APPLIED PSYCHOLOGY: AN INTERNATIONAL REVIEW, 2004, *53* (1), 61–86

Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods

Susana Rubio,* Eva Díaz, Jesús Martín and José M. Puente

Universidad Complutense de Madrid, Spain

Cette recherche mesure plusieurs propriétés psychométriques (l'ingérence, la sensibilité, la valeur diagnostique et la validité) de trois instruments multidimensionnels de l'évaluation de la charge de travail subjective: le NASA Task Load Index (TLX), le Subjective Workload Assessment Technique (SWAT) et le Workload Profile (WP). Sujets ont réalisé deux tâches de laboratoire séparément (tâches simples) et simultanément (tâches doubles). D'après l'analyse de variance, les trois instruments ne présentent pas de différences au niveau de l'ingérence, mais WP bénéficie d'une sensibilité exceptionnelle aux manipulations des différentes tâches. On a fait appel à une analyse canonique discriminante pour apprécier la valeur diagnostique de chacun des trois instruments. Les résultats de l'analyse ont prouvé que les trois évaluations multidimensionnelles avaient fourni une information diagnostique sur la nature des exigences des tâches qui était cohérente avec leur description *a priori*. Toutefois, la valeur diagnostique du WP s'est révélée nettement supérieure à celles du TLX ou du SWAT. Pour évaluer la validité concurrente de chaque instrument avec la performance aux tâches, on a calculé les corrélations de Pearson entre chaque performance et chaque mesure de la charge subjective. On a enfin calculé les corrélations de Pearson entre les trois mesures de charge subjective pour évaluer la validité convergente des instruments. Les trois coefficients ont été positifs et proche du maximum, soulignant ainsi la forte validité convergente des trois outils retenus pour cette recherche. On a aussi comparé les conditions d'application et l'acceptabilité par les sujets. On mentionne pour terminer les implications pratiques de ces trois sortes d'évaluation.

The present research evaluates several psychometric properties (intrusiveness, sensitivity, diagnosticity, and validity) of three multidimensional subjective workload assessment instruments: the NASA Task Load Index (TLX), the Subjective Workload Assessment Technique (SWAT), and the Workload Profile (WP). Subjects performed two laboratory tasks separately (single task)

* Address for correspondence: Susana Rubio, Department of Differential and Work Psychology, Faculty of Psychology, Universidad Complutense de Madrid, Campus de Somosaguas, 28223 Madrid, Spain. Email: srubiova@psi.ucm.es

and simultaneously (dual task). The results of the ANOVAs performed showed that there are no differences with regard to the three instruments' intrusiveness, and that among the three subjective workload instruments WP has an outstanding sensitivity to the different task manipulations. To evaluate the diagnosticity of each of the three instruments canonical discriminant analysis was used, and this demonstrated that the three multidimensional ratings provided diagnostic information on the nature of tasks demands that was consistent with the a priori task characterisation. However, the diagnostic power of WP was clearly superior to that obtained using TLX or SWAT. Pearson correlations between each performance and each subjective workload measure were calculated to evaluate the concurrent validity of each instrument with task performance, and to assess the convergent validity of the instruments. The three coefficients were positive and near to one, showing the high convergent validity of the three instruments considered in this research. Implementation requirements and subject acceptability were also compared. Finally, practical implications on the three assessment approaches are mentioned.

1. INTRODUCTION

Currently, the evaluation of mental workload is a key point in the research and development of human–machine interfaces, in search of higher levels of comfort, satisfaction, efficiency, and safety in the workplace. These are the major goals of ergonomics.

In order to ensure the safety, health, comfort, and long-term productive efficiency of the operator, a reasonable goal is to regulate task demands so that they neither underload nor overload an individual. Although the dangers of overload have long been recognised, many of our recent concerns are with the stress of underload and boredom (Becker, Warm, Dember, & Hancock, 1991; Hancock & Warm, 1989), particularly as operations become the subject of progressively increased automation.

Applied research has paid much attention to mental workload study during the last few years. Many questions may be posed. To what extent is the operator involved? How complex is the task? Is the operator able to perform additional tasks at the same time as the main one? Is s/he able to respond to any particular stimuli? How is the operator feeling at the time of performing his/her tasks? Psychology has long been trying to find the answer to these questions, making a great contribution to the study and evaluation of mental workload (Wickens, 1992), which is commonly defined as the difference between cognitive demands of a particular job or task and the operator's attention resources.

A number of tools for the evaluation and prediction of mental workload exist. Most of these methods fall into the three following categories (Meshkati, Hancock, & Rahimi, 1992): (a) performance-based measures, (b) subjective measures, and (c) physiological measures. The performance-based measures are grounded on the assumption that any increase in task

difficulty will lead to an increase in demands, which will decrease performance. Subjective procedures assume that an increased power expense is linked to the perceived effort and can be appropriately assessed by individuals. Physiological indexes assume that the mental workload can be measured by means of the level of physiological activation.

The suitability of the procedures for the evaluation of mental workload depends on the extent to which they meet the following requirements (Eggemeier, Wilson, Kramer, & Damos, 1991):

1. *Sensitivity:* A tool's power to detect changes in task difficulty or demands.
2. *Diagnosticity*: This involves not only the identification of changes in workload variation but also the reason for those changes.
3. *Selectivity/Validity*: The index must be sensitive only to differences in cognitive demands, not to changes in other variables such as physical workload or emotional stress, not necessarily associated with mental workload.
4. *Intrusiveness*: The measure should not interfere with the primary task performance, the load which is the actual object of evaluation.
5. *Reliability*: The measure must reflect consistently the mental workload.
6. *Implementation requirements*: Including aspects such as time, instruments, and software for the collection and analysis of data.
7. *Subject acceptability*: This refers to the subject's perception of the validity and usefulness of the procedure.

Subjective measures are becoming an increasingly important tool in system evaluations and have been used extensively to assess operator workload. The reasons for the frequent use of subjective procedures include their practical advantages (ease of implementation, non-intrusiveness) and current data which support their capability to provide sensitive measures of operator load. As human–machine systems have become more complex and automated, evaluations based on the operator's performance have become prohibitively difficult, and the need to assess subjective mental workload has become critical.

Many subjective procedures exist to measure mental workload. The most outstanding among them are the Cooper-Harper Scale (Cooper & Harper, 1969), the Bedford Scale (Roscoe, 1987; Roscoe & Ellis, 1990), the SWAT (Subjective Assessment Technique) (Reid & Nygren, 1988) and the NASA-TLX (Task Load Index) (Hart & Staveland, 1988). Recently Tsang and Velazquez (1996) have proposed a new multidimensional subjective workload assessment instrument (Workload Profile), which portends to be a technique with an elevated diagnosticity. This vast range of procedures and techniques for the evaluation of subjective mental workload sets confusion among psychologists who very often lack the information to choose the assessment

technique that best fits the situation under study. In an attempt to overcome these difficulties, this research pursues two main goals:

1. To study the psychometric and methodological characteristics of three instruments for the subjective evaluation of mental workload, the NASA-TLX, the SWAT, and the Workload Profile (WP), comparing each with the others in terms of their sensitivity, diagnosticity, validity, intrusiveness, implementation requirements, and operator acceptability. The main goal is to compare more established measures (TLX and SWAT) with the new WP.
2. To issue some guidelines for the appropriate use of the assessment instruments under consideration. This could help to choose the most suitable tool for a particular situation.

2. METHOD

2.1 Subjects

A sample of 36 students of psychology in the Complutense University of Madrid volunteered to participate in the study. Ages ranged from 20 to 24 years and all were right-handed. About two-thirds were females, and the rest were males. Subjects were randomly assigned to three groups of the same size: 12 subjects filled out the TLX questionnaire, another 12 answered the SWAT questions, and the remaining 12 filled in the Workload Profile.

2.2 Workload Measures

2.2.1 NASA Task Load Index (TLX). The NASA Task Load Index (Hart & Staveland, 1988) uses six dimensions to assess mental workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Figure 1 shows the definitions of NASA-TLX dimensions. Twenty-step bipolar scales are used to obtain ratings for these dimensions. A score from 0 to 100 (assigned to the nearest point 5) is obtained on each scale. A weighting procedure is used to combine the six individual scale ratings into a global score; this procedure requires a paired comparison task to be performed prior to the workload assessments. Paired comparisons require the operator to choose which dimension is more relevant to workload across all pairs of the six dimensions. The number of times a dimension is chosen as more relevant is the weighting of that dimension scale for a given task for that operator. A workload score from 0 to 100 is obtained for each rated task by multiplying the weight by the individual dimension scale score, summing across scales, and dividing by 15 (the total number of paired comparisons).

TITLE	ENDPOINTS	DESCRIPTIONS
MENTAL DEMAND	Low/High	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	Low/High	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	Low/High	How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
PERFORMANCE	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals?
EFFORT	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	Low/High	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

FIGURE 1. Rating scale definitions and endpoints from the NASA Task Load Index.

Development of the TLX has implied an important and vast program of laboratory research (Hart & Staveland, 1988), and the instrument's sensitivity has been demonstrated using a great variety of tasks. TLX has been applied successfully in different multitask contexts, as for example in real (Shively, Battiste, Matsumoto, Pepiton, Bortolussi, & Hart, 1987) and simulated flight tasks (Battiste & Bortolussi, 1988; Corwin, Sandry-Garza, Biferno, Boucek, Logan, Jonsson, & Metalis, 1989; Nataupsky & Abbott, 1987; Tsang & Johnson, 1989; Vidulich & Bortolussi, 1988), in air combat (Bittner, Byers, Hill, Zaklad, & Christ, 1989; Hill, Byers, Zaklad, Christ, & Bittner, 1988; Hill, Byers, Zaklad, & Christ, 1989), and using remote-control vehicles (Byers, Bittner, Hill, Zaklad, & Christ, 1988). Sawin and Scerbo (1995) used the TLX technique to analyse the effects of instruction type and boredom proneness on vigilance tasks performance.

2.2.2 Subjective Workload Assessment Technique (SWAT). The Subjective Workload Assessment Technique (Reid & Nygren, 1988) is a subjective rating technique that uses three levels: (1) low, (2) medium, and (3) high, for each of three dimensions of time load, mental effort load, and psychological stress load to assess workload. Figure 2 shows the SWAT rating scale dimensions. It uses conjoint measurement and scaling techniques to develop a single, global rating scale with interval properties.

I. Time Load
1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
2. Occasionally have spare time. Interruptions or overlap among activities occur infrequently.
3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.
II. Mental Effort Load
1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainly, unpredictability, or unfamiliarity. Considerable attention required.
3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.
III. Psychological Stress Load
1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.
3. High to very intense stress due to confusion, frustration, or anxiety. High extreme determination and self-control required.

FIGURE 2. Subjective Workload Assessment Technique (SWAT) rating scale dimensions.

The use of SWAT entails three distinct steps. The first is called scale development. All possible combinations of three levels of each of the three dimensions are contained in 27 cards. Each operator sorts the cards into the rank order that reflects his or her perception of increasing workload. Conjoint scaling procedures are used to develop a scale with interval properties. The second step is the event-scoring, that is the actual rating of workload for a given task or mission segment. In the third step, each three-dimension rating is converted into numeric scores between 0 and 100 using the interval scale developed in the first step.

The sensitivity of SWAT has been shown using a variety of tasks (memory tasks, manual control tasks, display monitoring). Reid and Nygren (1988) expound all of the laboratory studies performed to develop the instrument. SWAT has been applied successfully in the mental workload assessment of several aircraft multitask conditions (Battiste & Bortolussi, 1988; Corwin, 1989; Corwin et al., 1989; Gawron, Schiflett, Slater, Miller, & Ball, 1987; Haworth, Bivens, & Shively, 1986; Kilmer, Knapp, Bursal, Borresen, Bateman, & Malzahn, 1988; Nataupsky & Abbott, 1987; Schick & Hahn, 1987; Skelly & Purvis, 1985; Thiessen, Lay, & Stern, 1986), in nuclear plant simulations (Beare & Dorris, 1984), and using military tank simulators (Whitaker, Peters, & Garinther, 1989). Also, SWAT has been

used, at the same time as other measures, to assess the mental workload of different systems of air defense (Bittner et al., 1989) and remote control vehicles (Byers et al., 1988).

2.2.3 Workload Profile (WP). Tsang and Velazquez (1996) have introduced and evaluated a new multidimensional instrument to assess subjective mental workload, based on the multiple resource model of Wickens (1987). Their instrument (Workload Profile) tries to combine the advantages of secondary task performance based procedures (high diagnosticity) and subjective techniques (high subject acceptability and low implementation requirements and intrusiveness). As Tsang and Velazquez recognised, the Workload Profile technique needs to be the object of more detailed and extensive research about its properties.

The Workload Profile (Tsang & Velazquez, 1996) asks the subjects to provide the proportion of attentional resources used after they had experienced all of the tasks to be rated. The tasks to be rated are listed in a random order down the column and the eight workload dimensions are listed across the page (see Figure 3). The workload dimensions used in this technique can be defined by the resource dimensions hypothesised in the multiple resource model of Wickens (1987): perceptual/central processing, response selection and execution, spatial, processing, verbal processing, visual processing, auditory processing, manual output, and speech output. Subjects have available to them the definition of each dimension at the time of the rating. In each cell on the rating sheet, subjects provide a number between 0 and 1 to represent the proportion of attentional resources used in a particular dimension for a particular task. A rating of "0" means that the task placed no demand on the dimension being rated; a rating of "1" means that the

Workload Dimensions								
	Stage of processing		Code of processing		Input		Output	
Task	Perceptual/ Central	Response	Spatial	Verbal	Visual	Auditory	Manual	Speech
m2								
m2s1								
m2s3								
m4								
m4s1								
m4s3								
s1								
s3								

FIGURE 3. Workload Profile rating sheet.

task required maximum attention. The ratings on the individual dimensions are later summed for each task to provide an overall workload rating.

2.3 Tasks

Two experimental tasks were used: a Sternberg's memory searching task, and a tracking task. Both were implemented through a computer program.

2.3.1(a) Sternberg's Memory Searching Task. Subjects were asked to memorise a set of consonants at the beginning of each trial. Two levels of objective difficulty were set according to the number of letters to be memorised: 2 letters (m2) in the easiest set, 4 letters (m4) in the more difficult set. The letters were randomly chosen for each trial. During the experimental trials, subjects were asked if the letter displayed was the one they had memorised. Subjects responded with their left hand. For each trial, data were collected for hits, errors, and response time. Since there were not enough omissions for analysis and there was no evidence of a speed–accuracy tradeoff, only the response time measure was used as the Sternberg task dependent variable.

2.3.2(b) Tracking Task. Subjects had to keep the cursor within a moving path, using the right and left cursor keys on the keypad. The width of the path was handled to objectively measure the task difficulty: level 1, the narrowest path (the most difficult, s1), level 3, the widest path (the easiest, s3). For this task subjects were asked to use their right hand. The dependent variable for the tracking task was Root Mean Square Error (RMSE).

2.3.3(c) Dual Tasks. A combination of the aforementioned tasks gave rise to four dual tasks: m2s1, m2s3, m4s1, and m4s3. Subjects were asked to evenly pay attention to both tasks trying to do the best they could.

2.4 Variables and Procedure

Four variables were taken into account: (a) the difficulty of the memory task (m2m4), (b) the complexity of the tracking task (s3s1), (c) the task condition (single vs. dual), and (d) the tool used to measure subjective mental workload (TLX, SWAT, WP). The first three variables were repeated measures and the last variable was a between-subjects measure. There were two dependent variables: performance measures for each of the tasks, and the subjective measures of the mental workload scales.

The sample was randomly divided into three experimental groups, one for each workload instrument. Data were collected in the Work Psychology Laboratory of the Faculty of Psychology of the Complutense University of Madrid. All of the participants performed all the experimental tasks following

the same order: m2, m4, s3, s1, m2s3, m4s3, m2s1 and m4s1. Subjects were informed of the tasks condition at the beginning of the trial and they were instructed to treat the two tasks in a dual-task condition as a unit. For each group, data were collected in one different experimental session of about one hour.

For each single task, one measure of the subject performance and the mental workload were collected. For each dual task, two measures of the subject performance and one of the mental workload were obtained. In all conditions, mental workload measures were taken immediately after the task was performed. The time needed to apply the three mental workload instruments was similar for all groups (1 hour for TLX and WP, and 70 minutes for SWAT). The scale development phase of SWAT was quite tiring for the subjects.

3. RESULTS

3.1 Intrusiveness

An ANOVA was carried out on performance measures in order to check the existence of performance differences associated with the workload assessment method. No significant differences ($p > .05$) were found for any of the performance variables used in the study (see Tables 1 and 2). Bearing in mind that the three instruments measuring mental workload are paper-and-pencil techniques, administered once the subject has performed the experimental task, we can assume the intrusiveness of these scales to be almost 0. In addition, this result could show the similarity of mean capacity of the three groups of subjects.

TABLE 1
Results of ANOVAs Comparing the Performance Levels for the Three Groups

	F	*Sig.*
M2_time	0.121	0.887
M4_time	0.479	0.623
M2S3_time	0.127	0.881
M4S3_time	0.459	0.636
M2S1_time	2.288	0.070
M4S1_time	2.876	0.061
S3_rmse	1.101	0.344
S1_rmse	0.470	0.629
M2S3_rmse	1.550	0.227
M4S3_rmse	0.086	0.917
M2S1_rmse	0.906	0.414
M4S1_rmse	1.122	0.338

TABLE 2
Performance Means and Standard Deviations for Each Group

		Mean	*SD*
M2_time	TLX	142.18	56.52
	SWAT	150.11	94.93
	WP	136.15	53.34
M4_time	TLX	131.66	39.70
	SWAT	138.82	69.84
	WP	119.36	28.21
M2S3_time	TLX	135.77	28.22
	SWAT	130.13	43.79
	WP	139.97	35.27
M4S3_time	TLX	144.39	29.31
	SWAT	130.31	29.08
	WP	145.37	56.97
M2S1_time	TLX	155.80	33.83
	SWAT	122.16	24.79
	WP	142.50	38.96
M4S1_time	TLX	164.31	37.92
	SWAT	133.46	36.71
	WP	143.85	25.22
S3_rmse	TLX	2.74	3.11
	SWAT	3.16	3.62
	WP	4.66	3.23
S1_rmse	TLX	5.79	1.88
	SWAT	6.68	3.04
	WP	5.90	2.30
M2S3_rmse	TLX	4.16	1.49
	SWAT	3.71	2.83
	WP	6.78	7.31
M4S3_rmse	TLX	3.26	2.95
	SWAT	3.24	3.39
	WP	3.70	2.73
M2S1_rmse	TLX	7.47	4.30
	SWAT	6.51	2.23
	WP	5.63	3.17
M4S1_rmse	TLX	6.11	2.47
	SWAT	5.02	1.69
	WP	6.35	2.65

TABLE 3
Workload Means and Standard Deviations for Each Instrument

		Mean	*SD*
M2	TLX	7.44	6.58
	SWAT	4.07	1.84
	WP	1.24	0.79
M4	TLX	8.81	7.55
	SWAT	4.09	1.33
	WP	1.60	0.91
S3	TLX	16.53	11.01
	SWAT	5.70	2.37
	WP	1.72	0.81
S1	TLX	25.97	10.95
	SWAT	7.32	1.83
	WP	2.61	0.93
M2S3	TLX	32.42	18.30
	SWAT	9.73	2.39
	WP	2.96	0.97
M4S3	TLX	33.17	17.06
	SWAT	10.94	3.15
	WP	3.30	0.82
M2S1	TLX	39.30	16.40
	SWAT	13.53	3.01
	WP	3.44	0.82
M4S1	TLX	46.28	17.61
	SWAT	14.51	4.13
	WP	4.32	1.02

3.2 Sensitivity of Mental Workload Instruments

Analyses of variance were performed to show the sensitivity of each mental workload assessment tool. The aim was to find out to what extent the global indices of mental workload varied as a function of objective changes in the difficulty of both single and dual tasks. Table 3 displays means and standard deviations for overall scores of mental workload and Table 4 shows the results of ANOVAs for each instrument—TLX, SWAT, and WP.

The effects of memory set size and path width and their interaction resulted in significant WP scores in all cases, in both single and dual task conditions. The results obtained for TLX and SWAT ratings were different. In the single task condition, TLX and SWAT were sensitive only to path width manipulation. In the dual task condition, there were no significant effects of interaction for these two instruments.

TABLE 4
Summary of ANOVAs for Each Instrument

Task	*Variable*	*Instrument*	F(1,11)	p
Single	Memory set size	TLX	1.09	0.319
		SWAT	0.00	0.976
		WP	28.72	0.000**
	Path width	TLX	48.86	0.000**
		SWAT	14.20	0.003**
		WP	50.43	0.000**
Dual	Memory set size	TLX	13.81	0.003**
		SWAT	4.52	0.057
		WP	70.39	0.000**
	Path width	TLX	43.67	0.000**
		SWAT	18.14	0.001**
		WP	30.93	0.000**
	Interaction	TLX	4.86	0.500
		SWAT	0.03	0.864
		WP	6.22	0.030*

* $p < .05$; ** $p < .01$.

TABLE 5
Pearson Correlation Coefficients Between the Three Global Scores of Mental Workload

	TLX	*SWAT*	*WP*
TLX	1.0000	0.9817	0.9863
SWAT	0.9817	1.0000	0.9720
WP	0.9863	0.9720	1.0000

All were significant with $p < .001$.

3.3 Validity of Mental Workload Instruments

3.3.1 Convergent Validity. An average of mental workload values for each instrument and for each task was estimated and Pearson correlation coefficients were computed between the global scores obtained from the three instruments. Results are shown in Table 5. All correlation coefficients were positive and statistically significant ($p < .001$). Therefore, convergent validity appears to be very high for the three instruments.

3.3.2 Concurrent Validity. For each instrument, Pearson correlation coefficients were computed between global workload scores and each

TABLE 6
Correlation Coefficients Between Mental Workload and Performance

	Time	*RMSE*
TLX	0.751**	0.653**
SWAT	0.792**	0.292*
WP	0.727**	0.300*

* $p < .05$; ** $p < .01$.

performance measure. Table 6 shows these correlation coefficients. According to these results, concurrent validity was demonstrated by significant correlations between each of the performance measures and the three overall subjective measures. However, the results obtained were different for each performance measure. In considering the response time, the correlation values were high and very similar, but correlations with RMSE were different among the three overall ratings. Thus, RMSE had significantly higher correlation with NASA-TLX ratings than with the SWAT and the WP ratings. In short, TLX was the instrument that correlated higher with performance.

3.4 Diagnosticity

Stepwise discriminant analyses were performed to determine to what extent mental workload profiles allow discrimination between tasks. Workload profiles for each task were obtained through subject evaluation of every dimension of the assessment instruments. Both single and dual tasks were taken into account in the same analysis. Results are presented separately for each instrument.

3.4.1 NASA-TLX Diagnosticity. Figure 4 shows mental workload profiles for each task. In the figure, it can be seen that, in general, the values assigned to all the dimensions were higher as the task difficulty increased. However, this enlargement was smaller for physical demand. Mental demand received the highest estimations for all the tasks, followed by effort and temporal dimensions. Although time to perform every task was not limited, subjects estimated temporal demand as an important source of workload, especially for the m4s1 task, the most difficult task.

Figure 5 shows the centroids for each task in the space delimited by both discriminant functions. The first discriminant function (Rc = 0.7225, per cent of variance = 98.02%) discriminates between both task conditions, single and dual. All single tasks are located at the left-hand side of the origin (negative values along the horizontal axis), whereas dual tasks are located at the right-hand

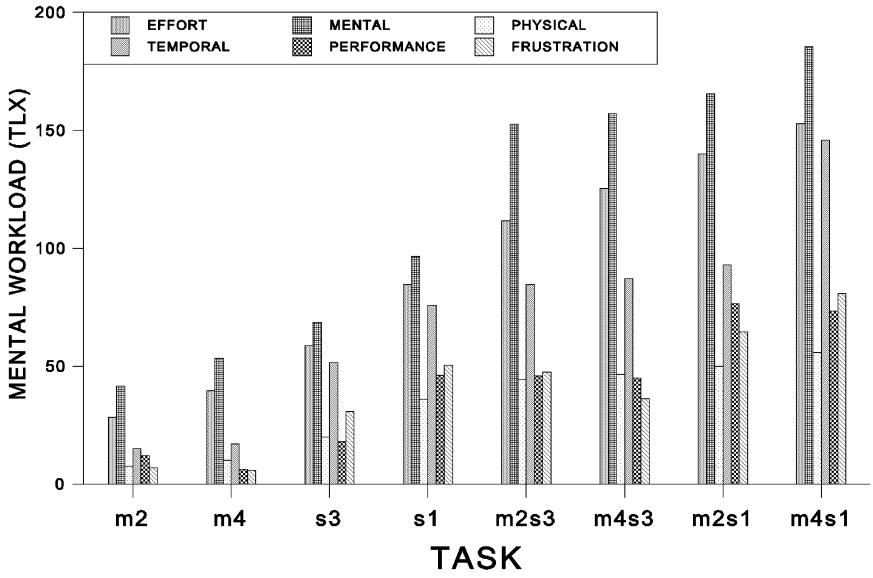

FIGURE 4. Mental workload profiles obtained using TLX for each task.

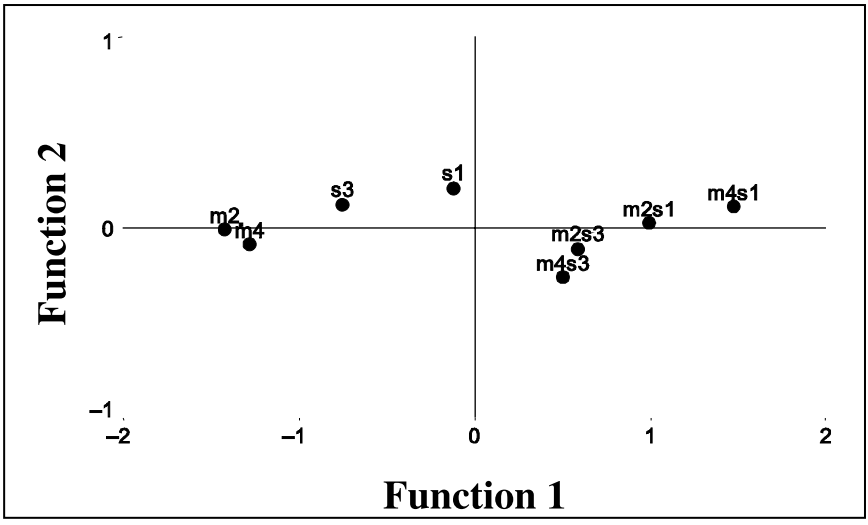

FIGURE 5. Discriminant function centroids for different tasks conditions (TLX).

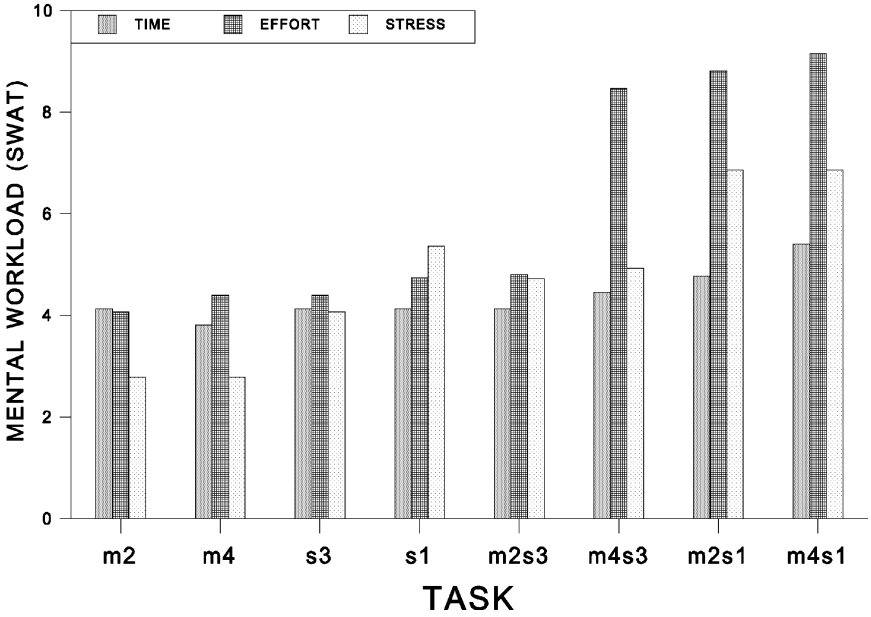

FIGURE 6. Mental workload profile (SWAT) for each task.

side (positive values along the horizontal axis). The insignificant discriminant power of the second function is clear attending to the proximity of the eight tasks to the horizontal (similar coordinates on the vertical axis) and the per cent of variance explained (Rc = 0.1469, per cent of variance = 1.98%).

3.4.2 SWAT Diagnosticity. Figure 6 shows the mental workload profiles obtained for each task. In this case, the time dimension received similar values for all single and dual tasks. The variations observed for stress dimension were smaller than those obtained for effort, but they follow a similar trend. In all single task and in most easy dual task (m2s3) conditions, the effort dimension received significantly smaller values compared to the three most difficult dual tasks.

Centroids for each task in the bi-dimensional space appear in Figure 7. As with the TLX, the first function discriminates between both conditions of task, single and dual (Rc = 0.8809, per cent of variance = 89.45%). All single tasks are located at the left-hand side of the origin (negative values along the horizontal axis), whereas dual tasks are located at the right-hand side (positive values along the horizontal axis). The second function allows us to distinguish between memory and tracking tasks but its discriminant power is smaller (Rc = 0.5386, per cent of variance = 10.55%).

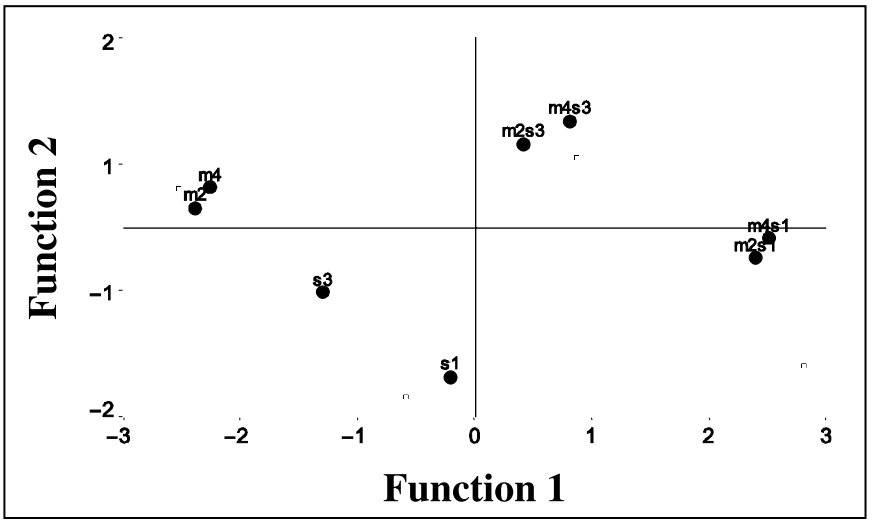

FIGURE 7. Discriminant function centroids for the different tasks (SWAT).

3.4.3 WP Diagnosticity. Likewise with the other two assessment instruments, the first step consisted of obtaining mental workload profiles for each task. Figure 8 shows those profiles. In single memory task conditions, the six dimensions received more elevated values in m4 than in m2 and in the two cases subjects estimated higher perceptual/central, verbal, and visual than manual, response, and spatial dimensions. For tracking single tasks, the more difficult version (s1) obtained greater estimations in all the dimensions with the exception of verbal. Response, spatial, visual, and manual were estimated to be significantly higher than the other two dimensions.

The profiles for dual tasks showed several aspects. First, the most difficult dual task (m4s1) received the highest estimations for the six dimensions. With regard to the perceptual/central processing dimension, the two dual tasks of moderate difficulty (m4s3 and m2s1) received intermediate estimations compared to m2s3 (the least difficult dual task) and m4s1 (the most difficult dual task). The response dimension obtained the smallest value in the m2s1 condition. The two dual tasks, m2s3 and m2s1, showed similar ratings in the verbal dimension, which was significantly higher for m4s3 and m4s1 tasks. In general, spatial, visual, and manual dimensions increased as tracking difficulty was more elevated.

Centroids for each task in the space delimited by both discriminant functions are depicted in Figure 9. It can be seen in the figure that it is the second function again which discriminates between task conditions: single

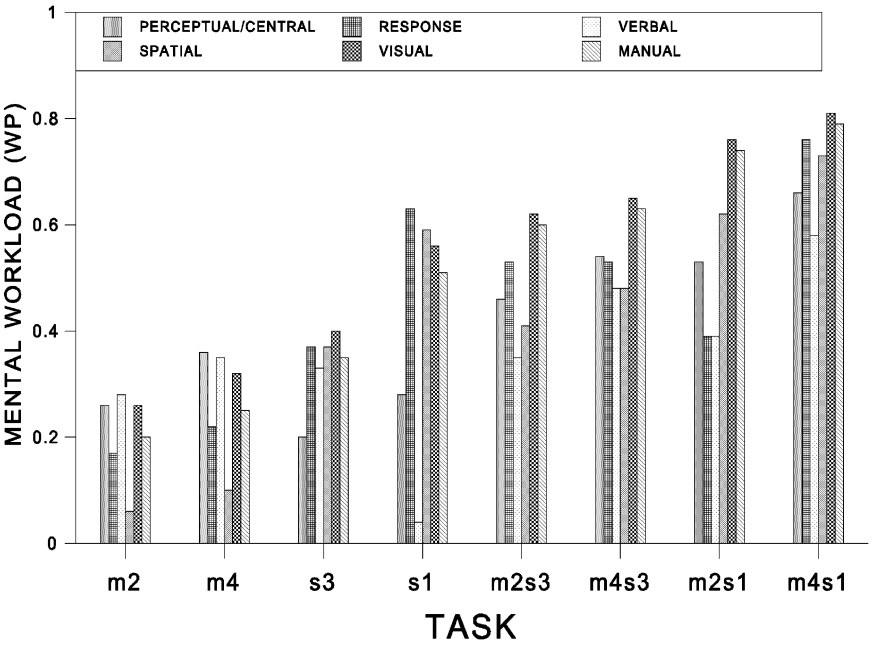

FIGURE 8. Mental workload profile (WP) for each task.

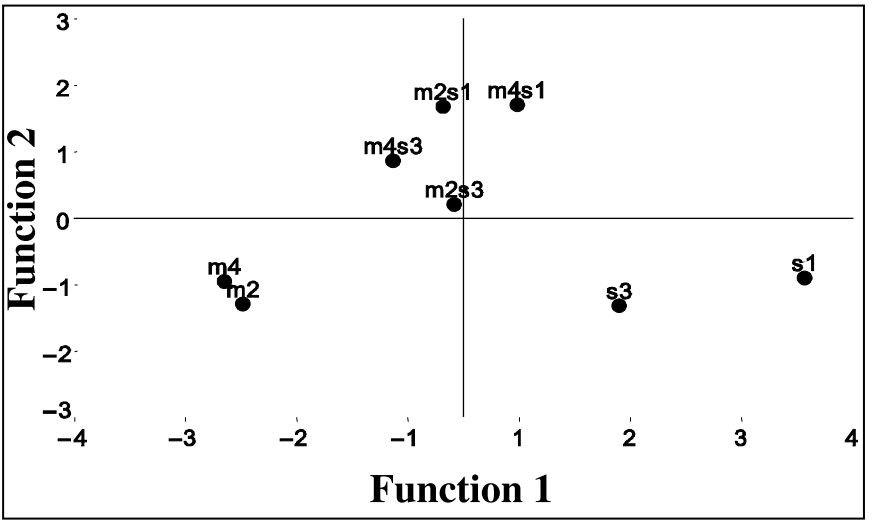

FIGURE 9. Discriminant function centroids (WP).

vs. dual (Rc = 0.7855, per cent of variance = 37.97%). All single tasks are located under the axis origin (negative values along the vertical axis), whereas dual tasks are located in the upper half (positive values along the vertical axis). Within the single task condition, the first discriminant function (horizontal axis) (Rc = 0.8587, per cent of variance = 62.03%) separates memory task (negative values along the horizontal axis) from tracking tasks (positive values along the vertical axis).

Concerning the dual task condition, differentiation among the four tasks by the first discriminant function is lower: the most complex dual task (m4s1) is the only one to obtain positive values in this dimension (horizontal axis), while values for the other three dual tasks are negative.

4. DISCUSSION

The main goal of this study was to compare the properties of three multidimensional subjective workload instruments: two rating scales which currently dominate the literature—NASA-TLX and SWAT—and a new method based on the multiple resources model proposed by Wickens (1984)—Workload Profile (WP). Several researchers have previously compared the two first techniques (Hendy, Hamilton, & Landry, 1993; Hill, Iavecchia, Byers, Bittner, Zakland, & Christ, 1992; Nygren, 1991; Vidulich & Tsang, 1985, 1986), whereas the last has been evaluated in only one study (Tsang & Velazquez, 1996).

The majority of the research comparing the psychometric properties of the different subjective workload instruments has been mainly focused on the evaluation of sensitivity and predictive/concurrent validity. In this paper, the three techniques were compared with respect the following issues: intrusiveness, sensitivity, validity, diagnosticity, implementation requirements, and operator acceptability. In general, most researchers assume that in subjective techniques intrusiveness does not represent a significant problem; most applications require rating scale completion subsequent to task performance and, therefore, present no intrusion problem. However, differences in the application procedure among the three instruments considered could have an effect on subject's performance due to two aspects: (a) differences between the subject's task required by TLX and SWAT in the pair comparisons phase (TLX) and the scale development phase (SWAT), and (b) the presence (in TLX and SWAT) or absence (from WP) of a previous phase to task performance. In this sense, the results of the ANOVA performed permit us to conclude that there are no differences as regards the three instruments' intrusiveness.

Taking into account diagnosticity, subjective mental workload instruments are traditionally not considered diagnostic. However, studies suggest that multidimensional techniques provide valuable diagnostic information about

sources of mental workload (Hart & Staveland, 1988; Reid & Nygren, 1988), but these instruments do not permit us to obtain data about the attentional resources demanded by a particular task.

In the development of the Workload Profile, one of the most important goals was to obtain an instrument that would provide a diagnostic workload profile useful to describe the way in which the task is demanding. In contrast to previous research, we examined SWAT and TLX diagnosticity as the degree to which the ratings in each dimension of workload allow us to discriminate among the different task conditions; in other words, if they permit us to distinguish the attentional resources demanded by each task condition. Following this process we could also compare the diagnosticity of the three instruments.

As in Tsang and Velazquez (1996), diagnosticity was examined using discriminant analysis. With respect to WP, these authors found that from the subjective workload profiles it was possible to differentiate among the several tasks in ways that were consistent with the multiple resource model. Their results suggested that subjects were able to report adequately about the nature of the resources that a particular task demands.

The results obtained in our study were very similar to the ones found by Tsang and Velazquez (1996). In this sense, the workload profiles resulting from WP revealed that:

- Tracking tasks demanded mainly spatial resources whereas the demand of memory tasks was principally verbal.
- Variations in objective difficulty proportionally increased the demands in perceptual/central and response processing.
- Dual task conditions increased the perceptual/central, verbal, and response processing demands compared to the single task conditions.

With respect to TLX and SWAT, the results also showed that these techniques have a certain degree of diagnosticity similar to, but less powerful than, that the provided by WP. In this sense, TLX was able to distinguish only between single and dual tasks, whereas SWAT and WP were able to distinguish also between memory and tracking tasks. The per cent of variance explained by the two discriminant functions obtained showed that SWAT discriminates better between task conditions (single–dual), whereas the discriminant usefulness of WP is higher if we are interested in task differentiation (memory task–tracking task).

Regarding the instruments' validity, the relations among the global scores provided by the three techniques were positive and near to one, indicating that all of them assess the same theoretical concept. Haworth, Bivens, and Shively (1986) included both SWAT ratings and NASA Bipolar scale ratings (precursor of the NASA-TLX) in an investigation of combat nap-of-the-earth flight in a helicopter simulator and, like us, they found that the

ratings resulting from the two procedures were significantly correlated with each other.

Furthermore, the concurrent validity of subjective workload assessment was examined by the degree of agreement between the subjective workload and performance measures. In this sense, NASA-TLX obtained an acceptable correlation with the two performance measures, indicating that this instrument shows a good concurrent validity. However, SWAT and Workload Profile global scores were more highly associated with response time than with tracking error. A possible explanation, but not the only one, for the small correlation between RMSE and SWAT and WP scores could be that subjects estimated workload focusing more on the time needed to perform the tasks than on the path deviations, but this hypothesis needs to be confirmed. These results emphasise the need for the development of better human performance models upon which subjective mental workload metrics can be based and, as Tsang and Velazquez (1996) have pointed out, it is not yet sufficiently clear how the information elicited by subjects should be combined for predicting performance.

Finally, sensitivity of the three instruments was examined using analyses of variance. The results obtained in the present research are consistent with findings provided by most of the previous studies. For example, Battiste and Bortolussi (1988) compared SWAT and NASA-TLX sensitivity with easy and difficult flight scenarios and segments within those scenarios. Both instruments resulted in significantly different ratings between scenarios and flight segments, but TLX proved sensitive to some mental workload differences not discriminated by SWAT. The Corwin et al. (1989) study in a part-task commercial aircraft simulator showed that both techniques produced essentially the same pattern of sensitivity because the two procedures discriminated the workload associated with the two flight levels distinguished in the evaluation, and also demonstrated differences among some flight phases within the more difficult flight condition. Nataupsky and Abbott (1987) reported similar data from a flight simulation experiment in which post-flight SWAT and TLX ratings successfully discriminated workload levels associated with different flight path conditions. Hill et al. (1992) in a flight simulation task compared four subjective workload rating scales, including SWAT and TLX, and found that TLX had the highest sensitivity among the scales, followed by Overall Workload (OW), and the Modified Cooper-Harper Scale (MCH) and SWAT.

In the laboratory environment, Vidulich and Tsang (1986) compared SWAT and NASA Bipolar ratings in both single and dual task versions of a one-dimensional compensatory tracking task and a spatial transformation task and they concluded that both techniques demonstrated essentially the same results, and therefore the use of either instrument as an operational tool appeared justified. Hancock and Warm (1989) have also reported

agreement between SWAT and TLX ratings in a laboratory experiment conducted to evaluate the effects of practice on performance and workload in a compensatory tracking task.

The present data show the same pattern of results found in the research mentioned above, regarding NASA-TLX and SWAT. Sensitivity of the WP has been evaluated on just one occasion (Tsang & Velazquez, 1996), in a laboratory experiment using a memory search task and a tracking task. In their study the WP was introduced and evaluated against two unidimensional instruments: Bedford and Psychophysical Scaling. Among the three subjective procedures, the psychophysical ratings were the most sensitive to task demand manipulations. The overall ratings of the WP were the least sensitive to task demand manipulations, but the multivariate effect size was more impressive. In contrast to the Tsang and Velazquez study, our research compared the WP with two well-established multidimensional techniques and, under that condition, WP sensitivity was higher than that obtained with SWAT and TLX.

In summary, the present results demonstrate that, among the three subjective procedures, WP was the one which bears the highest sensitivity. The other two instruments showed similar sensitivity, NASA-TLX being slightly more sensitive than SWAT according to the F value.

5. CONCLUSIONS

The analysis of mental workload in a certain job leads to a number of practical implications for the training plans, the selection process, and task designing and redesigning. If we are to process a variety of information at the same time, to make appropriate decisions, to efficiently solve emergency problems, and to adapt to technological changes we have to bear a substantial increase in cognitive complexity in many operations. Therefore, a major goal of work psychology is the analysis of task demands in order to design jobs that bring about a lower mental workload. This in turn will lead to lower stress levels and accident rates, and to a decrease in the likelihood of errors as well. Hence, the importance of mental workload evaluation.

This research attempted to analyse and compare the characteristics of three measures of subjective mental workload: NASA-TLX, SWAT, and WP. The study focused on the characteristics listed below:

Intrusiveness

No differences were found in performance due to the workload evaluation instrument. As questionnaires were administered following the task performance, it can be concluded that interference of these instruments with performance is almost negligible.

Sensitivity

Although the three mental workload instruments used yielded global workload indices sensitive to changes in the objective difficulty of tasks, the Workload Profile (WP) was the only one to reveal differences caused by both factors of task complexity and by their interaction. WP is then, among the three instruments, the one which bears the highest sensitivity. The other two instruments show similar sensitivity, NASA-TLX being slightly more sensitive according to the F test.

Convergent Validity

The study yielded positive correlation coefficients near to 1 between the three measures. Thus, there is a very high convergent validity between them.

Concurrent Validity

NASA-TLX shows a high correlation with performance. In contrast, the correlations with performance were lower for SWAT and WP.

Diagnosticity

This index assesses the extent to which mental workload indices discriminate between tasks. Both NASA-TLX and SWAT produced very similar task clusters, the latter having a much higher discriminant power. As for the Workload Profile, the first discriminant function discriminates between single memory tasks and single tracking tasks. The coordinates of the tracking task in the bi-dimensional space suggest that this task demands response processing resources and has a spatial processing code. However, the memory task demands perceptual/central processing resources and has a verbal processing code.

The second discriminant function discriminates between single and dual tasks. Bearing in mind that single tasks yield negative scores and dual tasks yield positive scores, and that structure matrix coefficients are positive and high across all dimensions of mental workload, it can be concluded that dual tasks demand processing resources of all kinds, whereas single tasks demand only some of them.

This shows the high diagnostic power of WP. Mental workload profiles yielded by this technique allowed the accurate detection of differences in the kind of attention resources demanded by each task, according to the multiple resources model.

Implementation Requirements and Acceptability

Although no computation of differences among the three instruments was made concerning implementation and acceptability, several advantages were noticed in the separate administration of each. As they are paper-and-pencil instruments, implementation requirements are the very least. The only difference to emphasise is the administration time. The SWAT took a little longer to administer (70 min.) than the other two (60 min.).

Subjects accepted willingly the three instruments although there were some problems concerning comprehension of the dimensions in the WP. As for the SWAT, the ranking task prior to the performance of the experimental tasks proved wearisome.

To sum up, some basic recommendations can be given concerning the evaluation of mental workload in applied settings, depending on the goals:

- If the goal is a comparison between the mental workload of two or more tasks with different objective levels of difficulty, then the assessor should choose the Workload Profile.
- If the goal is to predict the performance of a particular individual in a task, then NASA-TLX is recommended.
- If what is needed is an analysis of cognitive demands or attention resources demanded by a particular task, then the best choice would be WP or, as an alternative, SWAT.

Finally, we want to emphasise the need for continued research in subjective mental workload that serves to develop better human information processing models, to design new measure procedures, and to improve the properties of the existing assessment instruments. The present experiment represents only a small step on this journey. Many things might be done, but it would be particularly interesting to test the reliability of the three instruments considered and to replicate the experiment using different tasks. Moreover, we think that the Workload Profile instrument—like TLX and SWAT—can be applicable in complex real-world tasks, but to be able to establish definitive conclusions about this it would be necessary to perform a study specifically designed for that.

REFERENCES

Battiste, V., & Bortolussi, M. (1988). Transport pilot workload: A comparison of two subjective techniques. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 150–154). Santa Monica, CA: Human Factors Society.

Beare, A.N., & Dorris, R.E. (1984). The effects of supervisor experience and the presence of a shift technical advisor on the performance of two-man crews in a nuclear power plant simulator. In *Proceedings of the Human Factors Society*

Twenty-Eighth Annual Meeting (pp. 242–246). Santa Monica, CA: Human Factors Society.

Becker, A.B., Warm, J.S., Dember, W.N., & Hancock, P.A. (1991). Effects of feedback on perceived workload in vigilance performance. In *Proceedings of the Human Factors Society Thirty-Fifth Annual Meeting* (pp. 1491–1494). Santa Monica, CA: Human Factors and Ergonomics Society.

Bittner, A.V., Byers, J.C., Hill, S.G., Zaklad, A.L., & Christ, R.E. (1989). Generic workload ratings of a mobile air defence system (LOS-F-H). In *Proceedings of the Human Factors Society Thirty-Third Annual Meeting* (pp. 1476–1480). Santa Monica, CA: Human Factors Society.

Byers, J.C., Bittner, A.C., Hill, S.G., Zaklad, A.L., & Christ, R.E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 1145–1149). Santa Monica, CA: Human Factors Society.

Cooper, G.E., & Harper, R.P. (1969). *The use of pilot ratings in the evaluation of aircraft handling qualities* (NASA Ames Technical Report NASA TN-D-5153). Moffett Field, CA: NASA Ames Research Center.

Corwin, W.H. (1989). In-flight and post-flight assessment of pilot workload in commercial transport aircraft using SWAT. In *Proceedings of the Fifth Symposium on Aviation Psychology* (pp. 808–813). Columbus, OH: Department of Aviation, Ohio State University.

Corwin, W.H., Sandry-Garza, D.L., Biferno, M.H., Boucek, G.P., Logan, A.L., Jonsson, J.E., & Metalis, S.A. (1989). *Assessment of crew workload measurement methods, techniques, and procedures: Process, methods and results. Report WRDC-TR-89-7006*. Wright-Patterson Air Force Base, OH: Wright Research and Development Centre, Air Force Systems Command.

Eggemeier, F.T., Wilson, G.F., Kramer, A.F., & Damos, D.L. (1991). General considerations concerning workload assessment in multi-task environments. In D.L. Damos (Ed.), *Multiple task performance* (pp. 207–216). London: Taylor & Francis.

Gawron, V.J., Schiflett, S.G., Slater, T., Miller, J., & Ball, J. (1987). Concurrent validation of four workload and fatigue measures. In *Proceedings of the Fourth Symposium on Aviation Psychology* (pp. 609–615). Columbus, OH: Ohio State University.

Hancock, P.A., & Warm, J.S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*, 519–537.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: North-Holland.

Haworth, L., Bivens, C., & Shively, R. (1986). An investigation of single-piloted advanced cockpit and control configurations for nap-of-the-earth helicopter combat mission task. In *Proceedings of the 1986 Meeting of the American Helicopter Society* (pp. 657–672). Washington, DC: American Helicopter Society.

Hendy, K.C., Hamilton, K.M., & Landry, L.N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors*, *35*(4), 579–601.

Hill, S.G., Byers, J.C., Zaklad, A.L., & Christ, R.E. (1989). Subjective workload assessment during 48 continuous hours of LOS-F-H operations. In *Proceedings of the Human Factors Society Thirty-Third Annual Meeting* (pp. 1129–1133). Santa Monica, CA: Human Factors Society.

Hill, S.G., Byers, J.C., Zaklad, A.L., Christ, R.E., & Bittner, A.C. (1988). Workload assessment of a mobile air defences system. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 1068–1072). Santa Monica, CA: Human Factors Society.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklad, A.L., & Christ, R.E. (1992). Comparison of four subjective workload rating scales. *Human Factors*, *34*, 429–439.

Kilmer, K.J., Knapp, R., Bursal, C., Borresen, R., Bateman, R., & Malzahn, D. (1988). Techniques of subjective assessment: A comparison of the SWAT and the Modified Cooper-Harper scales. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 155–159). Santa Monica, CA: Human Factors Society.

Meshkati, N., Hancock, P., & Rahimi, M. (1992). Techniques in mental workload assessment. In J. Wilson & E. Corlett (Eds.), *Evaluation of human work. A practical ergonomics methodology* (pp. 605–627). London: Taylor & Francis.

Nataupsky, M., & Abbott, T.S. (1987). Comparison of workload measures on computer-generated primary flight displays. In *Proceedings of the Human Factors Society Thirty-First Annual Meeting* (pp. 548–552). Santa Monica, CA: Human Factors Society.

Nygren, T.E. (1991). Psychometric properties of subjective workload techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, *33*(1), 17–33.

Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P.A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185–218). Amsterdam: Elsevier.

Roscoe, A.H. (1987). *The practical assessment of pilot workload*, AGARD-AG-282. Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development.

Roscoe, A.H., & Ellis, G.A. (1990). *A subjective rating scale assessing pilot workload in flight. A decade of practical use*. Royal Aerospace Establishment, Technical Report 90019. Farnborough, UK: Royal Aerospace Establishment.

Sawin, D.A., & Scerbo, M.W. (1995). Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. *Human Factors*, *37*, 752–765.

Schick, F.V., & Hahn, R.L. (1987). The use of subjective workload assessment technique in a complex flight task. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload*, AGARD-AG-282 (pp. 37–41). Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development.

Shively, R., Battiste, V., Matsumoto, J., Pepiton, D., Bortolussi, M., & Hart, S.G. (1987). In flight evaluation of pilot workload measures for rotorcraft research. In *Proceedings of the Fourth Symposium on Aviation Psychology* (pp. 637–643). Columbus, OH: Department of Aviation, Ohio State University.

Skelly, J., & Purvis, B. (1985). B52 Wartime mission simulation: Scientific precision in workload assessment. Paper presented at the 1985 Air Force Conference on Technology in Training and Education.

Thiessen, M.S., Lay, J.E., & Stern, J.A. (1986). Neuropsychological workload test battery validation study. Report FZM 7446, for Harry G. Amstrong Aerospace Medical Research Laboratory. Fort Worth, TX: General Dynamics Corporation.

Tsang, P.S., & Johnson, W.W. (1989). Cognitive demand in automation. *Aviation, Space, and Environmental Medicine*, *60*, 130–135.

Tsang, P.S., & Velazquez, V.L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381.

Vidulich, M.A., & Bortolussi, M.R. (1988). A dissociation of objective and subjective workload measures in assessing the impact of speech controls in advanced helicopters. In *Proceedings of the Human Factors Society Thirty-Second Annual Meeting* (pp. 1471–1475). Santa Monica, CA: Human Factors Society.

Vidulich, M.A., & Tsang, P.S. (1985). Assessing subjective workload assessment: A comparison of SWAT and the NASA-bipolar methods. In *Proceedings of the Human Factors Society Twenty-Ninth Annual Meeting* (pp. 71–75). Santa Monica, CA: Human Factors Society.

Vidulich, M.A., & Tsang, P.S. (1986). Techniques of subjective workload assessment: A comparison of SWAT and the NASA-Bipolar methods. *Ergonomics*, *29*, 1385–1398.

Whitaker, L., Peters, L., & Garinther, G. (1989). Tank crew performance: Effects of speech intelligibility in target acquisition and subjective workload assessment. In *Proceedings of the Human Factors Society Thirty-Third Annual Meeting* (pp. 1411–1413). Santa Monica, CA: Human Factors Society.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & D.R. Davis (Eds.), *Varieties of attention* (pp. 63–102). Orlando, FL: Academic Publishers.

Wickens, C.D. (1987). Information processing, decision making, and cognition. In G. Salvendy (Ed.), *Cognitive engineering in the design of human–computer interaction and expert systems*. Amsterdam: Elsevier.

Wickens, C.D. (1992). *Engineering psychology and human performance*. New York: HarperCollins.