REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
April 2004

2. REPORT TYPE
Technical Paper

3. DATES COVERED *(From - To)*
April 2004

4. TITLE AND SUBTITLE
~~Flight Tests, Analyses, and Rapid Tools for Advanced Laser Targeting Pods United States Air Force Academy TG-14A Motorglide High Density Altitude Operation Limited~~

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
George, Edward J.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
418 FLTS/DOEF
Air Force Flight Test Center (AFFTC)
Edwards AFB, CA 93524

8. PERFORMING ORGANIZATION REPORT NUMBER
PA-04030

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
418 FLTS/DOEF
Air Force Flight Test Center (AFFTC)
Edwards AFB, CA 93524

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

12. DISTRIBUTION / AVAILABILITY STATEMENT
A Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
CC: 012100 CA: Air Force Flight Test Center Edwards AFB

14. ABSTRACT

The intent of this paper is to encourage reevaluation of the unidimensional workload scaling used in aerospace test and evaluation applications. The more specific intent is to encourage reevaluation from a structured psychometric viewpoint. The end goal is to facilitate a uniformly higher standard of measurement quality in unidimensional scaling having complex scale step descriptors. The basic principles and methods of psychometrics have been accessible in the technical literatures for decades. Even so, they have not been consistently applied to the design and verification of scaling for aerospace crew station usability evaluations. Psychometric verification should be performed on every scale employed as a test and evaluation tool. To this end, a simple but powerful psychometric method is demonstrated via two case studies performed at the United States Air Force Flight Test Center (AFFTC), Edwards Air Force Base California. Study 1 assessed the AFFTC revised United States Air Force School of Aerospace Medicine (USAFSAM) workload scale, while Study 2 assessed the Bedford workload scale. As was the case with the original USAFSAM, the Bedford was found to be psychometrically deficient, although the revised USAFSAM was verified to be psychometrically satisfactory.

15. SUBJECT TERMS
United States Air Force School of Aerospace Medicine (USAFSAM)
Test and Evaluation (T&E)
Combined Test Force (CTF)
Test Pilot School (TPS)
Standard Deviations (SD)
Bedford Workload Scale

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
			Unclassified Unlimited	29	Edward George
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED			661-277-7190

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

20040930 015

The Psychometric Anatomy of Two Unidimensional Workload Scales

By

Edward J George

AFFTC, Edwards AFB, California
October 2004

Abstract

The intent of this paper is to encourage reevaluation of the unidimensional workload scaling used in aerospace test and evaluation applications. The more specific intent is to encourage reevaluation from a structured psychometric viewpoint. The end goal is to facilitate a uniformly higher standard of measurement quality in unidimensional scaling having complex scale step descriptors. The basic principles and methods of psychometrics have been accessible in the technical literatures for decades. Even so, they have not been consistently applied to the design and verification of scaling for aerospace crew station usability evaluations. Psychometric verification should be performed on every scale employed as a test and evaluation tool. To this end, a simple but powerful psychometric method is demonstrated via two case studies performed at the United States Air Force Flight Test Center (AFFTC), Edwards Air Force Base California. Study 1 assessed the AFFTC revised United States Air Force School of Aerospace Medicine (USAFSAM) workload scale, while Study 2 assessed the Bedford workload scale. As was the case with the original USAFSAM, the Bedford was found to be psychometrically deficient, although the revised USAFSAM was verified to be psychometrically satisfactory.

TABLE OF CONTENTS

Background ..... Page 3

Basic Concepts and Devices of Psychometrics ..... Page 4

Simple Unidimensional Scaling ..... Page 5

Pair Comparison and Successive Intervals .....Page 6

Complex Unidimensional Scaling ..... Page 9

The Rank-Order Method ..... Page 10

Task I: Rank Order Sort ..... Page 10

Task II: Interval Estimation ..... Page 11

Case Study 1: USAFSAM Workload Scale ..... Page 13

Methods .....Page 13

Results ..... Page 15

Case Study 2: Bedford Workload Scale .....Page 16

Methods ..... Page 19

Results .....Page 20

Discussion .....Page 28

References .....Page 29

The Psychometric Anatomy of Two Unidimensional Workload Scales

Background

As a working definition, psychometrics has the purpose of systematically ensuring that a rating scale is composed of distinctly different levels of intensity on the scale's dimension of interest. For example, if workload were the dimension of interest, then the sequence of scale steps should represent an unambiguous ordinal hierarchy of workload intensity levels from low to high. If the only thing that distinguished one scale step from another were a number to identify their relative positions on the scale's measurement continuum, then psychometrics would have little application. However, if word phrases are used to anchor the meaning of each step to some specific definition of workload intensity, then psychometrics does apply. In other words, psychometrics has to do with establishing a descriptive phrase's ability to characterize a distinct level of intensity on the dimension of interest (reference 1). Thurstone established the benchmark principles and methods of psychometrics more than 75 years ago (reference 2). However, in aerospace crew station test and evaluation (T&E), structured psychometric methods are still only inconsistently applied at best. Crew station usability surveys are often designed and pressed into service without any certainty that meaningfully quantifiable information will result. This is a notable omission considering the weight that aircrew opinion tends to carry and also because surveys are so frequently employed when crew station usability is formally evaluated. The reasons for the shortfall are deduced to be both economic and cultural.

Historically, the emphasis in psychometrics has been in the areas of intelligence and personality testing, social and political attitudes, consumer preferences, and employee satisfaction for example. In contrast, issues of how psychometrics might be rigorously applied to testing of human-machine systems, has thus far attracted considerably less attention. The primary reason is that the fundamental principles of psychometrics are not common knowledge among aerospace T&E professionals. Quite legitimately, the traditional engineering emphasis has been on objective measures. Crew station interfaces are, however, sufficiently complex to make it prohibitively difficult to address all critical usability issues with objective measures alone. In spite of this, issues of subjective measurement quality do not always receive the attention they deserve. Compounding the situation, crew station usability testing is usually characterized by marginal control over test conditions and abysmally small numbers of aircrew participants. These limitations, coupled with sometimes-misconceived notions about human response variability, tend to foster a view that little if anything quantitatively useful can be done with aircrew survey data. Therefore, as the reasoning may go, high quality in subjective rating scales is largely superfluous.

To the contrary, a usability survey can produce hundreds if not thousands of scores encompassing all aspects of the crew station operation. A survey is therefore a potentially rich source of information, or misinformation, about crew station capabilities

and limitations. Scale induced response variability will have a cumulatively adverse impact on the entire database. The use of very small numbers of subjects and the existence of difficult to control variables only make it more imperative that the scaling itself not be a contributor to measurement error. The latter potentiality can be largely mitigated with a simple but powerful psychometric verification method. The method can be effectively used even when subject resources are limited, perhaps amounting only to immediate project aircrew and T&E engineering personnel. More generally, the method will prove useful both for verification and refinement of new or existing scaling for which no prior psychometric data exists.

The following discussion revisits the fundamental concepts and devices of psychometrics. This is intended as a lead-in to describing the proposed method itself. The method's practical utility is then demonstrated with two case studies performed at the United States Air Force Flight Test Center (AFFTC), California. Study 1 was a revision and verification of the USAFSAM general crew station workload scale, while the Study 2 was a critical evaluation of the Bedford pilot workload scale. The paper concludes with a discussion of the key psychometric issues governing the design, implementation, and application of unidimensional scaling composed of complex scale step descriptors.

Basic Concepts and Devices of Psychometrics

In the purest sense, 'unidimensional' refers to scaling designed to obtain ratings on a single dimension of interest (reference 2). Unidimensional scaling is desirable in aerospace T&E because the rater need only supply a single score for any given test point or question item. In human-machine systems the overriding dimension of interest is workstation usability. However, issues of usability in aerospace T&E tend to be focused on sub dimensions like aircraft handling qualities, control and display operability, mission utility, situational awareness, and crew workload to name a few. A typical survey will be composed of an itemized list of mission tasks or crew station interface components. Ideally, either type of listing will represent the essential capabilities required of the crew station system for the intended mission to be accomplishable. Each item is rated either for the absolute level of usability it demonstrates or usability in comparison to some alternative crew station system or configuration.

The focus of this paper is constrained to the scaling used to obtain item ratings. In the final analysis, classic psychometrics is just as strongly focused on the metric relationships among the question items themselves. In the latter case, the central concern is the metric differences or equivalences among items relative to their standing or importance to the survey's critical dimension of interest. Certainly, issues of question item weighting are just as important to crew station survey design as they are to more traditional survey applications. Typically, however, they receive even less structured psychometric attention then do the scales actually used to score them. Although the present discussion is limited to the rating scale component alone, the importance of question item metrics in the larger psychometric scheme of things is nevertheless worthy of note. Indeed, survey item psychometrics represents an extensive subject area for

aerospace T&E where little if any research has been systematically documented. Still, the fundamental principles and devices of psychometrics are essentially the same regardless of whether rating scale or survey item design is the issue of interest. As such, the following discussions provide a thumbnail review of basic psychometric principles and devices as they might be applied to both.

Simple Unidimensional Scaling

A rating scale for a usability survey might directly express a range of states from *totally usable* to *totally unusable*. In actual practice, the scaling is more likely to employ other more general dimensions like *satisfactory-unsatisfactory, acceptable-unacceptable, adequate-inadequate, effective-ineffective, agree-disagree, more-less,* or *better-worse* to obtain ratings. There are several reasons for this. First, these general dimensions are adaptable to a variety of question item types representing different evaluation dimensions. For example, it makes complete sense to rate a crew station's operability or mission utility in terms of how adequate, acceptable, or satisfactory it is. Second, the psychometric contingencies of matching the above stated general dimensions with intensity level modifiers like *slightly, moderately, substantially, very, totally,* and others, were formally researched and documented several decades ago. A variety of general-purpose scales were subsequently devised and repeated use has earned them common acceptance. Typically, however, similar research has not been accomplished for more specialized terms like usability, operability, situation awareness, workload or a host of others of specific interest to the aerospace T&E mission. This latter omission characterizes a key psychometric problem in aerospace usability testing.

The psychometrics of rating scale design has the fundamental purpose of quantifying the differences in intensity of meaning that distinguish one scale step descriptor from another. The goal is to ensure that an unambiguous ordinal hierarchy exists among them. To illustrate, consider a situation where the tester wants to capture aircrew judgments about the operability of a radio communications panel and decides to use a bipolar adequate/inadequate scale for question item scoring. Intuitively, it is quite certain that a scale step descriptor like *totally adequate* will be universally perceived as representing a stronger state of dimensional intensity then *slightly adequate*. However, issues of proper rank order between descriptors become more difficult to resolve when comparing alternatives like *totally inadequate* to *wholly inadequate* or *slightly adequate* to *barely adequate* for example.

In a well-designed scale, each step descriptor will represent a clear difference in dimensional intensity relative to all others. In addition, it is also desirable to have the subjective distances between adjacent step descriptors approximate a continuum of equal appearing intervals. When the subjective distances between scale steps approach this ideal, an approximate to a linear measurement continuum is achieved. In such case, the highest possible certainty exists that no scale step will ever get confused with another as to either the relative or absolute level of dimensional intensity it represents.

The classic and methodologically rigorous way to accomplish this is to first develop an exhaustive list of candidate descriptors that could be included in the final version of the scale. For example, a listing of candidate descriptors might include: *totally adequate, wholly adequate, decidedly adequate, completely adequate, strongly adequate, very strongly adequate, substantially adequate, moderately adequate, somewhat adequate, barely adequate, slightly adequate, marginally adequate,* as well as others. Once an exhaustive listing is developed, the next task is to quantitatively establish how they compare in terms of the relative intensity of adequacy represented. Pair comparisons and successive intervals are the most frequently used psychometric methods, but other serviceable methods are also used (reference 3).

Pair Comparison and Successive Intervals

Pair comparison is a method where multiple subjects evaluate every possible combination of two descriptors that can be made from the exhaustive listing of candidates. For each combination, the subject identifies a point on a multi-step continuum that represents the amount of subjective intensity, if any, distinguishing the two descriptors. For example, Ames (reference 4) devised a variant of the pair-comparison method to evaluate subjective scaling in T&E. If it were used to evaluate candidates for a simple bipolar adequate/inadequate scale it would have been employed as shown in figure 1.

Consider each descriptor pair below. If equal, put a check in the left-most column. If unequal, circle the letter of the descriptor describing the higher level of adequacy and rate the degree of unequalness by checking one of the other eight columns

		RELATIVE AGREEMENT DOMINANCE			
	DESCRIPTOR PAIR	EQUAL	A Little More	A Good Deal More	Very Much More
(a.)	Wholly Adequate			✓	
b.	Strongly Adequate				

Figure 1: Example of a Pair Comparisons Method

The method of successive intervals serves the same purpose, but achieves it by having the subjects rate the position of each candidate relative to the extreme poles of an interval continuum (reference 5). In the case of the adequate/inadequate dimension, the poles might be defined as being *most adequate* and *most inadequate* respectively as shown in the hypothetical example in figure 2.

Place a checkmark above the number that best identifies the level of adequacy represented by the phrase "Strongly Adequate".

-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
								✓		

Most Inadequate — **Neither Adequate Nor Inadequate** — **Most Adequate**

Figure 2: Example of a Successive Intervals Method

Pair comparison is the more rigorous of the two methods because just 10 candidate descriptors will require 45 individual comparisons while 20 candidates require 190. If the desired scale is bipolar, then every descriptor's antonym must also be included and the number of required comparisons becomes much larger. In contrast, successive intervals require only one estimate per candidate descriptor. As such, 10 candidate descriptors require only 10 ratings. Because the two methods appear to correlate well (references 2 and 5) and substantially less labor is required with successive intervals, it is most commonly chosen when exhaustive studies of candidate descriptors are undertaken.

Table 1: Partial Tabulation of Adequate/Inadequate Candidate Descriptors

Candidate Descriptor	Mean	Standard Deviation
Totally Adequate	4.620	0.846
Absolutely Adequate	4.540	0.921
Very Adequate	3.420	0.851
Slightly Inadequate	1.200	0.566
Barely Adequate	0.627	0.928
Neutral	0.00	0.00
Barely Inadequate	-1.157	0.638
Slightly Inadequate	-1.380	0.772
Moderately Inadequate	-2.157	1.107
Very Inadequate	-3.735	0.777
Wholly Inadequate	-4.784	0.676
Totally Inadequate	-4.900	0.412

(Values taken from Matthews, Wright, and Yudowitch, 1975, reference 6)

Having accomplished the tabulations, the next task is to build a scale where the mean distances separating the steps are as equal as possible, but also with each having small standard deviations so that little if any distributional overlap occurs between them.

A hypothetical solution for a six-point adequate/inadequate scale, showing means and standard deviations, is presented in figure 3.

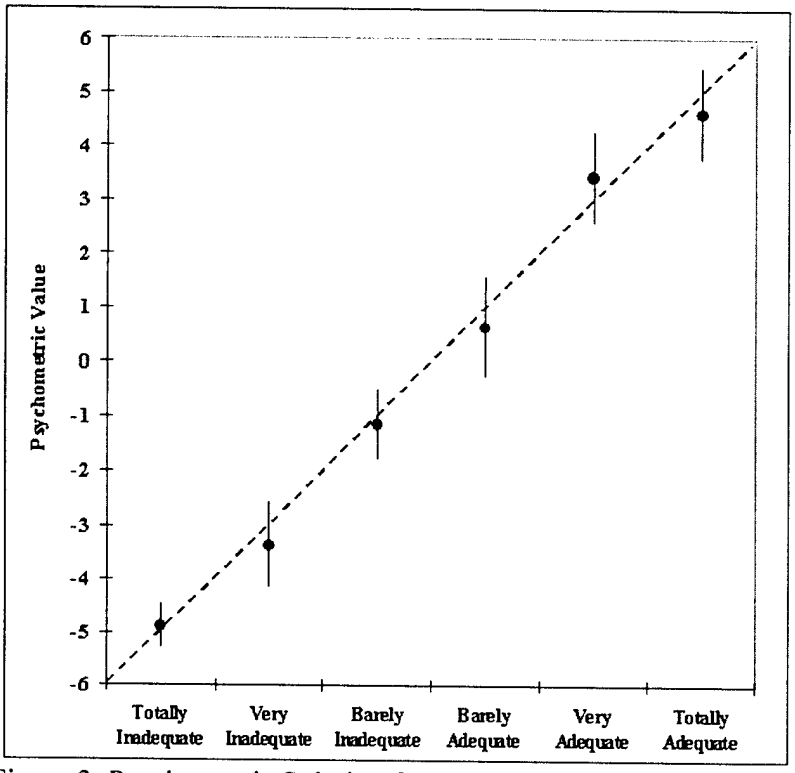


Figure 3: Psychometric Solution for an Adequate/Inadequate Scale

Ideally, the tabulations will be based on data from as many subjects as possible, all randomly selected from the target population of survey subjects. Once an initial set of steps has been extracted from the tabulation, additional data should be obtained to verify that a strong approximate to a linear continuum has been captured. The additional data might show the final set of step descriptors to be a closer approximate then predicted by the tabulated values. If the new data shows otherwise, then the proper recourse is to go back and reevaluate which among the candidates should actually be included in the final cut of the scale. All this can amount to a lengthy iterative process. Many survey designers, particularly in the applied world, are hampered by a lack of time, subjects, and or knowledge of psychometric methods. Consequently, exhaustive preparatory studies are typically not accomplished as precursor to most usability surveys.

More typically, a scale is adapted from previous testing or from a survey design handbook. This allows the survey designer to avoid exhaustive psychometric groundwork. As a further accommodation, rank-order tabulations like the one shown in table 1 are available in the technical journals for a variety of common scaling dimensions. A few handbooks go a step further by providing inventories of tabulations for the most popular dimensions like *adequate-inadequate*, *agree-disagree*, *good-bad*, *acceptable-unacceptable*, and others (references 7 and 8). With these data available, scales with simple two-term descriptors can be adapted or even custom tailored with relative ease.

It is notable, that most of the published tabulations are seldom supported with data from more than a few hundred subjects and sometimes less. For any given target population, adequate sample size is always a matter of interest. Still, the relatively small samples found in most published studies, suggests that large numbers of subjects are not always necessary for reliable psychometric values. This is most likely if the dimension of interest and its modifying adjectives are very common in English usage, the subjects are truly representative of the target population, and their minimum education level is relatively high. The results of the two case studies presented herein tend to support the validity of this assumption. In any case, with the availability of documented rank-order tabulations, the survey designer need only collect enough additional data to verify that a good approximate to a linear continuum was captured with the chosen set of descriptors.

Complex Unidimensional Scaling

The situation becomes more difficult when the dimensions of interest are specialized and/or the scale step descriptors are complex. In such case there will be no published tables to rely on. This is frequently the case in aerospace usability testing where specialized dimensions like workload are involved. For these applications, survey designers often want their scaling to possess a greater precision of meaning then simple two-term descriptors can provide. Two general situations may subsequently arise.

The survey designer may begin with an existing scale solution composed of simple two-term descriptors of known psychometric properties, but then proceed to enhance each descriptor with a supporting auxiliary definition. For example, in a control and display operability survey, a scale step descriptor of *totally unacceptable* might be further qualified by the phrase *"unusable; mission essential operations cannot be accomplished."* This is done in an attempt to anchor the meaning of the step to the specific information needs of the evaluation. In other cases a stand-alone set of descriptors with multiple sub dimensions in the phraseology is customized from scratch. For hypothetical example, the designer of a task-oriented scale might desire one step to be defined as *"task easily completed, plenty spare time, no input errors"* versus another step defined as *"task completed with some difficulty, little spare time, very few input errors"*. Assuming the scale will ultimately consist of between 5 and 10 steps, the designer faces a clear challenge in attempting to account for all possible combinations of the sub dimensions while still managing to adjust the within-step intensities to approximate to an interval continuum. Success also presupposes that meaningful real-world correspondences in intensity level exist among the sub dimensions.

In both situations, scale design will likely not follow the classical process of first developing an exhaustive listing of candidate descriptors and then gathering psychometric data as precursor to selecting a final set among them. Rather, a single set of definitions is drafted at the onset and then armchair refined until the designer is satisfied. Typically in such case, the definitions are refined in accordance with the designer's own personal sense of dimensional intensity, frequently aided only with informal input from a few peers and aircrew at best.

Starting with just enough descriptors to fill out the desired number of steps need not be an unsound approach. In situations where complex and highly specialized multi-element descriptors are required, it may be the most correct way to proceed. Psychometrics must nevertheless be integral to the process. In such case, however, a different regimen of psychometric refinement and verification is required. That is, at the beginning of the refinement process, psychometric data on the prototype scale is iteratively collected using very small numbers of subjects. When the most current iteration indicates that a psychometrically sound measurement continuum has been resolved, a larger number of subjects are then used to provide final verification. Unfortunately, a number of scales have been designed and pressed into service without psychometric refinement or verification of any kind. Because of their repeated appearance in technical reports, some of these scales have become established scales of choice. This is problematic because any data generated from their use will always be questionable to a degree. For example, if any of the step definitions turn out to be confusable or transposable with one another in terms of the level of intensity they represent, then the scale is technically no better than nominal. In such case, scale users will not be able to effectively discriminate between the descriptors, thus making the task of selecting the most representative step a matter of guesswork. The consequence is scale induced response variability, which is ultimately likely to translate into measurement error both within and between users. In turn, this stands to compromise the sensitivity and reliability of the survey database as a whole.

The two case studies herein will show that the survey designer cannot rely on his or her own sense of denotative and connotative meaning to establish scale step intensity norms for the subject population. This can only be done with structured psychometric methods. Consequently, a tangible need exists for a simple but effective psychometric method that can be used in applied situations where time and subject resources are limited. The method described in the following sections was devised to satisfy that need.

The Rank-Order Method

Task 1: Rank Order Sort

The proposed method employs a conventional rank order sort as its basis. It is administered in questionnaire survey format, typically composed of two to four tasks for the subjects to complete. In Task I, the subject takes a deck of shuffled flashcards, each card having one of the scale's descriptors printed on it. The subject is instructed to sort the cards from lowest to highest level of dimensional intensity. Assume for the present example a ten-step workload scale. The size of the flashcards is somewhat arbitrary, but 2.5 by 3 inches should accommodate the length of most descriptors. Task I is completed as per the instructions shown in figure 4.

There is nothing new or unique about the rank-order method. To paraphrase Guilford (reference 3), it is one of the most popular and practical in the psychometric inventory. It possesses several decided advantages over both pair comparisons and

successive intervals. First, the subjects are forced to make the same number of discriminations between descriptors as pair comparisons requires, so the maximum amount of discriminatory information is obtained. Second, it allows the subject to perceive the scale as an integrated whole, which is consistent with how it would be perceived during applied scoring. Third, it forces the subject to decide on a dominant rank order among the step descriptors. This is advantageous, because valid scaling must ultimately demonstrate a very high level of agreement, if not total agreement among subjects as to the dominant rank order. A critical assumption of the present methodology is that if a dominant rank order does exist, then it should demonstrate itself even among a very small number of subjects. Conversely, if there is no dominant rank order, then this should also be evident even among a very small number of subjects.

Each card has a letter of the alphabet affixed in the upper left hand corner. When you are satisfied that you have sorted the cards correctly, write their respective letters in the ten boxes provided below, one letter for each box, with the letter for the lowest level of workload in the first box, the next higher level of workload in the second box, and so on.

FIRST	SECOND	THIRD	FORTH	FIFTH	SIXTH	SEVENTH	EIGHTH	NINTH	TENTH

Figure 4: Rank-Order Method- Task I

Task 2: Interval Estimation

The rank order sort alone has limitations because it provides only ordinal information. This falls short of establishing relative interval distances between adjacent steps. To compensate, Task II is included as shown in figure 5. Task II is the most unique aspect of the present methodology. Having already accomplished a rank ordering, the subject is then positioned to directly estimate relative interval distances between the sorted step descriptors. The forced choice as to rank order in Task I was at the expense of some information loss. If the subject perceived any two descriptors as being the same or nearly the same in dimensional intensity, the rank order sort would not reveal it. On the other hand, the interval distance estimates required in Task II would.

The primary diagnostic benefit of Task II is that mean intensity estimates and standard deviations can be calculated for each descriptor. This information is vitally useful during scale refinement for detecting weaknesses in descriptor phraseology and to ensure that a solid approximate to a linear function is captured by the final iteration. In addition, it provides a metric device to evaluate how well a given scale stacks up against alternative scaling and against established psychometric standards.

As a methodological constraint, the first or lowest step in the subject's rank ordering is assigned a value of 1 while the tenth or highest step is assigned a value of 100. For any workload scale, it is likely that additional subjective space exists between the lowest step in the rank order and a state of zero total workload. In turn, some space might exist between the highest step and a state of total catastrophic overload. If points 1 and 100 on the ruled line (figure 5) were defined as total zero workload and total catastrophic overload, then each subject would be free to choose a position for all 10 steps within those extremes. This would defeat the essential purpose of the task. Each subject would then be using their own personal yardstick to define the subjective space occupied by the scale step continuum relative to the numerical boundaries of 1 and 100. As a virtual certainty, this would artificially increase the variance for all step definitions even if there were total agreement about the proper ranking among step descriptors.

Estimate the amount of workload defined by each of the ten scale step descriptors. Imagine that the lowest level of workload in your rank order sort (first step) was placed on the line below at the point checked as "1" and the highest level of workload (tenth step) at the point checked as "100". With this in mind, check a point on the line between 1 and 100 for each of the other eight steps according to where you think they belong on the line relative to one another.

Figure 5: Rank-Order Method- Task II

The exclusive purpose of Task II is to identify differences in perceived relative intensity among the step descriptors themselves. Meaningful data will only result if all subjects use the same subjective yardstick for the scale step continuum. Setting the position of the highest and lowest descriptors to values of 1 and 100 was to achieve that end. If there is total agreement among subjects as to which descriptors are the highest and lowest, then the sample variances for the two will be zero. If there is not total agreement, then variance will be present. If there is total agreement as to the proper ranking among all step descriptors, then their means will be regularly spaced across the 100 units of subjective distance. In turn, if little or no confusion exists between descriptors, then there will be little or no overlap between their sample distributions. The less agreement among subjects the more irregularity of spacing between means and the greater overlap between their sample distributions. The two case studies that follow demonstrate the validity of these assumptions.

Tasks II and I are the core elements of the method, but additional tasks may be added depending on the amount of detailed information desired. For example, two additional tasks are recommended. These require the subjects to identify any

combination of step descriptors that are perceived as confusable. If any confusion is identified, then the fourth task is to explain the source. The latter kinds of information are particularly useful for scale refinement as it provides clues about how descriptor phraseology might be changed to achieve more reliable differences in dimensional intensity.

Case Study 1: USAFSAM Workload Scale:

The original United States Air Force School of Aerospace Medicine (USAFSAM) 7-point workload scale is shown in the left portion of figure 6. It was devised in the late 1970's at the USAF School of Aerospace Medicine, Brooks Air Force Base, Texas, and subsequently employed as a companion to a 7-point fatigue scale. Together, the two scales combined to provide a system of measurement called the Crew Status Survey (reference 4). The step definitions for the workload scale were devised through a process of iterative refinement with the aid of inputs from USAF aircrew. However, structured psychometric verification was not accomplished. The scaling was used in a variety of vehicle transport applications chiefly to evaluate military aircrew workload (references 9 and 10). In the early 1990's several exploratory studies at the AFFTC confirmed that the original USAFSAM was essentially ordinal, but detailed ambiguities among scale step descriptors nevertheless detracted from its linearity. Some steps showed considerable variance relative to the others (references 4 and 11). This initiated a structured psychometric refinement and verification effort, which produced the revised form of the scale as shown in the right portion of figure 6 (reference 11). The following describes the methods and results of that revision effort.

Methods

Two psychometric methods were used to support revision and verification of the USAFSAM. These were the pair comparisons method identified in figure 1 and rank order method described in the previous section. Both were employed to collect psychometric data from a pool of AFFTC personnel consisting of both pilot and non-pilot aircrew and T&E engineers. Both methods used a conventional paper questionnaire as survey materials. The pair comparisons survey, which required 21 pair-wise comparisons, was administered via e-mail and returned at the convenience of the subjects. Subjects were recruited throughout the AFFTC T&E workforce. In contrast, the rank order survey was administered via one-on-one interview. This allowed the survey administrator to obtain detailed clarification if any confusion between step descriptors was identified. Recruits for the rank order survey were obtained from a tactical transport and special operations aircraft combined test force (CTF).

The USAFSAM was revised through a process of several iterations of step descriptor adjustments. The adjustments were made with the aide of small samples of data typically amounting to five subjects each. New subjects were used whenever a new iteration of the scale was tested. The actual refinement process was supported with several late versions of Webster's dictionary and the inventory of rank-order tabulations found in Babbitt and Nystrom (reference 7). Although no tabulation directly

corresponded to the dimensionality of the USAFSAM, they were still useful in determining which adjective modifiers would likely produce the desired differences in intensity level among descriptor sub elements. As evidenced by figure 6, the USAFSAM sub elements predominantly consist of activity level, time, and systems management. Notably, the revised scale employed activity level as the primary descriptive dimension by emphasizing it in upper case letters. This was the single most notable departure from the original scale.

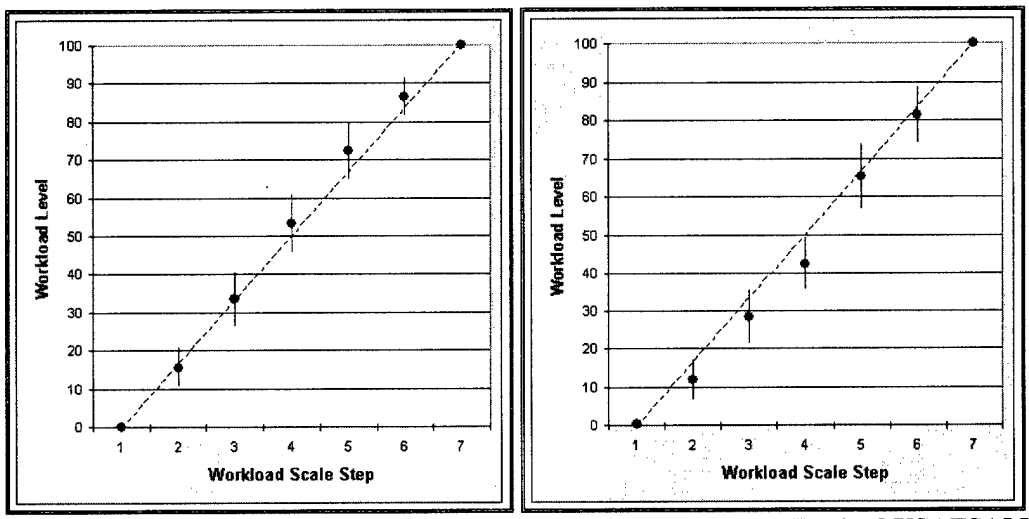
Nothing to Do; No System Demands	NOTING TO DO; No system demands
Little to Do; Minimum System Demands	LIGHT ACTIVITY; Minimum system demands
Active Involvement Required; But easy to Keep Up	MODERATE ACTIVITY; Easily managed; Considerable spare time
Challenging But Manageable.	BUSY; Challenging but manageable; Adequate time available.
Extremely Busy; Barely Able to Keep Up.	VERY BUSY; Demanding to manage; Barely enough time.
Too Much to do; Overloaded; Postponing Some Tasks	EXTREMELY BUSY; Very difficult; Non-essential tasks postponed.
Unmanageable; Potentially Dangerous; Unacceptable	OVERLOADED; System unmanageable; Essential tasks undone; unsafe

Figure 6: Original and AFFTC Revised USAFSAM Workload Scale

Subject comments obtained from the iterative sampling were essential to the success of the revision process. At the onset, the step descriptors were first adjusted according to designer insight and then the first sample of survey data was collected to evaluate the effects of the adjustments. Based on the results, more adjustments were made leading to more sampling and more adjustments. This process continued for several iterations. Descriptor adjustment was stopped when the sample data showed diminishing returns for improving the linear quality of the scale. At that time, a mixture of 49 aircrew and non-aircrew subjects was recruited for the final verification survey. Twenty subjects provided data for the rank-order method, while the other 29 provided data for the pair comparisons method. All subjects in the final verification survey were new subjects.

Results

Figures 7 and 8 show the resulting scale-step functions side-by-side for the two methods, showing intensity level mean estimates and their standard deviations for each of the seven steps. The data from the two methods were set to a common scalar standard for comparison. As evident from figures 7 and 8, the results for the two methods were quite similar, both showing a good approximate to a linear function. Babbitt and Nystrom (reference 7) established the criterion that no scale step distribution should overlap with the means of the adjacent step distribution out to 1 standard deviation. Figures 7 and 8 show the revised USAFSAM to have exceeded the criterion by nearly a full standard deviation for all scale steps.



Figures 7 and 8: Rank Order and Pair Comparisons Data for the Revised USAFSAM

Among the 20 rank-order subjects there was 100-percent agreement as to the appropriate ranking among the 7 descriptors. The pair comparisons methodology took a more circumspective and thus more rigorous approach to establishing both rank-order preference and relative intensity level differences. Even so, 26 out of the 29 pair comparisons subjects were in total agreement with the prescribed order, while the other 3 deviated only slightly. For example, one the three showed an inversion between scale steps 1 and 2, while the other two showed a zero difference in estimated intensity; one between steps 2 and 3 and the other between steps 3 and 4.

Overall, the results confirmed the achievement of a solidly ordinal continuum with low within-step variance. The AFFTC revised USAFSAM was subsequently used during the AC-130U Gunship program in parallel with the Finegold multi-dimension workload scale (reference 13). The Finegold dimensions consisted of time stress, mental effort, physical effort, environmental stress, and psychological stress with each dimension requiring a separate rating. The comparative results from these two workload measures showed good correlation (Pierson 0.85). Although the USAFSAM is facially valid, the statistical results confirmed its validity as a workload measure (references 4 and 12).

It is notable that the level of correspondence between the rank order and pair comparisons final verification data was very high even though the methods were distinctly different. In the pair comparisons survey, the subjects were not told how many scale steps there were and they made 21 comparative ratings on unnumbered descriptors (figure 1). Scale step ordinal ranking and the intervals between steps were determined mathematically from the subject ratings. In addition, the subjects had no direct feedback about their performance. In the rank order survey, the subjects knew exactly how many scale steps were being evaluated, that the descriptors were intended as an ordered continuum, and that the two descriptors selected as highest and lowest were to be anchored at the ends of the scale for Task II (figure 5). The rank order subjects had immediate visual feedback as to the spacing they assigned between scale steps, and could correct their responses if desired. As different as the two methods were, the obtained results showed a Pearson correlation of +0.994. By itself, this is not particularly notable because strong positive correlations between rank order and pair comparisons methods were previously reported (reference 3). What is notable is the relatively small number of subjects (20 versus 29) used to obtain the present results. This supported the assumption that under reasonably controlled circumstances, reliable psychometric data can be obtained using a relatively small number of subjects.

Case Study 2: Bedford Workload Scale

Since its conception in the 1980's at the Royal Air Force Academy, Bedford, England (references 14 and 15), the Bedford workload scale has been used in simulation and in flight test (figure 9). It is best classified as a Cooper-Harper type scale because it contains a 10-step inner scale continuum combined with an outer decision tree that divides the inner continuum into four workload impact classifications. The actual Cooper-Harper was conceived in the late 1960's to support assessment of aircraft handling qualities. The first scale of its kind in aerospace T&E, the Cooper-Harper became a standard device for flying quality evaluations and was subsequently embedded in the associated military standards (reference 16). The widespread acceptance of the Cooper-Harper eventually motivated other scale designers to copy its 10-step and decision tree format. These other scales simply used different dimensional phraseology in the scale step descriptors. For better or worse, a variety of Cooper-Harper "look-alikes" was devised with each employing different dimensional concepts to express workload. Although most of them have not gained popularity, the Bedford is one that did.

Like the USAFSAM, the Bedford is classified as unidimensional. However, it shares this distinction only because a single score is required of any given rating. Beyond this, the notion of unidimensional is somewhat difficult to justify. The Bedford's inner scale is intended to identify 10 different levels of workload intensity, using spare capacity, effort, and attention as the key interrelated sub dimensions. The outer decision tree is a unique sub dimension in its own right as it provides four levels of workload impact ranging from *workload satisfactory without reduction* to *could not complete the task*. This arrangement divides the inner 10-step scale into an unbalanced bipolar continuum with the first three steps designated as satisfactory while the other seven steps

are divided into three levels of unsatisfactory of increasing severity. Prescribed usage of the Bedford is to employ the decision tree to first select one of the impact classifications and then select a scale step within to further characterize the absolute level of workload involved. Thus, a minimum of two decisions is required for each scoring. The upper juncture of the decision tree queries the rater as to whether the experienced workload was *satisfactory without reduction* while the lower two junctures query as to whether the workload was *tolerable for the task* versus whether it was *possible to complete the task.* This inconsistency notwithstanding, the scale was intended to be a task specific and pilot oriented. The author recommended that its use be constrained to tasks that are well defined and of short duration (reference 8).

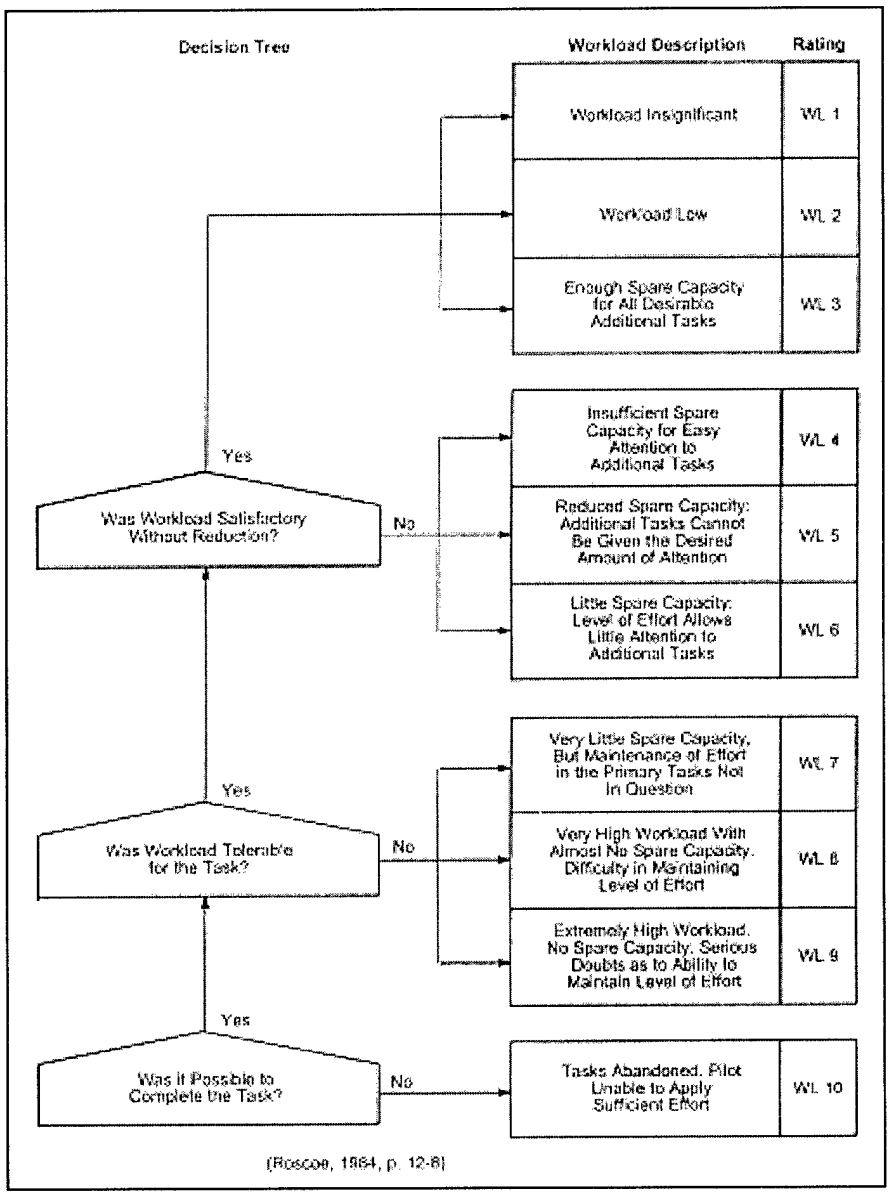


Figure 9: Bedford Workload Scale (from reference 10)

Several published studies indicate that Bedford scores do tend to raise and fall in parallel with gross changes in task load. This confirms that the scale's mix of sub dimensions do correlate with the human experience of workload. In balance, however, there is the equally important issue of reliability, which can also involve the underlying issue of measurement quality. There have been conflicting reports regarding this. Some reports claim good reliability while others evidence inconsistent scoring relative to objective differences in task load (references 16 and 17). The inconsistencies suggest the possibility of deficiencies in the scale's design and implementation, which could promote scale-induced response variability. If present, the deficiencies would likely stem from ambiguous scale step descriptors. The practical consequences are that some scale step descriptors would be ambiguous or indistinguishable from others in terms of the state of workload represented. Alternatively, some descriptors in the scale's numerical ranking might actually denote less workload then a descriptor occupying a lower numerical rank. Ambiguities about the absolute level of workload represented by any step descriptor would compromise the ordinal integrity of the scale and thus degrade the reliability of data generated from its use.

Much like the original USAFSAM, the Bedford was the product of armchair refinement with non-quantitative inputs from some pilots. Psychometric procedures were not used to either refine or verify the scaling. The absence of psychometric verification, coupled with a number of undocumented complaints about the difficulty of using the scale to supply meaningful ratings, triggered verification testing at the AFFTC in 2003.

Methods

This study employed a total of 48 AFFTC T&E personnel as survey subjects. They included 25 pilots, 3 navigators, 2 flight engineers, 3 loadmasters, 1 air refueling boom operator, and 14 flight test engineers. Twenty of the subjects were recruited from a tactical transport and special operation aircraft CTF and the other 28 were Air Force Test Pilot School (TPS) students. The previously described rank order method was exclusively used for this testing. The pair comparisons method was not used, partly because of limited subject resources, but also because the two methods were known to produce comparable results (figures 7 and 8). The same rank-order survey questionnaire used during the USAFSAM testing was administered to the 20 CTF subjects. The only difference was that the survey task items were configured for a 10-point rather than a 7-point scale. The CTF subjects received the questionnaire via one-on-one survey interview. When the questionnaire was completed, the subjects were queried to clarify their responses and to explain any sources of confusion between scale step descriptors. Although no time limit was set, the CTF subjects required about 15 minutes on average to complete the survey.

The same questionnaire was administered to the 28 TPS personnel, but was augmented with two additional tasks. One of the additional tasks required the correspondences between the Bedford's decision tree classifications and the scale's inner 10 steps to be evaluated. The details of this task are described in the following results.

The questionnaire was group administered to two classes of TPS students during regularly scheduled class time. This venue did not allow the students to be queried individually about any sources of confusion between scale step descriptors. To compensate, the second additional task requested an explanation in writing if any confusion was identified. The TPS students took the questionnaire under a half-hour time limit, but were able to complete it in about 20 minutes on average. The questionnaire was administered first to the CTF personnel. It was then administered to the TPS students to confirm the results from the CTF data and to extend the scope of the questionnaire to include the Bedford's outer decision tree. As it turned out, both TPS classes previously received 6 hours of human factors lecture. This included a brief introduction to the Bedford. None of the subjects had prior knowledge of what the survey intended.

Results

The fundamental assumption of rank-order estimation is that sound scaling will demonstrate a very high level of agreement if not total agreement among subjects as to the proper hierarchal (ordinal) relationship among step definitions. In contrast, table 2 shows a very high level of disagreement. All shaded cells in the table identify instances where a subject's rank order deviated from the Bedford prescription. As shown, 38 out of 48 subjects did not rank the definitions as the Bedford prescribed, and there was very little between-subject agreement as to what the proper alternative ranking should be. Perceptions of the appropriate rank order among steps WL-4 through WL-7 were particularly inconsistent. More than half the subjects found two or more scale step definitions confusing. Several rank orderings contained both scale step inversions and multi-step translocations relative to the prescribed order. In addition, there were large differences of judgment as to the actual intensity of workload that any descriptor represented relative to the others.

Overall, eight of the 10 subjects who sorted according to the Bedford prescription were pilots while the other 2 were non-pilots. This difference in outcome motivated a combinatorial analysis on the data (reference 18). If producing a rank order sort consistent with the Bedford were just as likely for non-pilots as for pilots, then the odds against getting a 2 to 8 ratio difference by chance alone would be slightly greater than 1 in 20. In detail, in the CTF data pilots produced 4 of the 5 rankings that were consistent with the Bedford prescription, as was the case for 4 of the 5 consistent rankings in the TPS data. In the CTF data, the ratio of pilots to non-pilots was 11 to 9 versus 14 to 14 in the TPS data. More generally, only about 1 in 3 pilots versus 1 in 12 non-pilots overall managed to produce a sort consistent with the Bedford, which makes for a rather striking difference. The reliability of any inference made from these results is, however, diminished because the comparison was unplanned and based on data from just 25 pilots versus only 23 non-pilots overall. Moreover, there was no deductively transparent reason why there should be any difference between pilots and non-pilots on this count. Still, the detailed trends in the data were strong enough to entertain the possibility for additional study.

Table 2: How the Bedford Scale Step Descriptors Were Ranked by the 48 subjects

	WL 1	WL 2	WL 3	WL 4	WL 5	WL 6	WL 7	WL 8	WL 9	WL 10
P	1	2	3	7	4	6	5	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	4	5	6	7	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	7	5	4	6	8	9	10
P	1	2	3	5	6	7	4	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	4	5	6	7	8	9	10
P	1	2	3	4	5	6	7	8	9	10
P	1	2	3	6	4	5	7	8	9	10
P	1	2	3	5	4	7	6	8	9	10
P	1	2	3	4	5	6	7	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	5	4	6	7	8	9	10
P	1	2	3	6	5	7	4	8	9	10
P	1	2	3	4	5	7	8	6	9	10
P	1	2	3	4	5	6	7	8	9	10
P	1	2	3	5	4	7	6	8	9	10
P	1	2	3	4	5	6	7	8	9	10
P	1	2	3	4	6	5	7	8	9	10
P	1	2	3	4	6	6	7	8	9	10
P	2	1	3	5	6	7	8	4	9	10
P	1	2	3	4	5	6	7	8	9	10
NAV	1	2	3	5	6	7	4	8	9	10
NAV	1	2	3	5	6	4	7	8	9	10
NAV	1	2	3	7	6	10	5	4	8	9
LM	1	2	3	7	4	6	5	8	9	10
LM	1	2	3	4	5	6	7	8	9	10
LM	1	4	2	3	7	6	5	8	9	10
FTE	1	2	3	6	7	4	5	8	9	10
FTE	1	2	3	5	6	4	7	8	9	10
FTE	1	2	3	5	6	7	8	4	9	10
FTE	1	2	3	5	6	7	8	9	4	10
FTE	1	2	3	6	4	5	7	8	9	10
FTE	1	3	2	4	5	6	7	8	9	10
FTE	2	1	3	5	6	7	4	8	9	10
FTE	1	2	3	5	7	6	4	8	9	10
FTE	1	2	3	4	5	6	7	8	9	10
FTE	1	2	3	5	6	7	4	8	9	10
FTE	1	2	3	5	6	7	4	8	9	10
FTE	1	2	3	7	6	4	7	8	9	10
FTE	1	2	3	5	7	6	4	8	9	10
FTE	1	2	3	5	6	4	7	8	9	10
FE	1	2	3	6	7	5	4	8	9	10
FE	2	1	6	3	7	4	5	10	9	8
BOOM	1	3	2	7	4	8	6	5	9	10

Notes: 1. The shaded cells identify rank orderings that deviated from the Bedford's prescribed rank order.
2. P = pilot, FE = flight engineer, LM = loadmaster, FTE = flight test engineer, NAV =navigator, and BOOM = boom operator.

In any event, figure 10 shows the average numerical level of workload that the 48 subjects assigned to each descriptor on the 100-unit ruled line (Task II, figure 5) along with their respective standard deviations (SD). These data are shown in comparison to an ideal linear function (dashed line). Variability among subjects as to the relative amount of workload represented by each descriptor was severest for the descriptors occupying the center of the scale. Still, there was substantial variance even for descriptors at the high and low ends. Although as a matter of central tendency, the Bedford appears to retain stable ordinal quality at the high and low ends of its 10-step continuum, the scale was no better than nominal in its center. That is, descriptors WL-4 through WL-7 are not reliably distinguishable in terms of the absolute level of workload they represent.

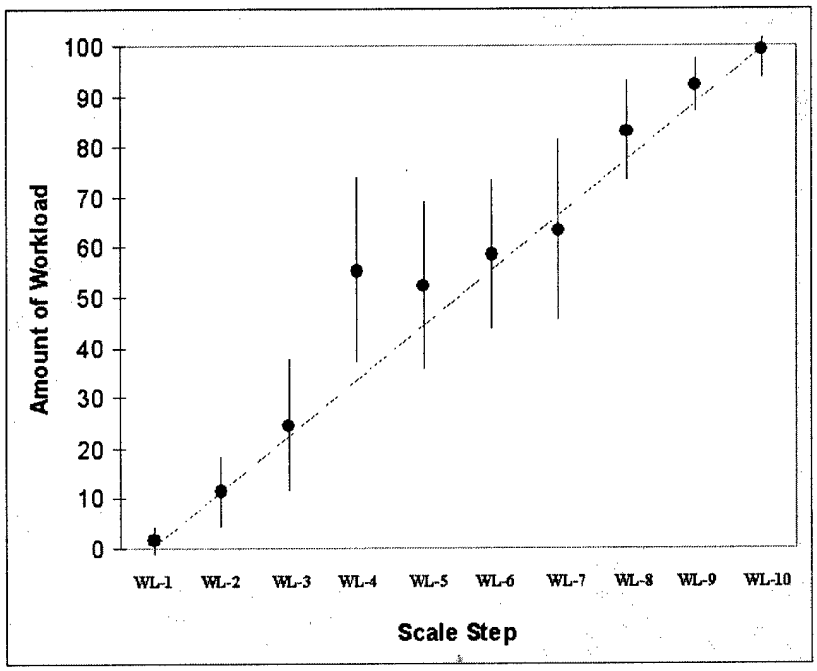


Figure 10: Mean Workload Estimates and Standard Deviations For the Bedford Scale Steps

Recalling the Babbitt and Nystrom criterion (reference 7), there should be no distribution overlap with adjacent scale step means out to one SD. When this criterion is met, there is a reasonable confidence that about 84 percent of the subject population will agree as to the proper ranking of workload intensity levels among the descriptors. Figure 10 shows that this relatively liberal criterion was approached only for steps WL-1 through WL-4 and WL-7 through WL-10. In comparison, the distribution overlaps for steps WL-4 through WL-7 was near total. The conservative conclusion is that no practical differences exist between descriptors WL-4 through WL-7 in terms of the level of workload they express. This was particularly notable because the Bedford's outer decision tree classifies steps WL-4 through WL-6 as *workload not satisfactory without reduction,* whereas, WL-7 is classified as *workload not tolerable for the task.*

Bedford users assume that raters can be trained to accept the prescribed rank order. That is, to accept that the descriptor for a rating of WL-7, for example, intends to characterize a higher level of workload than the descriptor for a WL-4, even though the rater's internal sense of word meaning might suggest otherwise. This, however, amounts to a forcing of the issue. The differences between scale step descriptors should be sufficiently transparent that forced agreement is not necessary. Training is not likely to adequately compensate. Inter-rater response variability will persist because the step descriptors remain ambiguous in spite of the mandated numerical position they occupy. The differences in workload intensity represented by the wording of adjacent scale step descriptors must be clearly distinguishable in terms of verified dependable norms of English meaning. Otherwise the scale falls short of providing a truly ordinal response

continuum. In such case, scale induced confusions will occur as to which step best fits the workload of a given task or mission segment, thus leading to scale induced response variability.

Table 3: Descriptor Combinations Most Frequently Cited as Confusable

Descriptor Combination	Number of Subjects
WL-4 with WL-5	14
WL-4 with WL-6	11
WL-5 with WL-6	9
WL-1 with WL-2	8
WL-6 with WL-7	6
WL-2 with WL-3	4
WL-8 with WL-9	4

Of 41 out of 48 subjects who identified confusable descriptors, the majority cited at least two pairs, but the range was between one and four pairs. The subjects who provided sorts consistent with the Bedford prescription were just as likely or unlikely to cite confusable combinations as those who produced an alternative sort. Table 3 shows the most frequently cited pairs and the number of subjects who cited them. Most of the cited confusions involved steps in the center of the scale with WL-4 versus WL-5 being cited most often. Nevertheless, instances of confusability were evident even for the steps at the high and low ends of the scale. These data substantiate the multiplicity of confusing phrase elements in the Bedford's step descriptors.

The underlying causes of the confusions are a complex semantic issue involving the miss-application of adjectives to modify the meaning of the descriptor's root dimensional terms. The use of more scale steps then the Bedford's workload concept can effectively hold could also be an intervening factor. This latter issue involves the relative distinctiveness of adjacent descriptors versus their combined ability to characterize changes in workload level. The less difference in conveyed meaning that exists between adjacent descriptors the greater the potential sensitivity of the scale for changes in workload level. In counter balance, the closer the descriptors are in meaning the potentially more confusable they become, thus limiting the number of steps a scale can effectively contain.

The 48 subjects in the present study provided numerous detailed comments about the sources of confusion between descriptors. Depth discussion of these data is beyond the scope of the present writing. Still, for notable example, consider the distinguishing differences between *Insufficient Spare Capacity for Easy Attention to Additional Tasks* (WL-4), versus *Reduced Spare Capacity; Additional Tasks Cannot be Given the Desired Amount of Attention* (WL-5). Figure 11 shows that a statistical inversion occurred between these two scale steps. This indicated that the denotative difference and thus ordinal relationship between them is vague and equivocal at best. Subject commentary indicated that the difference in dimensional intensity between *Insufficient Spare Capacity* and *Reduced Spare Capacity* was far from intuitive. Consider also the denotative relationship between *Reduced Spare Capacity* (WL-5) and *Little Spare Capacity* (WL-6).

When evaluated in isolation, the two are definitely ambiguous in terms of the relative intensity of spare capacity they represent. Strong similar denotative problems exist between all steps in the center of the scale. The combination of subject comments and numerical results substantiate that the Bedford's inner 10-step scale is not reliably ordinal across its continuum. Without clearly distinctive differences between the definitions, the scale's metric integrity breaks down. In such case, it becomes a matter of speculation what the rater's criteria for score selection will be. The technical consequences are that neither the raters nor the data analysts can effectively use the definitions as intended. That is, to provide a collectively understood anchoring definition of absolute workload level to mediate both scoring and data analysis.

Find below four general classifications of workload level. The top to bottom sequential ordering among them is by random assignment in this questionnaire.

1. Identify the proper rank order that should exist among them. The classification representing the lowest workload level is to be assigned a value of "1" on the line provided to the right of its definition. The definition representing the highest level of workload will consequently be assigned a value of 4 while the other two receive a value of either "2" or "3" accordingly.

2. Then, using the rank order sort you accomplished in Task 1, identify which among the ten scale-step definitions belong in each classification. Do this by writing their letter keys in the boxes provided below, low to high, left to right. Since we do not know how many definitions you will assign to any classification, you are provided ten boxes for each.

Not possible to complete tasks — Classification ___

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Workload satisfactory without reduction — Classification ___

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Workload not tolerable for the task — Classification ___

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Workload not satisfactory without reduction — Classification ___

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Figure 11: TPS Survey Instructions for the Bedford's Outer Decision Tree

The CTF data provided the initial evidence that psychometric problems existed with the Bedford's inner 10-step continuum. The TPS data subsequently confirmed these results and extended the scope of the study to the Bedford's outer decision tree. Figure 11 shows the most notable additional task that the TPS subjects performed. This item required them to rank the 4 outer decision tree classifications and then indicate which among the inner 10 descriptors should be assigned to each. This task was included to first verify that the outer decision tree itself represented an ordinal continuum and second to explore whether the subjects would classify the inner steps according to the Bedford 'hard-wire' decision tree prescription.

Table 12 shows the results obtained from this task. To briefly recap, the Bedford's outer decision tree affords a choice between four general workload impact classifications. These are:

(1) Workload satisfactory without reduction,
(2) Workload not satisfactory without reduction,
(3) Workload intolerable for the task, and
(4) Not possible to complete tasks.

The TPS students were instructed to first rank-order the classifications from highest workload to lowest, with 1 being lowest and 4 being highest. After that, they identified which of the 10 scale steps belonged to each classification. Each copy of the questionnaire presented the four classifications in a different random order.

Table 12: Steps Assigned to the Decision Tree Classifications by Each Subject

	1	1	1	2	2	2	3	3	3	4
	WL 1	WL 2	WL 3	WL 4	WL 5	WL 6	WL 7	WL 8	WL 9	WL 10
P	1	1	1	2	2	2	2	3	3	4
P	1	1	1	1	1	1	2	2	3	4
P	1	1	1	2	1	2	2	3	4	4
P	1	1	1	2	2	2	2	3	3	4
P	1	1	1	3	2	2	2	3	4	4
P	1	1	1	2	2	3	3	3	4	4
P	1	1	1	2	2	2	3	3	4	4
P	1	1	1	2	2	2	2	3	3	4
P	1	1	1	2	2	2	2	2,3	3,4	3,4
P	1	1	1	2	2	3	2	3	3	4
P	1	1	1	2	2	2	3	3	4	4
P	1	1	1	2	3	2	2	3	3,4	4
P	1	1	1	2	2	2	1	3	4	4
NAV	1	1	1	3	2	2	2	3	3	4
NAV	1	1	1	2	2	2	2	3	3	4
FTE	1	1	1	2	3	2	2	3	3	4
FTE	1	1	1	2	2	2	2	2	3	4
FTE	1	1	1	3	2	2	2	2	2,3	2,3,4
FTE	1	1	1	4	2	2	2	3	3	4
FTE	1	1	1	2	2	1	2	3	3	4
FTE	1	1	1	2	2	2	2	3	4	4
FTE	1	1	1	3	2	2	2	3	3	4
FTE	1	1	1	3	2	2	2	3	3	4
FTE	1	1	1	2	2	2	2	3	3	4
FTE	1	1	1	3	2	2	3	4	4	4
FTE	2	1	1	4	2	3	3	4	4	4

Notes: The shaded cells identify rank orderings that deviated from the Bedford rank order
P = pilot, FTE = flight test engineer, NAV = navigator, and BOOM = boom operator

The top-most row of table 12 identifies the steps that the Bedford assigns to each of the four classifications. The gray cells identify any instance where a student made an assignment that was different. Twenty-six of the 28 TPS students provided data for this item. The other two students arrived late to class and did not have time to complete this final task. Among the 24 who did, there was 100 percent agreement that the outer

decision tree classifications were ordinal as prescribed, but no student was in total agreement with the Bedford as to which scale steps belonged to each classification. Most notably, 23 out of 26 disagreed that scale step 7, *Very Little Spare Capacity but Maintenance of Effort in the Primary Tasks Not in Question*, should belong under Classification 3 *Workload Not Tolerable for the Task*, although, that is the Bedford prescription. In addition, there were 8 instances where WL-4 showed up either in Classification 3 *Not Tolerable* or Classification 4 *Not Possible* rather than classification 2 *Not Satisfactory*, and WL-9 was 11 times assigned to *Not Possible* rather than *Not Tolerable* in further contradiction to the Bedford prescription.

These data show that the prescribed correspondences between the decision tree classifications and the scale's inner 10 steps are also ambiguous. For example, 20 out of 26 subjects did not visualize WL-7 *Very little spare capacity but maintenance of effort in the primary tasks not in question* as corresponding with a state of *Workload Intolerable for the Task*, although that is what the Bedford prescribes. This was the most prominent inconsistency but others were also prominently evident in the data. This suggests that the Bedford's system of decision tree and inner scale steps does not provide a reliable generalized representation of possible real-world workload conditions. Further, it would likely fail on this count even if the inner scale step descriptors were reliably ordinal, which they are not. In fact, two subjects assigned several of the same steps to multiple classifications. In contrast, another assigned all of the steps WL-1 through WL-6 to classification 1, *workload satisfactory without reduction*. Ambiguities between the decision tree classifications and their inner scale steps likely confound the rater's ability to select a truly representative score. It is generally agreed that the relationship between task demand (workload level) and performance is not purely linear (reference 20). Considering this, there is no substantive reason to assume a purely linear relationship between absolute workload impact and absolute workload level either. The present data would seem to strongly indicate that workload impact and workload level cannot be reliably accounted for with a single unified rating, although this is what the Bedford attempts to do.

The tangible consequences are that any given score can amount to an over-estimation or under-estimation of the absolute workload associated with a particular task or mission element. The individual aircrew rater will likely always know the state of workload they want to express with their scoring. Still, if the scale's step descriptors and or their correspondence with the associated decision tree classification are ambiguous, then the ability to render a valid score is compromised. This creates analysis problems because a rating of WL-4 for example, may actually denote nearly the same level of workload as a WL-5, WL-6, or WL-7, although the data analyst has little choice but to assume that the numerically higher score designates a higher workload state. In turn, the analyst has little choice but to assume that the associated decision tree classification provides a valid characterization of the workload impact that the rater intended to convey, which it may not.

The present results do fly in the face of a number of studies that claim validity and reliability for the Bedford (references 15 and 16). The primary issue therefore pertains

to what level of validity and reliability can be realistically ascribed to the scale. Undoubtedly, the outer decision tree's workload impact classifications do amount to an ordinal continuum from low to high workload impact. This coupled with the use of numerical identifiers (WL-1 through WL-10) for the inner scale steps evidently do allow the Bedford to function as an ordinal measure in some rough sense. Nevertheless, the purpose of employing scale step descriptors is to anchor each to a specific and commonly understood definition of absolute workload level. If the scale's inner steps were reliably descriptive of a continuum of 10 distinctively different workload levels, than they would possess that quality independent of the numerical identifiers attached to them. The results from the rank-order sorts clearly show that they do not. Steps WL-4 through WL-7 is not reliably distinguishable in terms of the workload intensity levels they represent. Consequently, the descriptors cannot be used to identify what the absolute change in workload level is between ratings of WL-4 through WL-7 for example. As such, the inner scale can only support scoring in the relative sense of the word. That is, if a rater decided that his or her experienced workload impact was low unsatisfactory, then evidently they would select WL-4. In turn, if it were high unsatisfactory then WL–6 would be selected regardless of what the wording of the associated step descriptor indicated. Given the ambiguousness of the descriptors, this is likely how the scale steps are frequently used.

The fundamental measurement problem is that usage like this amounts to a functional abandonment of the descriptors, thus defeating the only purpose for which they could be intended. That is, to provide clear descriptions of absolute workload level to mediate the score selections of all raters. Without functionally usable anchoring, the raters will be inclined to resort to their own internal criteria for what low, medium, or high workload amounts to. The condition of the step descriptors in the center of the scale is such that the steps may as well have no descriptors at all. As a matter of consistency in scoring behavior, this stands to affect scoring across the satisfactory, unsatisfactory, and, intolerable workload classifications rather equally. The ambiguous correspondences between the decision tree classifications and the inner scale steps can only further compound the problem. For example, subjects who faithfully use the outer decision tree to mediate their scoring may end up abandoning use of the scale step descriptors more or less completely. On the other hand, it has been observed that scores are sometimes assigned without the rater consulting the outer decision tree (reference 14). Although this was not reported in criticism of the scale, it nevertheless suggests that some of the scoring may not always validly reflect correspondence with a specific workload impact classification.

At the time of this writing, the impact of the Bedford's deficiencies on data reliability had not been quantified. Nevertheless, there should be little dispute that the measurement quality of the Bedford is degraded by those deficiencies. If descriptors are integral components in a scale's design, then each must be meaningfully different in terms of the absolute level of dimensional intensity they represent. The Bedford's descriptors certainly do not square well with that standard. In addition, the Bedford's unbalanced bipolar design may be a matter of concern in its own right. Although the scale's four workload impact classifications are facially ordinal, the use of three classifications of unsatisfactory versus only one for satisfactory can cause the difference

between *workload intolerable* and *workload not satisfactory without reduction* to be treated equivocally. Since the outer decision tree itself does not specify the practical operational differences between the two, the associated inner scale should. Otherwise the raters are left to make an unanchored determination about the matter. The indistinctiveness of the inner scale step descriptors makes the determination all the more difficult to accomplish.

Discussion

This paper demonstrates the technical value of using structured psychometric procedures to guide the design and verification of subjective scaling having complex scale step descriptors. A variation on classical rank-order methods was introduced. Its application was demonstrated with case studies of the revised USAFSAM and Bedford workload scales. The two studies showed the method to be flexible, and usable even when subject resources are relatively limited, thus making it suitable for applied situations. Although pair comparisons and successive intervals will accomplish essentially the same ends, the rank order variation presented herein tends to combine the most desirable properties of both.

The use of rating scales is a persistent attribute of T&E where operator-workstation systems are issues of interest. Employed by qualified weapon systems operators, they can be powerful tools for accurately characterizing the strengths and limitations of workstation interfaces, and for identifying and qualifying the impact of usability issues. In order to obtain best value, the scale step descriptors must have clear meaning to the test engineers and data analysts as well as all workstation operators who have occasion to employ the scale. To this end, some general observations about complex unidimensional scaling are worthy of mention. Comparison of figures 6 and 9 show the Bedford descriptors to be wordier then the revised USAFSAM descriptors. During revision of the USAFSAM, the criterion size for the descriptors was informally set at "lucky number seven" plus or minus one. For comparable inter-rater scoring to occur, it must be mediated by the consistent use of descriptor content, not just the numerical position the descriptors occupy on the scale.

All other things being equal, if the descriptors are wordy, the more difficult they are to assimilate and the less likely they are to be regularly used. In addition, informal accounts of rater scoring behavior suggest that descriptor usage may be compromised if the scale steps have numbers affixed to them. That is, the raters can be tempted to use the numbers based solely on what is remembered about the descriptors from prior scoring. The more complex and wordy the descriptors the less reliable the rater's memory is likely to be. Accurate recall stands to be hampered even more if the rater must supply scores in the busy environment of a high fidelity simulation or during an actual flight test sortie. This makes it questionable whether any scale employing descriptors to anchor the scale steps should ever be presented to the rater with the steps numbered. The functional purpose of the numbering should primarily be for statistical convenience during data reduction and analysis. In the absence of numbers, the rater is forced to supply scores based solely on descriptor content. When queried for a score for example, the rater

would then be required to articulate the chosen descriptor in its entirety. For example, when using the revised USAFSAM, the rater would respond to a query with *"Busy, challenging but manageable, adequate time available"* or *"Very busy, difficult to manage, barely able to keep up"* rather than simply responding with, "that was a 4" or "that was a 5". Unnumbered descriptors increase the probability that each score actually represents a well-considered estimate. Of course, this presupposes that the scale step descriptors are well crafted and then rendered in a font size that makes them easily readable under the conditions they are employed. In parallel, if the survey employs a questionnaire to obtain postflight ratings for an array of mission elements, then here also, the scaling should be presented in unnumbered form.

In closing, the two case studies demonstrated the difference between a psychometrically will-refined scale and one that is less well refined. More generally, the resulting data provided some basic insights about what is required for high quality scaling when complex scale step descriptors are used. In scale design, one of the fundamental problems that must be addressed is the tradeoff between scale sensitivity to changes in dimensional intensity versus having scale step descriptors that are unambiguously distinguishable. Selection of descriptive terminology to characterize differences in intensity level between scale steps is not always as intuitive as it might seem. This is particularly so when multiple sub dimensions are incorporated into the descriptor phraseology. The Bedford's attempt to unify workload impact and workload level together in a single scaling solution is a notable case in point. Consulting the published psychometric data should be considered an essential part of any scale design and refinement process. Two survey handbooks that consolidate these kinds of data were cited (References 7 and 8). In any case, some form of structured psychometric verification should be performed before pressing any new or undocumented scale into service. Therefore, the employment of previously used scales not supported with psychometric verification data should be reconsidered in this context.

REFERENCES

1. Cliff, N. "Adverbs As Multipliers." *Psychological Review*, Vol. 66, No. 1, 1959.
2. McIver, J.P. and E.G. Carmines. "Unidimensional Scaling." *Series: Quantitative Applications in the Social Sciences*, Sage Publications Inc; 1981.
3. Guilford, J.P. "Psychometric Methods." New York: McGraw-Hill Book Company, Inc, 1954.
4. Ames, L.L. and E.J. George. *Revision and Verification of a Seven-Point Workload Estimate Scale,* AFFTC TIM-93-01; Air Force Flight Test Center (AFFTC); Edwards AFB; California, 1993.
5. Edwards, Allen L. *Techniques of Attitude Scale Construction* (Chapter 5), New York: Appleton- Century-Crofts, Inc. 1978.
6. Matthews, J. J., C. E. Wright and K. L. Yudowitch *Analysis of the Results of the Administration of Three Sets of Descriptive Adjective Phrases.* Palo Alto: Operations Research Associates (prepared for the U.S. Army Research Institute

for the Behavioral and Social Sciences, Fort Hood, Texas under Contract DAHC19-74-C-0032).

7. Babbitt, B.A. and Nystrom, C.O. *Questionnaire Construction Manual,* U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia; ARI Research Project 89-20, June 1989.

8. Ross, K.C, et el. *Air University Sampling and Survey Design Handbook*, HQ AU/XOPA, Maxwell Air Force Base, Alabama, April 1996.

9. Gawron, V.J. et al *The effect of pyridostigmine bromide on inflight aircrew performance* (USAFSAM-0TR-87-24) Brooks Air Force Base, Texas, School of Aerospace Medicine, January 1988.

10. George, E.J., M. Nordeen, and D. Thrumond. *Combat Talon II Human Factors Assessment* (AFFTC TR 90-36). AFFTC, Edwards Air Force Base, CA, 1991.

11. George, E.J. and S. Hollis. *Scale Validation in Flight Test*. Air Force Flight Test Center, Edwards Air Force Base, California; December 1991.

12. George, E.J. and L.L. Ames, *AC-130U Gunship Workload Evaluation,* AFFTC TR-95-79, Air Force Flight Test Center, Edwards AFB, California 1996.

13. Rosco, A.H. *Assessing pilot workload in flight. Flight Test Techniques*. Proceedings of NATO Advisory Group for Aerospace Research and Development (AGARD-CP-373). Neuily-sur-Seine, France: AGARD, 1984.

14. Roscoe, A.H. *In-Flight assessment of workload using pilot ratings and heart rate*. In A.H. Roscoe (Ed.) The practical assessment of pilot workload. AGARDograph No. 282 (pp. 78-82). Neuilly-sur-Seine, France: AGARD, 1987.

15. MIL-F-6785C, *Military Specification for Flying Qualities of Piloted Airplanes*, U.S. Department of Defense, November 1980, Superseding, Mil-F-8785PhlL-F-8785P, 7 August 1969.

16. Vidulich, M.A. *The Bedford Scale: Does I measure spare capacity?* Proceedings of the American Helicopter Society National Specialist's Meeting: Automation Applications for Rotorcraft, 1988.

17. Vidulich, M.A and M.R. Bortolussi *Control Configuration Study*. Proceedings of the American Helicopter Society National Specialist's Meeting: Automation Application for Rotorcraft, 1988.

18. Steel, R.G.D and J.H. Torrie P*rinciples and Procedures of Statistics: A Biometrical Approach*, McGraw-Hill Book Company, 1980

19. Tsang, P.S. and W. Johnson. *Automation: Changes in cognitive demands and mental Workload.* Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, Ohio: Ohio State University, 1987.