fine points.

Harper Scale

R. P. Harper Jr. used a pilot opinion scale (Fig. 2) for evaluating the handling qualities of a variable stability aircraft in 1959.[6] The Harper Scale was developed honoring the stipulations of question formulation but with a concept quite different from the Cooper Scale. Harper was interested in evaluating pilot-vehicle performance, but found this extremely difficult because of pilot adaptability. Instead, a scale was devised to evaluate pilot opinion with respect to alterations in the stability derivatives and thereby arrive at a pilot preference: a most suitable aircraft stability.

To ensure reliability and compensate for scale vocabulary deficiencies, test pilots wire-recorded their subjective comments during the evaluation and recorded their scale rating following each evaluation. This was, perhaps, the best aspect of the testing procedure. The pilot rating was kept simple and subordinate to the subjective evaluation. Because of this reliance on subjective comments made during the tests, the pilot rating was utilized as a cursory index to the evaluation and not as an end in itself.

In evaluating the handling qualities with respect to the rating scale, the pilot considered four handling qualities categories: Acceptable and Satisfactory, Acceptable but Unsatisfactory, Unacceptable, and Unflyable. The separation between these categories occurred at 3.5, 6.5, and 9.5, respectively. Within each category, the pilot further defined his opinion in terms of a single, though sometimes ambiguous, adjective (Fig. 2).

Harper Scale Adaptations

In contrast to the Cooper Scale, the Harper Scale (often cited as Cornell or CAL Scale because of its extensive use by Cornell Aeronautical Laboratory Inc.) was designed as an index for evaluating particular handling qualities restricted by the nature of the tests being conducted. Efforts to adapt this Scale to the evaluation of aggregate handling qualities met with varied success.

One such example was the application made by M. L. Parrag[7] in 1967 in studying the effects on handling qualities of higher-order response characteristics against a background of varying conditions and associated mission tasks.

To facilitate more reliable and consistent pilot comments, the test pilots were provided with a comment check list for the two flight conditions, and instructed to make subjective comments following each test run. After all tasks were completed, a comprehensive subjective report was required incorporating all the salient features of each configuration. Finally, an objective report using the comment check list was made.

Here, as in Ref. 4, emphasis was placed on subjective comments. Task-oriented objective comments were used to provide consistency and point out features of each task which might otherwise have been over-looked. Although the CAL Scale was used as an index to pilot opinion, it was, for all practical purposes, insignificant in evaluating the handling qualities investigated.

Cooper-Harper Scale

With wide and independent usage of the Cooper and Harper Scales, the problems cited for each were sources of confusion in application. It became increasingly apparent that an acceptable composite rating system incorporating the best features of each scale would be advantageous.

To this end Cooper and Harper jointly advanced a revised rating scale in 1966.[8] This scale (Fig. 3), hereafter referred to as the Cooper-Harper Scale, enjoyed general acceptance and preference over the previous scales; however, the various implementing institutions voiced a need for clarification in semantics and in application. In 1969 an explicitly comprehensive joint report was published to modify and clarify the Cooper-Harper Scale.[9] The report precisely defined flight evaluation terminology and discussed the aspects of question formulation and scale data application.

Based on the voluminous data and comments available from international audiences of the Cooper and Harper Scales, the Cooper-Harper Scale was excellently designed as a two-part procedure of evaluation. A pilot, in evaluating a handling quality, systematically chose between two alternatives which channeled his consideration into a rating category or into another two-part decision with the same channeling result. Through this simplified procedure (compare with the relative complexity of previously discussed procedures) three of four existing categories were eliminated without ever considering the applicable descriptive adjectives.

The inverted ten-point scale was retained in the interests of consistency. Although an ordinal sequence increasing in magnitude with the degree of "goodness" may have been more appropriate, users of the previous scales had become accustomed to the inverted scale and a reordering of the numerical indices would have resulted in unnecessary confusion. To further ease the transition from previous scales, the boundaries of 3.5, 6.5, and 9.5 were retained.

CATEGORY	ADJECTIVE DESCRIPTION WITHIN CATEGORY	NUMERICAL RATING
ACCEPTABLE AND SATISFACTORY	EXCELLENT	1
	GOOD	2
	FAIR	3
ACCEPTABLE BUT UNSATISFACTORY	FAIR	4
	POOR	5
	BAD	6
UNACCEPTABLE	BAD*	7
	VERY BAD**	8
	DANGEROUS***	9
UNFLYABLE	UNFLYABLE	10

*REQUIRES MAJOR PORTION OF PILOT'S ATTENTION
**CONTROLLABLE ONLY WITH MINIMUM OF OTHER DUTIES
***CONTROLLABLE ONLY WITH COMPLETE ATTENTION

Fig. 2 Harper scale.

ADEQUACY FOR SELECTED TASKS	NUMERICAL RATING	AIRCRAFT CHARACTERISTICS
SATISFACTORY	1	EXCELLENT, HIGHLY DESIRABLE
	2	GOOD, NEGLIGIBLE DEFICIENCIES
	3	FAIR, SOME MILDLY UNPLEASANT DEFICIENCIES
DEFICIENCIES WARRANT IMPROVEMENT	4	MINOR BUT ANNOYING DEFICIENCIES
	5	MODERATELY OBJECTIONABLE DEFICIENCIES
	6	VERY OBJECTIONABLE BUT TOLERABLE DEFICIENCIES
DEFICIENCIES REQUIRE IMPROVEMENT	7	MAJOR DEFICIENCIES CONTROLLIBILITY NOT IN QUESTION
	8	MAJOR DEFICIENCIES, REQUIRES CONSIDERABLE PILOT COMPENSATION
	9	MAJOR DEFICIENCIES, REQUIRES INTENSE PILOT COMPENSATION
IMPROVEMENT MANDATORY	10	MAJOR DEFICIENCIES CONTROL WILL BE LOST

Fig. 3 Cooper-Harper scale.

It would appear that a satisfactory method for assessing the man-machine interface had been achieved; but not quite. Although the Cooper-Harper continues to be the most widely used evaluation system to date, it remains insensitive at the bad end and does not exhibit the desirable feature of linearity. Linearity is that feature of a rating scale which will allow the averaging of data ensembles without distorting the data sample interpretation.

McDonnell Scale

In 1968, J. D. McDonnell published his study of rating techniques[10] with the objective of evolving a rating scale which has an underlying linear structure to facilitate mathematical operations on the rater's data. This underlying structure was an intervaled psychological continuum.

If an objective measure is made upon some object, the resulting data must lie along some physically continuous values. On the other hand, if an evaluator estimates a measure with only the end points defined, e.g., bad and good, the measure is subjective and must lie along some psychological continuum. The relationship between these two continua, if it could be determined, would provide a means for linearizing the subjective scale.

To establish an intervaled psychological continuum, a list of sixty-four appropriately descriptive phrases was randomly submitted to sixty-three raters by McDonnell. For each phrase, the raters were instructed to indicate their impression of a hypothetical vehicle so described on a plot with the end points of "most favorable" and "least favorable." The data were then processed by the methods of psychophysics and successive intervals and assigned to relative standing on a scale of nine. The data were further reduced to the arbitrary seven-point McDonnell Scale depicted in Fig. 4.

The McDonnell Scale (often called the Global Rating Scale because it related aggregate handling qualities) was, therefore, presumed to be a linear scale reflecting the ability of the raters to distinguish and resolve sematic differences. Because it was related to a seven-point scale in contrast with the ten-point scales with which the users were familiar, it was not accepted with any noticeable exuberance.

The most important contribution made by McDonnell was the list of evaluation phrases related to an index of nine (the seven-point scale plus two end points) and reflecting psychological sensitivity. The phrases were divided into six categories: Handling Qualities, Control, Precision, Response Characteristics, Effects of Deficiencies, and Demands on Pilot. Through the use of this listing, specialized linear scales may be constructed to satisfy particular rating requirements.

Contemporary Research

In designing the washout circuitry for the NASA Ames All-Axis Motion Generator, it became a necessary expedient to solicit pilot opinion in determining the "best" set of parameters to use in a given configuration. To this end, S. F. Schmidt and Bjorn Conrad[11] use three nonordinal, relative rating scales in their evaluations (Fig. 5 a, b and c).

The questions related to each scale were particularly tailored to the descriptive adjectives shown and they were simple in nature. By using pilot comments as an index, the design providing the best over-all simulator characteristics was obtained. However, moderate changes in the washout circuitry initially selected did not alter pilot opinion during subsequent testing.

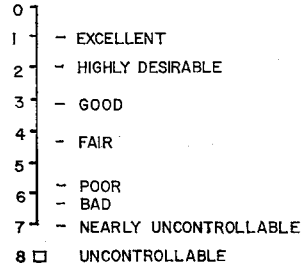

Fig. 4 McDonnell scale.

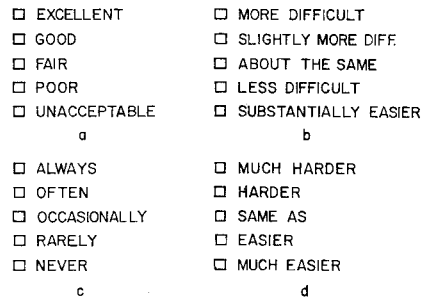

Fig. 5 Conrad scales.

It would appear that one or both of the following factors were responsible for the inability of rating pilots to distinguish minor changes in simulator characteristics. 1) The evaluation task was insensitive to minor changes in system response. 2) The rating scale adjectives were too widely separated on the mentally pictured, nondivided scale.

During an interview, Conrad discussed the work on which he had reported in Ref. 11. In determining the best washout circuitry the pilot ratings extracted from his scales were heavily supplemented with debriefs. It was primarily through this method of pilot interview that the best washout circuitry was obtained.

He observed that pilots rapidly adapted to minor configuration changes without altering their rating, and he described this lack of sensitivity as a rating plateau (Fig. 6).

He additionally noticed that a pilot's impression of his mean performance changed from day to day. This, therefore, required that at least one test run utilizing the "standard" washout circuitry be conducted to re-establish the pilot's mean performance, a time-consuming and costly procedure.

Conrad's present work, an extension of that above, tasks pilots with flying formation on the television display of a six-degree of freedom simulated tanker aircraft. It is his hope that this relative position task will prove to be sufficiently sensitive and thereby provide reliable pilot ratings on the scale depicted in Fig. 5d.

Summary

The rating scales which have been reviewed fall into the two categories, as distinguished according to purpose, of aggregate and relative handling qualities evaluations. The first category consists of the Cooper and Cooper-Harper Scales; whereas, the latter consists of the Harper, McDonnell and Conrad Scales.

Cooper's original scale was designed to eliminate the inadvertent misinterpretation of flight data while testing a variable stability aircraft. When Cooper presented his Scale at the annual meeting of the Institute of Aeronautical Sciences it was immediately accepted and internationally implemented as an aggregate evaluation scale. Though the Cooper Scale was not designed for this purpose, international usage determined its application.

In the collaborative effort to develop the Cooper-Harper Scale, Harper advocated a relative evaluation scale; however,

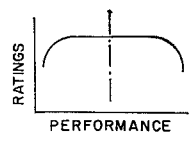

Fig. 6 Pilot response plateau.

the various implementing institutions preferred a scale applicable to aggregate evaluations and the two-part scale resulted.

The Harper and Conrad Scales were obviously designed to evaluate relative handling qualities and no further discussion is necessary.

The McDonnell (or Global) Scale was designed as an aggregate rating scale; however, because of simplicity of its adjective ratings, it could be applied only to relative evaluations. The sixty-three psychologically intervaled phrases resulting from McDonnell's research, however, were applicable to both aggregate and relative handling qualities evaluations.

In evaluations utilizing any of the rating scales except the Cooper-Harper Scale subjective pilot comment was required to provide meaningful evaluation data.

Human Response

Hypothesis

The Cooper-Harper Scale was excellently designed and remains the best aggregate rating scale in existence because of its dichotomous nature and its acceptance as the international standard. However, it was specifically designed so as not to facilitate the averaging of ratings.[9]

With the advent of greater sophistication of aircraft research and development, it has become increasingly important to evaluate the relative "goodness" of aircraft components and subsystems. It is assumed that a highly desirable aerospace vehicle may be designed and built; however, a rating scale capable of reliably determining the acceptance or rejection of one highly desirable system over another is yet to be evolved. It is the purpose of this section to investigate the possibility of such a rating scale.

For a scale to effectively reflect minor differences in performance, extreme sensitivity is desired. The inherent advantages of linearity are also desired to facilitate mathematical operations on a limited ensemble and thereby suppress research and procurement costs.

It was hypothesized that a linear rating scale with end points coincident with the psychological continuum would produce a sensitive and averageable scale. The psychological continuum has been investigated resulting in the McDonnell Scale,[10] but, as may be noted from Fig. 4 descriptive adjectives and/or phrases did not align cardinally. This, then, provided a source of confusion because the numerical value associated with the adjective might not coincide with linear points in the rater's mental scale. Were this source of adjective/numerical relationship eliminated, the rater could transpose his impression of performance directly to a rating scale and thereby relate his psychological continuum to a linear numerical index. Additionally, if allowed to fractionalize his rating, sensitivity would be limited only by the rater's descriminate dispersion[5] and frustrations.

Validation Tests

To investigate this hypothesis, a simple puzzle was selected and submitted to the analytically inclined students in the Department of Aeronautics of the Naval Postgraduate School. Upon completion of the test, or at the expiration of an allotted time, the subjects were asked to rate their impression of the difficulty they encountered in working the puzzle on three numerical scales; zero to ten, zero to four, and ten to zero.

The plastic Kohner EVEN-STEVEN solitaire puzzle was used as the testing device. It consisted of a base with eight equal depth holes that accepted eight equal length sleeves with variable interior depths into which fit eight variable length pegs. The puzzle had 40,320 (eight factorial) different solutions, one of which resulted in all pegs being even.

Before starting the exercise, the subjects were briefed in detail regarding the physical characteristics of the puzzle. Prior to each test the pegs and sleeves were removed from the base and mixed randomly within a box before the subject. The exercise was started on the proctor's "mark" with the subject's hands poised over the box. At test completion the time was recorded or, if the subject did not complete the test in 60 sec, the number of even pegs, regardless of height, was recorded. The elapsed time or number of even pegs was the basis for determining performance.

The subject was then asked to rate his impression of the difficulty he encountered in working the puzzle with respect to all three scales on the Rater Questionnaire and to indicate his rating in the box provided. This procedure was repeated so that each subject underwent three tests. When subjects inquired as to the degree of difficulty associated with scale end points, they were told that this determination was the rater's responsibility. By so doing, the rater's personal psychological continuum was used to establish the breadth of the scale.

Linearity

To facilitate detailed analysis and to justify raw data averaging, an individual correlation factor (r) was calculated for each of the thirty-one exercise subjects. In correlation factor calculations the time to exercise completion or the number of even pegs was used as the independent variable, and the subject's rating was used as the dependent variable.

The zero to ten (A) and zero to four (B) scales yielded correlation factors of which 90.9% were greater than 0.8 and 81.8% were greater than 0.9. The ten to zero (C) scale yielded 77% and 72%, respectively. The over-all correlation factors for scales A, B, and C were 0.928, 0.905, and 0.927 respectively. The high degree of performance-rating correlation confirmed linearity and sensitivity, and was an extremely strong indication that raters were able to relate their personal psychological continuum to a linear, nonadjectival, nonordinal rating scale. It additionally provided justification for the averaging of ratings.

Another feature of high correlation is that relatively few trials may be conducted with a high degree of confidence in the resulting data. This thereby reduces the time and cost expenditures associated with testing.

Rating Analysis

The test subjects' ratings fell into two groups as characterized by those who completed all tests during the allotted time (Group X) and those who completed two or less tests (group Y). As indicated in Fig. 7, group X experienced less difficulty than Y throughout the testing sequence; however, the rating curves of group X reflected decreased learning in contrast to the curves of group Y.

It should be noted that the rating curves of group Y did not remain parallel as did those of group X. This was, perhaps, an indication of the frustration experienced in not being able to complete each test. Such a factor would influence rating accuracy and, consequently, rating sensitivity.

By averaging the unweighted corresponding test ratings of both groups (there were more subjects in group Y), Fig. 8 was constructed. As may be observed, the average rating curves ranged about the numerical mean of each scale, and, in fact, the average ratings of scales A, B, and C were 5.00, 2.02, and 5.02, respectively.

Considering these two facts, it must be assumed that the test subjects discarded any "degree of difficulty" associated with the scale end points and related all of their ratings to the scale numerical mean. Consequently, all rating was a matter of judgment; a matter of relating their psychological continuum to what ever scales were presented. Whether test subjects consciously or subconsciously related to the scales' numerical means was beyond the scope of this study.

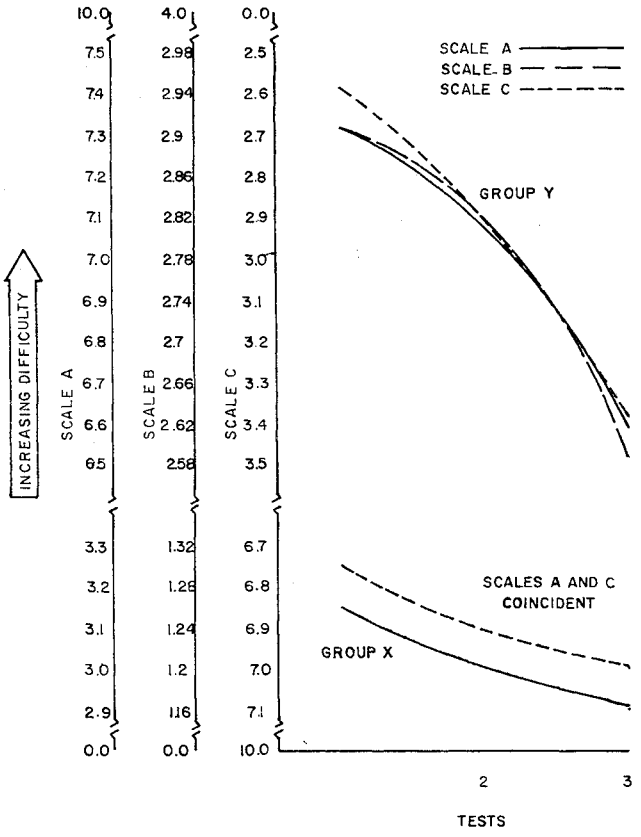

Fig. 7 Group rating curves.

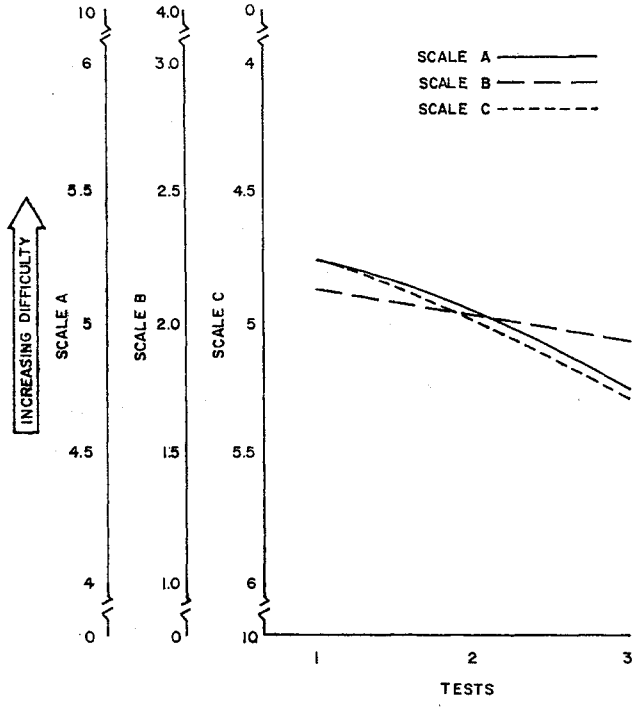

Fig. 8 Average rating curves.

Scale Preference

Of the 31 test subjects, 28 preferred scale A, two preferred scale B, and one preferred scale C. It was interesting to note that Scale A construction paralleled that of the Cooper-Harper Scale (i.e., increasing numerical index with increasing degree of "badness"); however, only 35% of the test subjects had ever been exposed to the Cooper-Harper Scale. Because the subjects were enrolled in a mathematically oriented curriculum, the preference for a decimal system based on ten seemed appropriate. As evidenced from the over-all correlation factors and the scale average ratings, the preference for scale A appeared valid.

The limited preference for scale B was believed to reflect exposure to the 4.0 Navy system.

Conclusions

The high correlation experienced during this investigation indicates that a rater may transpose his impressions of performance directly to a nonadjectival, nonordinal rating scale and thereby relate his psychological continuum to a linear numerical index. Because of the historical precedent and preference of a decimal, decade scale, this should be used for the nonadjectival modifier.

Although the Cooper-Harper scale is unequivocally accepted for its designed purpose, it is believed that its usefulness could be improved if the non adjectival scale were used as supplement.

For example, if the rater decided that the tested system was satisfactory (Pt. 1), but fair, with some mildly unpleasant characteristics (Pt. 2), he would assign a Cooper-Harper Rating of 3. This would then be the first digit of a series. The rater would then evaluate the relative "goodness" within this rating according to a non adjectival scale from one to ten based on his psychological continuum. If he decided that the latter "figure of merit" was 4.5, the final rating that he would assign would be 3.4.5. This number would, therefore, combine gross characteristics (from the Cooper-Harper Scale) with a fine tuning capability (from the nonadjectival continuum scale).

The use of this nonadjectival rating system could provide simplicity, linearity, averaging capability, high correlation and a high level of confidence for minimum testing. Such a combined scale, if used in contemporary testing, might greatly reduce evaluation costs.

References

[1] McRuer, D. T., "Human Pilot Dynamics in Compensatory Systems" AFFDL-TR 65-15, July 1965, Air Force Flight Dynamics Lab., Wright-Patterson Air Force Base, Ohio.

[2] Schultz, W. C., Newell, F. D., and Whitbeck, R. F., "A Study of Relationships Between Aircraft System Performance and Pilot Ratings," NASA CR-1643, July 1970, Cornell Aeronautical Lab. Inc., Buffalo, N.Y.

[3] Warner, E. P., *Airplane Design*, McGraw-Hill, New York, 1936, pp. 550–552.

[4] Gilruth, R. R., "Requirements for Satisfactory Flying Qualities," April 1941, Advanced Confidential Report, NACA.

[5] Cooper, G. E., "Understanding and Interpreting Pilot Opinion," *Aeronautical Engineering Review*, Vol. 16, No. 3, March 1957, pp. 47–51, 56.

[6] Harper, R. P., Jr., "In-Flight Simulation of the Lateral-Directional Handling Qualities of Entry Vehicles," AFFDL-TR-61-147, Nov. 1961, Cornell Aeronautical Lab. Inc., Buffalo, N.Y.

[7] Parrag, M. L., "Pilot Evaluations in a Ground Simulator of the Effects of Elevator Control System Dynamics in Fighter Aircraft," AFFDL-TR-67-19, Sept. 1967, Cornell Aeronautical Lab. Inc., Buffalo, N.Y.

[8] Harper, R. P., Jr. and Cooper, G. E., "A Revised Pilot Rating Scale for the Evaluation of Handling Qualities," AGARD Conference Proceedings 17, Sept. 1966.

[9] Cooper, G. E. and Harper, R. P., Jr., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities," TN-D-5153, April 1969, NASA.

[10] McDonnell, J. D., "Pilot Rating Techniques for the Estimation and Evaluation of Handling Qualities," AFFDL-TR-68-76, Dec. 1968, Systems Technology Inc., Hawthorne, Calif.

[11] Schmidt, S. F. and Conrad, B., "Motion Drive Signals for Piloted Flight Simulators," CR-1601, May 1970, NASA.