Diagnosticity and multidimensional subjective workload ratings

Article *in* Ergonomics · April 1996

Impact Factor: 1.56 · DOI: 10.1080/00140139608964470 · Source: PubMed

2 authors, including:

Pamela S Tsang
Wright State University
50 PUBLICATIONS **520** CITATIONS

ERGONOMICS, 1996, VOL. 39, NO. 3, 358–381

Diagnosticity and multidimensional subjective workload ratings

PAMELA S. TSANG

Department of Psychology, Wright State University, 3640 Col. Glenn HWY, Dayton OH 45435, USA

VELMA L. VELAZQUEZ

Usability Engineering, Intuit, PO Box 7850, Mountain View, CA 94039, USA

Keywords: Subjective mental workload; Diagnosticity; Multidimensional mental workload; Subjective assessment approaches.

A new multidimensional subjective workload assessment instrument—Workload Profile—was introduced and evaluated against two unidimensional instruments—Bedford and Psychophysical scaling. Subjects performed two laboratory tasks separately (single task) and simultaneously (dual task). The multidimensional procedure compared well with the unidimensional procedures in terms of sensitivity to task demands, concurrent validity with performance, and test-retest reliability. The results suggested that the subjective workload profiles would only have limited predictive value on performance. However, results of the canonical analysis demonstrated that the multidimensional ratings provided diagnostic information on the nature of task demands. Further, the diagnostic information was consistent with the a priori task characterization. This strongly supports the notion that mental workload is multidimensional and that subjects are capable of reporting the demands on separate workload dimensions. Theoretical implications on mental workload models and practical implications on the assessment approaches are discussed.

The present paper introduces and evaluates a multidimensional subjective workload assessment instrument—the Workload Profile. Implicit in all of the multidimensional workload procedures is the assumption that mental workload is multidimensional and multidimensional information is available to and accessible by the subjects. In contrast to a unidimensional approach, a multidimensional approach has the potential to offer diagnostic information concerning the nature of the task demand in addition to information on the intensity or levels of demand.

The benefits and necessity of developing theoretically based workload models have been advocated by many (Gopher and Donchin 1986, Gopher and Kimchi 1989, Kantowitz 1992). Several multidimensional instruments (such as the Subjective Workload Assessment Technique (SWAT), Reid and Nygren 1988, the NASA Task Load Index (TLX), Hart and Staveland 1988, the VACP model (visual, auditory, cognitive, psychomotor model), McCracken and Aldrich 1984) have been routinely used and found to provide useful information. However, most of these instruments do not have strong theoretical underpinnings. The workload dimensions in the Workload Profile procedure proposed here, in contrast, are based on a current theoretical model that has been subjected to much empirical testing. Theoretically based workload predictions can therefore be made a priori and tested. The diagnosticity of the Workload Profile dimensions can be judged against the theoretical predictions, independently of how they compared to other instruments.

1. Workload Profile

The Workload Profile procedure has in common with another multidimensional procedure—Workload Index (W/INDEX, North and Reiley 1988), the assumption that workload dimensions can be defined by the resource dimensions hypothesized in the multiple resource model of Wickens (1987). The various workload dimensions represent the different demands that can be imposed by a task: perceptual/central processing, response selection and execution, spatial processing, verbal processing, visual processing, auditory processing, manual output, and speech output. To the extent that these dimensions are a good representation of mental workload, the multidimensional ratings could provide a diagnostic workload profile that would describe the manner in which the task is demanding.

One major difference between W/INDEX and the Workload Profile procedure is that the latter asks the subjects directly to rate the *proportion of attentional resources* used for each task on several dimensions. Specifically, subjects are asked to provide difficulty ratings retrospectively (after they have performed all of the tasks) on the eight dimensions hypothesized in the Wickens' model (see also Wierwille and Eggemeier 1993). With W/INDEX, researchers knowledgeable about the tasks to be performed and familiar with the theoretical background of the resource dimensions (subject matter experts) would make a programmatic estimate a priori. The estimate is a workload index used to *predict* the eventual performance, especially time-shared performance (North and Reiley 1988, Sarno and Wickens 1991). The Workload Profile procedure examines the diagnosticity afforded by the multidimensional ratings provided directly by the subjects who have actually experienced the tasks and who are not necessarily familiar with the theoretical framework or predictions. In short, W/INDEX is a projective technique, whereas Workload Profile is an assessment technique.

As the Workload Profile procedure has not been previously examined, its performance is compared to two more established instruments. The comparison is made along three criteria: sensitivity to manipulation of task demands, concurrent validity with task performance, and test-retest reliability. A discussion on the importance of the various criteria for selecting workload assessment instruments can be found in Muckler and Seven (1992), O'Donnell and Eggemeier (1986), and Wierwille and Eggemeier (1993).

The natural choice of instrument for comparison would be another multidimensional instrument. Instead, two unidimensional instruments, the Bedford (Roscoe 1987) and Psychophysical (Gopher and Braune 1984) procedures, were chosen. There were two reasons for this. First, multidimensional ratings generally require more time to collect. As will be made clear below, time was severely limited given the scope of the present experiment. Second, rating on different dimensions for the different instruments in one setting could be unduly confusing to the subjects. Given that it was not practically feasible to compare the new instrument with other multidimensional instruments, comparing the new instrument with the unidimensional instruments would at least provide some indications of the sensitivity, reliability, and concurrent validity of the new instrument. Without these qualities, the diagnosticity information offered by the multiple dimensions would be of little value.

To summarize, the present paper introduces a multidimensional subjective workload assessment technique and examines the diagnosticity afforded by the multidimensional workload ratings. The task demand of the experimental tasks will first be characterized according to a multiple resource model and relevant empirical findings.

The actual subjective ratings are then gauged against the a priori characterization. Second, the new instrument is compared to two well-known subjective techniques to obtain some preliminary indications of its sensitivity, reliability and concurrent validity.

2. Method

2.1. *Subjects*

Sixteen right-handed college students (8 females and 8 males) between the ages of 18 and 28 years ($M = 21{\cdot}7$ years) served as subjects. Subjects were either remunerated at \$4·00 per hour or received research credits towards a course requirement.

2.2. *Tasks*

A discrete Sternberg memory task and a continuous tracking task were used. Each task could be performed alone (single task) or the two tasks could be performed simultaneously (dual task). The use of the continuous tracking task and the Sternberg memory task have been common in workload research (for example, see Gopher and Donchin 1986). These tasks have well-defined and documented difficulty manipulations, as well as face validity to real-world activities. The tracking task is commonly considered a generic laboratory analogue for a vehicular control task that requires constant monitoring and control in order to remain on course. The Sternberg task imposes a working memory demand that is typical of many in-flight operations. For example, a pilot will often have to maintain the air traffic controller's instructions (such as a new radio frequency) in memory amidst flight control activity. Generally, workload assessment techniques that have fared well in laboratory environments that used operational-relevant tasks have generalized well to the operational settings. For example, the Subjective Workload Dominance (SWORD) technique (Vidulich *et al.* 1991) was originally tested on a dual-tracking laboratory task (Vidulich and Tsang 1987) and has subsequently been successfully used in evaluating F-16 HUD designs (Vidulich *et al.* 1991) and the impact of crew-reduction in the KC-135 (Rueb *et al.* 1994).

2.2.1. *Sternberg task (SB):* Subjects learned a memory set of two (SB2) or four (SB4) letters at the beginning of each trial. Memory set size was manipulated as an objective difficulty parameter against which subjective workload measures could be compared. The letters were randomly chosen from the English consonants for each trial. During the trial, a single probe was presented at the centre of the screen where it remained until the subject responded. The Sternberg task had a fixed response-stimulus interval of 1·5 s. Subjects responded with the left hand by pressing the 'yes' button if the letter belonged to the memory set or the 'no' button otherwise. Subjects were asked to be as accurate and as fast as possible, but accuracy was emphasized over speed. The dependent variables were accuracy, omissions, and reaction time (RT, the median reaction time of the correct responses for the trial). Since there were not enough omissions for analysis and there was no evidence of speed-accuracy tradeoff, only the RT measure is discussed here.

2.2.2. *Tracking task (TR):* A one-dimensional continuous compensatory tracking task was used. The task was to maintain a moving cursor centred on a stationary reference line at the centre of the screen. Lateral random disturbance was generated with sum of sines. The leftmost possible cursor position and the rightmost possible

cursor position on the screen subtended a visual angle of 6° 50′. The control stick was a spring-centred, finger-controlled, joystick (Measurement System Inc., Model 531). The tracking task was performed with the right hand. Two orders of control were used: first order or velocity control (TR1) and second order or acceleration control (TR2). The order of control was manipulated as an objective difficulty parameter against which subjective measures could be compared. The dependent variable for the tracking task was Root Mean Square error (RMSE). (The joystick position was sampled 30 times per second. A RMSE was calculated over every 10 samples. A running average of the RMSEs was obtained over a 1-s sliding window. Finally, a mean RMSE of these running averages was calculated for the trial.)

2.2.3. *Dual tasks:* The four single tasks were factorially combined to produce four dual tasks: TR1SB2, TR1SB4, TR2SB2 and TR2SB4. The tracking task was presented directly below the Sternberg task. The two tasks were presented spatially close together to minimize visual scanning. The vertical visual angle subtended by the two tasks was 1° 10′; the horizontal visual angle subtended by the Sternberg stimulus and the furthest possible cursor position on one side was 3° 17′. Subjects were asked to pay equal attention to both tasks and to do their best on both tasks.

2.2.4. *A priori task characterization:* According to Wickens' multiple resource model and findings in the literature, the tracking task would be characterized as a spatial task and the memory task as a verbal task. The memory and tracking tasks had their input and output modalities in common (both were visually presented and responded to manually). The more difficult version of both tasks (TR2 and SB4) were expected to impose increased perceptual/central processing demand due to an increased demand on working memory in SB4 and the need to anticipate future position of the tracking cursor so as to generate leads for effective second order control (Isreal *et al.* 1980, Wickens *et al.* 1981). TR2 would also have increased response processing demand relative to TR1 owing to the need for more movement reversals for TR2 control. Jagacinski (1977) and Wickens (1986) provide an excellent overview of the parameters and demand of manual control. Among the dual tasks, resource competition or resource demand (and therefore dual-task performance decrement) was expected to be lowest for TR1SB2, intermediate for TR1SB4 and TR2SB2, and highest for TR2SB4 (see, for example, Tsang *et al.* 1995).

2.3. *Workload measures*

2.3.1. *Psychophysical scaling:* The Psychophysical scaling procedure (Gopher and Braune 1984) involves first assigning an arbitrary workload value to one task that is designated as the 'reference task'. In the present experiment, the first order tracking task was designated the reference task and assigned a workload value of 100. A somewhat easy task was used as a reference task so that the assigned value would not constrain the upper end of the ratings. The reference task was listed at the top of the page and all the other tasks were listed below in a random order. Subjects assigned one rating to each of the other tasks relative to the reference task, after having experienced all the task conditions. No restriction was placed on the range of possible ratings, hence each rating was transformed into a proportion of the range of ratings used by the subject (Gopher and Braune 1984). Once subjects were familiar with the procedure, it typically took only a few minutes to complete the ratings for all the tasks.

2.3.2. *Bedford scaling:* The Bedford scaling procedure has been used to evaluate aircraft systems in Britain (Roscoe and Ellis 1990) and in the United States (Corwin *et al.* 1989). Like the Psychophysical procedure, the Bedford procedure asks for a single unidimensional rating for each task condition. Immediately after performing the task to be rated, subjects went through a decision tree like that of a Cooper-Harper scale and rated the task by reporting the amount of spare attentional capacity not utilized by the task. Subjects had the Bedford decision tree available at the time of the rating, but few subjects consulted the tree after the first few trials. A rating of '1' corresponded to insignificant workload; a rating of '10' corresponded to extremely high workload with no spare capacity and unable to complete the task. It took only seconds to obtain a Bedford rating for each task.

2.3.3. *Workload Profile:* Conceptually similar to the Bedford scaling procedure, the Workload Profile asked the subjects to provide the *proportion of attentional resources* used after subjects had experienced all of the tasks to be rated. Figure 1 is a sample rating sheet. The tasks to be rated were listed in a random order down the column and the eight workload dimensions were listed across the page. Detailed explanation and examples for each of the workload dimensions were provided to the subjects (appendix A). Subjects had available with them the definition of each dimension at the time of the rating. In each cell on the rating sheet, subjects provided 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' meant that the task placed no demand on the dimension being rated; a rating of '1' meant that the task required maximum attention. Since there were no auditory or speech tasks, only six of the eight dimensions were analysed. The ratings on the individual dimensions were later summed for each task to provide an overall workload rating to be compared with the unidimensional Bedford and Psychophysical ratings. The ratings required 15 to 30 min to complete because the explanations were lengthy and because subjects had to rate multiple dimensions.

2.4. *Procedure*

All subjects performed the same tasks in six, 2·5 h sessions. The extent of practice was necessary for stabilizing performance and examining other psychological issues not relevant to the present paper. Although there were several blocks of task performance, subjective ratings were obtained only twice. The subjective ratings were collected once early in practice (Block 1 obtained in Session 1) to assess subjects'

	Workload Dimension							
	Stage of Processing		Code of Processing		Input		Output	
Tasks	Perceptual /Central	Response	Spatial	Verbal	Visual	Auditory	Manual	Speech
TR1								
TR1-SB4								
TR1-SB2								
SB2								
TR2-SB2								
TR2								
TR2-SB4								
SB4								

Figure 1. Workload Profile sample rating sheet.

initial reaction to the tasks and once late in practice (Block 2 obtained in Session 4) when they were 'skilled' at the tasks. Analyses were performed on the subjective ratings and performance obtained in these two blocks only.

There were three trials of each task condition in a row in Block 1 and two trials in a row in Block 2. Trial duration was 3 min. Subjects were informed of the task condition at the beginning of the trial. Subjects were asked to pay equal attention to both tasks in the dual-task conditions. Subjects were instructed to treat the two tasks in a dual-task condition as a unit and to report a Bedford rating that reflected the spare capacity not needed by either task. Bedford ratings were obtained at the end of the last trial of the same task condition, the Psychophysical and Workload Profile ratings were obtained at the end of the block.

3. Results

The predictive and diagnostic value of the ratings of the Workload Profile dimensions will first be examined. The overall Workload Profile ratings are then compared with the Psychophysical and Bedford ratings on their (a) sensitivity to objective task demands, (b) concurrent validity with task performances, and (c) test-retest reliability. Henceforth the overall subjective measures refer to the unidimensional Bedford and Psychophysical ratings and the sum of the ratings on the six Workload Profile dimensions. All *post-hoc* comparisons reported below used the Tukey Studendized Range statistic.

3.1. *Predicting performance with subjective workload ratings*

Multiple regression analyses were performed to obtain an estimate of the variance in performance accounted for by the subjective ratings on the six workload dimensions in the Workload Profile procedure. To the extent that the subjective ratings could account for the performance variance, the multiple regression equations could provide the appropriate weights (regression coefficients) for the various variables upon which future performance could be predicted. To assess the explanatory and predictive value of the subjective ratings, the amount of variance accounted for by the subjective ratings were, (a) tested to determine if they were significantly greater than zero, and (b) compared with that of objective demands. Several multiple regression models were examined. The top two models in Table 1 examined the relation between single-task performance and subjective ratings. The middle model examined the relation between the joint dual-task performance and dual-task subjective ratings. An estimate of the joint dual-task performance (ZSUM) was obtained by summing the standardized dual-task RT and RMSE. The joint performance allowed examination of the effects of task demand on both task performance regardless of the subject's allocation policy. In the bottom model, single-task ratings for each workload dimension were summed over the tracking and memory tasks and were used to predict the joint dual-task performance.

The full multiple regression models examined had a performance measure (i.e. RT, RMSE, or ZSUM) as the dependent variable; ratings on the six workload dimensions, the objective difficulty parameters (memory set size, tracking order), and block (practice) as independent variables. Subject was included as a classification variable because it was a repeated measure design. The first column of R^2 in table 1 shows that the objective difficulty parameters, practice, and the six subjective ratings accounted for a substantial amount of variance in performance in both single and dual-task conditions. Of note is that the sum of the single-task ratings

Table 1. Multiple regression analysis models and squared multiple correlations.

Multiple Regression model						R^2			
						Full model		Reduce model	
						With	Without	With	Without
Single ← RT	PC RP SP VB VS MN Single-task ratings	Memory Size			Block	·84****	·51****	·71****	·26****
Single ← RMSE	PC RP SP VB VS MN Single-task ratings		Track order		Block	·93****	·86****	·43****	·18***
Dual ← ZSUM	PC RP SP VB VS MN Dual-task ratings	Memory Size	Track order	M×T	Block	·81****	·79****	·56****	·15**
Dual ← ZSUM	PC RP SP VB VS MN Single memory + Single tracking ratings	Memory Size	Track order	M×T	Block	·82****	·79****	·30***	.10*

Notes. PC = perceptual/central, RP = response, SP = spatial, VB = verbal, VS = visual, MN = manual; full model included all independent variables listed, reduced model included only the subjective ratings for the six workload dimensions; with = with subject as a classification variable in the model, without = without the subject variable; M×T = Memory set size × tracking order interaction; ZSUM = sum of the standardized dual task RT and RMSE; **** $p < ·0001$ that R^2 is significantly different from zero, *** $p < ·001$, ** $p < ·01$, * $p < ·05$.

model (bottom model, $R^2 = ·82$) did just as well as the dual-task ratings model ($R^2 = ·81$).

The subject variable was statistically significant in the two single-task models ($p < ·0001$), suggesting considerable individual variability in single-task performance. Since the effect of a given individual is unknown and therefore could not be used for predicting performance, all the analyses were re-run without the subject variable. This was to determine which independent variable would make a significant contribution to the model without the influence of the subject variable. The second column of R^2 in table 1 shows that the subjective ratings, objective difficulty parameters, and practice together accounted for a substantial proportion of the variances in single- and dual-task performance, even without the subject variable.

The next logical step would be to examine the regression coefficients in order to determine the relative contribution of each variable. It would be useful to be able to distinguish for each task those components or specific workload dimensions that contributed significantly to changes in performance from those components that did not. However, the workload dimensions were found to be highly correlated among themselves. Table 2 displays the squared multiple correlation of each variable with the other independent variables in the full model without the subject variable. Multicollinearity among the workload dimensions was considerable and rendered the regression coefficients uninterpretable and not useful for prediction (Levine 1977, Pedhazur 1982). Multicollinearity among the workload dimensions however would not invalidate the R^2 or the proportion of variance in performance accounted for by the independent variables.

A second set of multiple regression analyses was therefore run with a reduced model. Only the ratings from the six workload dimensions served as the independent variables in the reduced models to assess the explanatory and predictive power of the

Table 2. Multicollinearity among the independent variables in the full multiple regression model without the subject variable.

Independent variable	R^2 (1 − Tolerance)		
	Single memory	Single tracking	Dual task
PC	·61	·49	·76
RP	·52	·69	·81
SP	·09	·78	.54
VB	·52	–†	·43
VS	·71	·72	·63
MN	·64	·76	·55
MEM	·16		.12
TRK		.19	·21
MEM×TRK			·00
BLK	.26	.37	·07

Notes R^2 = squared multiple correlation of each variable with the other independent variables in the model, PC = perceptual/central, RP = response, SP = spatial, VB = verbal, VS = visual, MN = manual, MEM = memory set size, TRK = tracking order, MEM×TRK = Memory set size× Tracking order interaction, BLK = block.

†VB was not correlated with the other variables because all subjects assigned a rating of zero for the tracking task.

subjective ratings by themselves. Comparing the first and third columns of R^2 in table 1, the reduced models clearly explained less performance variance than the full models that included the objective difficulty parameters, but the subjective ratings alone still accounted for a significant proportion of variance in performance. The proportion of variance accounted for was further reduced when the subject variable was not included in the model (last column). Statistically, the difference in R^2 between the reduced model with the subject variable and the reduced model without (between the third and fourth columns of R^2) was significant for all four models (single RT: ($F(15,106) = 11{\cdot}06$, $p < {\cdot}001$; single RMSE: $F(15,106) = 3{\cdot}05$, $p < {\cdot}001$; ZSUM with dual-task ratings: ($F(15,106) = 6{\cdot}54$, $p < 0{\cdot}001$; ZSUM with sum of single-task ratings: ($F(15,106) = 2{\cdot}01$, $p < {\cdot}05$). This indicated considerable individual variability in the subjective ratings.

The difference in R^2 between the full and reduced models (second and fourth columns in table 2) were all significant at ·001 level (single RT: ($F(2,119) = 30{\cdot}57$; single RMSE: ($F(2,119) = 287{\cdot}63$; ZSUM with dual-task ratings: ($F(4,117) = 86{\cdot}47$; ZSUM with sum of single-task ratings: ($F(4,117) = 97{\cdot}52$). Note that the difference between the full and reduced models could be attributed primarily to the objective difficulty parameters since the block variable accounted for only 5% of the variance in the single RT model and 2% or less in the other three models. Given that the objective difficulty parameters accounted for the predominant share of performance variance, the fourth model in table 1 shows that the single-task objective difficulty parameters could predict dual-task performance quite well. The sum of the single-task subjective ratings, on the other hand, did appreciably more poorly. Notwithstanding the difference between the full and reduced models, the six dimensions still accounted for a significant proportion of performance variance in all four models.

The main findings from the multiple regression analysis can be summarized as follows. First, subjective ratings on the six workload dimensions were highly

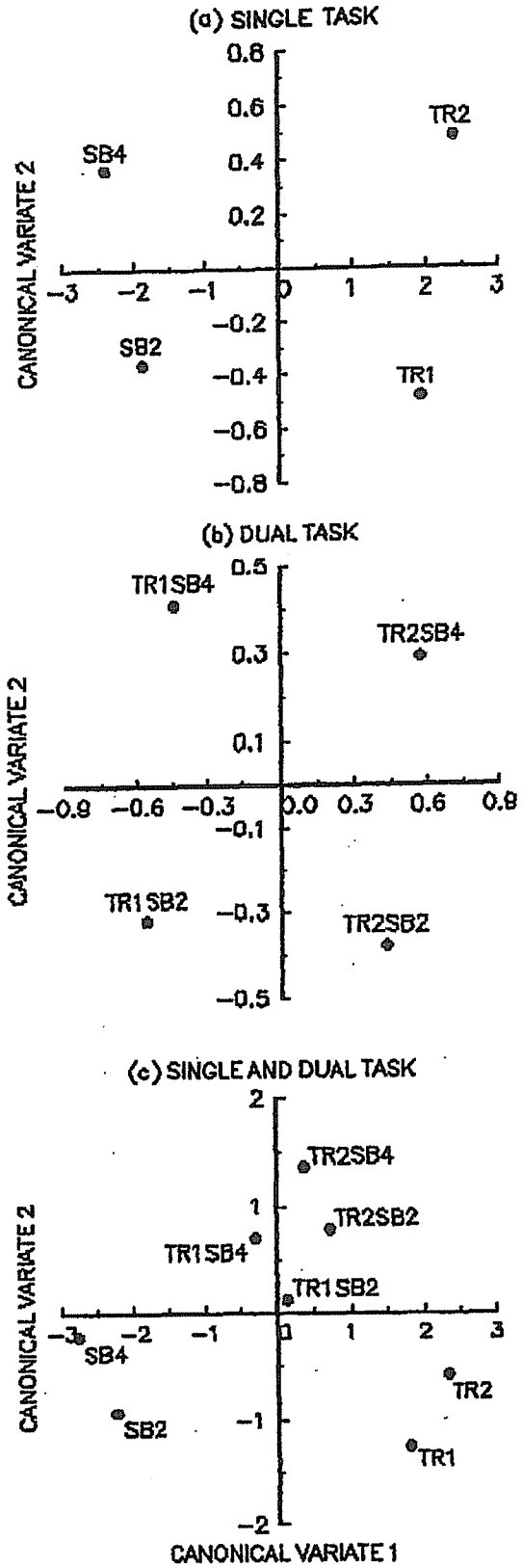

Figure 2. Discriminate function centroids of the different task conditions.

correlated with each other, rendering the regression coefficients uninterpretable. The ramifications of multicollinearity among the workload dimensions are further discussed below. Second, the objective difficulty parameters accounted for most of the variance in performance. Third, although the subjective ratings alone accounted

for a statistically significant proportion of variance in performance, subjective ratings would have limited predictive value because of the small proportion of variance that they accounted for.

3.2. *Diagnosticity of the Workload Profile*

A statistical approach that takes into account the intercorrelations among variables was used to examine the diagnosticity of the subjective workload ratings. The ratings on the six workload dimensions constituted a workload profile for each task condition (figures 3 and 4). Canonical discriminant analysis (SAS® procedure CANDISC) was used to examine the extent to which the workload profiles of the various task conditions could be distinguished. Second, canonical correlation analysis (SAS® procedure CANCORR) was used to examine the relationship between the dual-task workload profiles and the component dual-task performances (RT and RMSE).

As some readers may not be familiar with canonical analysis, a brief overview is provided here. Canonical analysis is used to study the relations between two sets of variables. The variables in each set are differentially weighted (expressed by standardized or structure coefficients) and a linear combination of the variables (known as the canonical variate) is formed for each set. The weights are determined in such a way that maximum possible correlation between the canonical variates of the two sets is obtained. Since there may be more than one pair of linear combinations that are highly correlated with each other, additional canonical variates are obtained. Each subsequent pair of combinations has a smaller canonical correlation than the preceding pair and is uncorrelated with all the preceding combinations. The

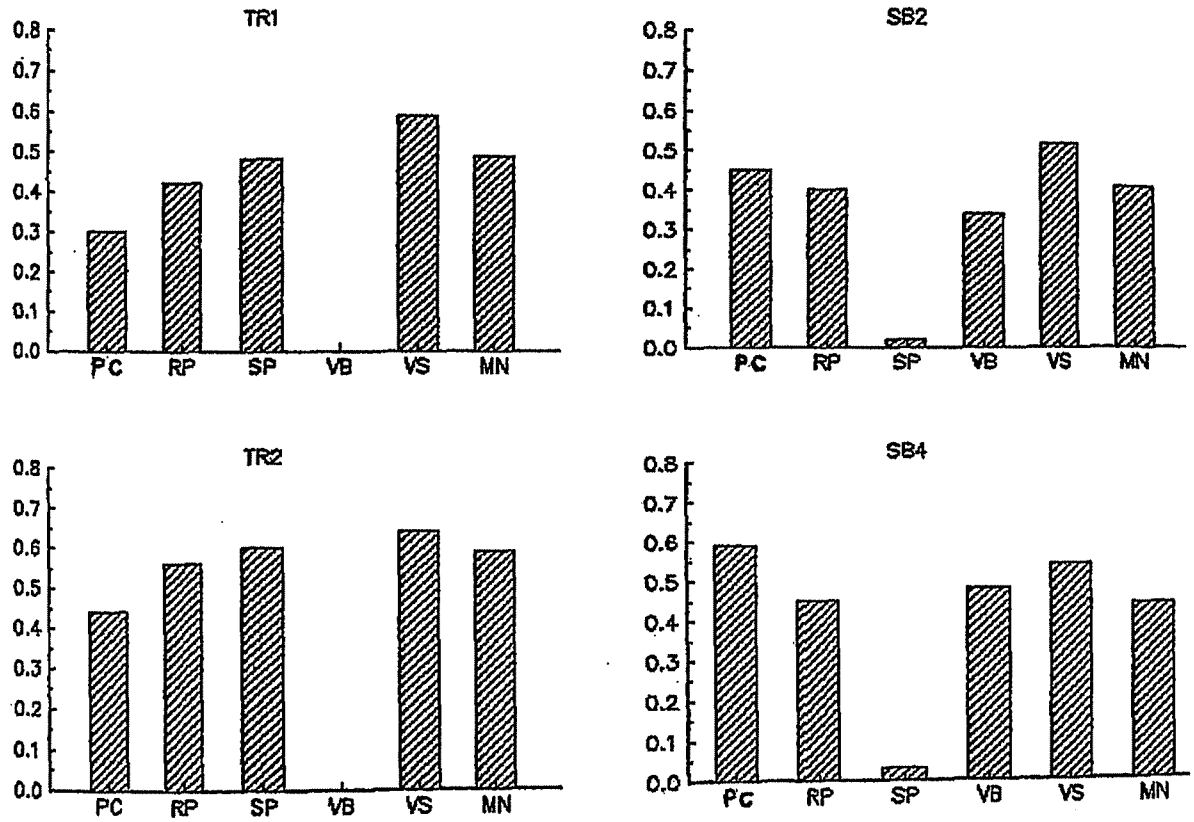

Figure 3. Single task mean subjective ratings on the six workload dimensions in the Workload Profile procedure. PC = perceptual/central, RP = response, SP = spatial, VB = Verbal, VS = visual, MN = manual.

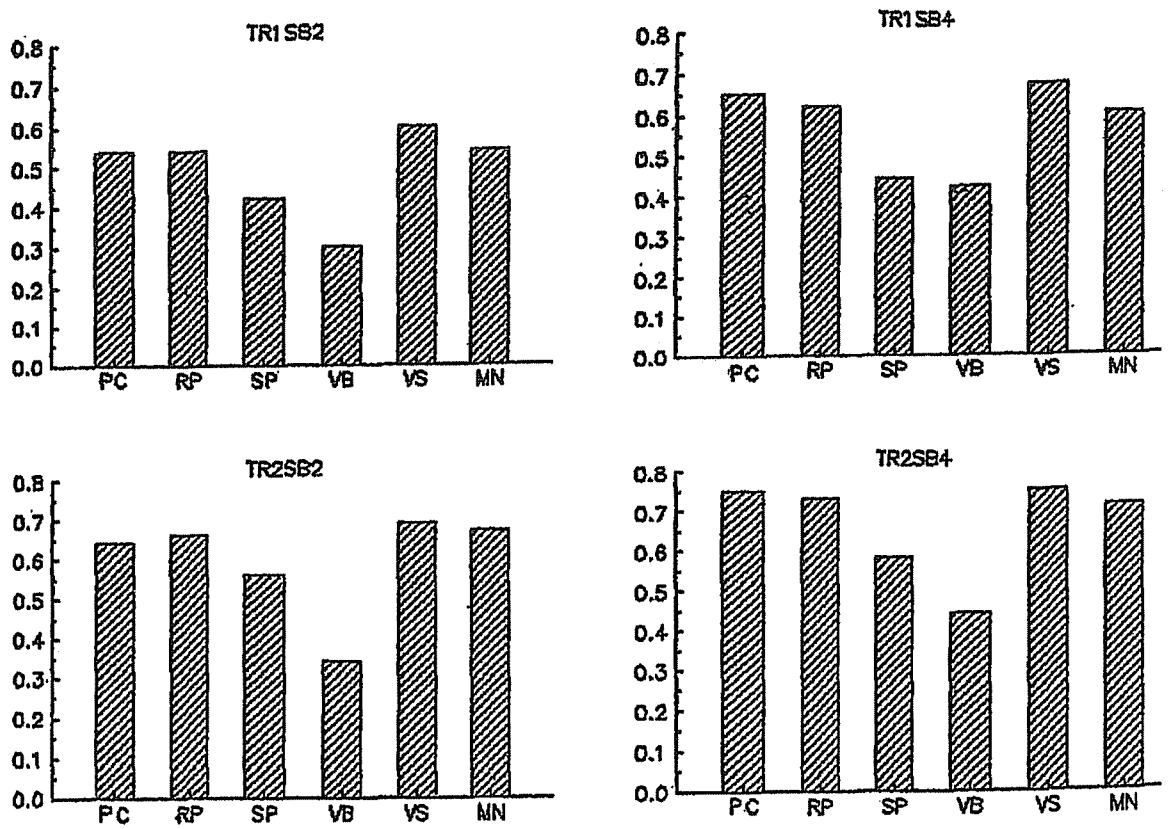

Figure 4. Dual task mean subjective ratings on the six workload dimensions in the Workload Profile procedure. PC = perceptual/central, RP = response, SP = spatial, VB = verbal, VS = visual, MN = manual.

maximum number of pairs of combination is constrained by the number of variables in the smaller set. The square of the canonical correlation (R_c^2) is an estimate of the variance shared by the two canonical variates.

In the present paper, the structure coefficients rather than the standardized coefficients from the canonical analysis are interpreted for the following reasons. Standardized coefficients are sample-specific and may not generalize across settings because they are related to the variances and covariances of the variables (Pedhazur 1982). Further, high multicollinearity among the variables (as is the case among the workload dimensions being studied) would result in broad confidence intervals around the standardized coefficients, and one variable may hide or suppress the importance of another variable correlated with the first (Levine 1977). Structure coefficients purportedly do not suffer from these two disadvantages. Structure coefficients are the correlations between the original variables and the canonical variates. They are obtained by multiplying the standardized coefficients by the correlation matrix (Pedhazur 1982). The structure coefficients are to be interpreted in the same manner as the loadings in factor analysis. The variables with the larger structure coefficients on a given canonical variate identify the dimensions on which they load. According to Pedhazur (1982), a rule of thumb is that only those structure coefficients $\geq \cdot 30$ and only those squared canonical correlations (R_c^2) $> \cdot 10$ are meaningful.

3.2.1. *Distinguishing among task conditions:* Three canonical discriminant analyses were performed to examine the diagnosticity of the subjective workload profiles in: (1) single tasks, (2) dual tasks, and (3) single and dual tasks in the same analysis. That is,

the set of ratings on the six workload dimensions were used to discriminate among the set of task conditions (a categorical variable). The squared canonical correlation and the structure coefficients for the significant canonical variates are reported in Table 3.

In the single task analysis (top proportion of table 3), subjective ratings on the six

Table 3. Canonical correlations between task conditions and workload profiles.

		Structure coefficients	
		Canonical Variate 1	Canonical Variate 2
Single-task workload dimensions	PC	−0·36	0·81
	RP	0·16	0·54
	SP	0·93	0·25
	VB	−0·88	0·27
	VS	0·15	0·17
	MN	0·23	0·34
Redundancy		·26	·03
Task means	TR1	1·92	−0·48
	TR2	2·37	0·48
	SB2	−1·89	−0·36
	SB4	−2·40	0·37
R_c^2		0.83	0·16
Dual-task workload dimension	PC	0·64	0·68
	RP	0·70	0·45
	SP	0·84	0·05
	VB	0·25	0·82
	VS	0·49	0·38
	MN	0·63	0·21
Redundancy		·08	·03
Task means	TR1SB2	−0·56	−0·32
	TR1SB4	−0·44	0·41
	TR2SB2	0·43	−0·38
	TR2SB4	0·57	0·29
R_c^2		0·21	0·13
Single- and dual-task workload dimension	PC	−0·17	0·84
	RP	0·19	0·66
	SP	0·84	0·50
	VB	−0·64	0·71
	VS	0.16	0.37
	MN	0·25	0·48
Redundancy		·16	·16
Task means	TR1	1·80	−1·25
	TR2	2·32	−0·58
	SB2	−2·26	−0·93
	SB4	−2·78	−0·22
	TR1SB2	0·13	0·12
	TR1SB4	−0·29	0·70
	TR2SB2	0·71	0·78
	TR2SB4	0·37	1·36
R_c^2		0·74	0·43

Notes. PC = perceptual/central, RP = response, SP = spatial, VB = verbal, VS = visual, MN = manual, TR1 = first order tracking, TR2 = second order tracking, SB2 = memory set size 2, SB4 = memory set size 4.

Table 4. Canonical correlations between dual-task performance and workload profiles.

		Structure coefficients of Canonical Variate 1			
		Dual task	R^2	Single tracking + single memory	R^2
Workload dimension	PC	0·71		−0·13	
	RP	0·67		−0·17	
	SP	0·54		−0·58	
	VB	0·25		0·26	
	VS	0·27		−0·21	
	MN	0·58		−0·01	
Redundancy		·15		·07	
Dual-task performance	RT	0.64	·10	0·82	·14
	RMSE	0·95	·21	−0·24	·01
R_c^2		0·23		0·20	

Notes. R_c^2 = squared canonical correlation, R^2 = squared multiple correlations between a canonical variate and performance (RT or RMSE); PC = perceptual/central, RP = response, SP = spatial, VB = verbal, VS = visual, MN = manual.

workload dimensions were used to discriminate among the four single tasks (TR1,TR2, SB2, SB4). The Wilks' Lambda was significant at ·0001 level (approx. $F(18,337) = 18·62$), showing that the workload profiles for the four single tasks were different. Two canonical variates were significant (canonical variate 1: approx. $F(18,337) = 18·62$, $p < ·0001$, $R_c^2 = ·83$; canonical variate 2: approx. $F(10,240) = 2·44$, $p < 0·01$, $R_c^2 = ·16$). The first canonical variate (first column, table 3) showed that the workload profiles for the tracking tasks (positive task means) and the memory tasks (negative task means) could be distinguished. The structure coefficients showed that the tracking tasks had high spatial and low verbal difficulty ratings, whereas the memory tasks had high verbal and low spatial ratings.

The second canonical variate (second column) showed that the workload profiles could also be distinguished by the objective task difficulty. According to the structure coefficients, the perceptual/central processing demand (·81) predominantly accounted for the difficulty of TR2 and SB4 (positive task means). High ratings in the response processing dimension (·54) was also associated with the more difficult tasks. The task means for each of the canonical variates are plotted at the top of figure 2. The task mean for each task was an average (over subjects) of the product of the coefficients in each canonical variate and the ratings for the corresponding dimension.

The first canonical variate accounted for 26% of the variance among the four single-task conditions (redundancy coefficient = ·26), the second variate accounted for 3% (redundancy coefficient = 0·3). Put differently, 26% of the total variance among the single-task conditions was explained by the first linear combination of the ratings on the six subjective workload dimensions.

Results of the canonical analysis on the four dual tasks are presented in the middle portion of table 3. The Wilks' Lambda was significant at ·001 level (approx. $F(18, 37) = 2·50$), showing that the four dual task workload profiles were different. The only significant canonical variate (approx. $F(18, 337) = 2·50, p < ·001$; $R_c^2 = ·21$) appeared to be primarily distinguishing between the dual second-order tracking tasks (positive task means) from those with first-order tracking (negative task means), irrespective of memory set size. The distinction laid heavily on the spatial dimension

that had the largest structure coefficient (·84). The structural coefficients further showed that the response (·70), perceptual/central (·64), and manual (·63) dimensions all had fairly large loadings and served to distinguish between first- and second-order tracking.

The second canonical variate was not significant ($p > \cdot 1$), but was displayed in table 3 anyway because it suggested a memory set size difference irrespective of tracking order. The primary dimensions distinguishing between the dual tasks with memory set size 4 (positive task means) and those with memory set size 2 (negative task means) were the verbal and perceptual/central dimensions. The first canonical variate accounted for 8% of the variance among the four dual-task conditions (redundancy coefficient = ·08). The second canonical variate accounted for 3%. The task means for each of the canonical variates are plotted in the middle of figure 2.

The third analysis examined the single- and dual-task workload profiles in the same analysis (bottom portion of table 4 and figure 2). Two canonical variates were significant (canonical variate 1: approx $F(42, 1143) = 14{\cdot}13$, $p < \cdot 0001$, $R_c^2 = \cdot 74$; canonical variate 2: approx. $F(30, 978) = 5{\cdot}37$, $p < \cdot 0001$, $R_c^2 = \cdot 43$). The first canonical variate (horizontal axis at the bottom of figure 2) primarily distinguished the workload profiles between the tracking and memory tasks. The single second-order tracking task had the highest positive task mean and the single-task memory set size 4 had the highest negative task mean. The dual tasks had intermediate task means, perhaps because each dual task had a tracking as well as a memory component. The high positive loading in the spatial dimension and the fairly high negative loading on the verbal dimension showed that the tracking tasks (positive task means) had high spatial and low verbal difficulty ratings whereas the memory tasks (negative task means) had high verbal and low spatial ratings.

The second canonical variate (vertical axis at the bottom of figure 2) distinguished between single (negative task means) and dual (positive task means) tasks. According to the structure coefficients, the perceptual/central dimension (·84) played a particularly important role in differentiating the single from dual-task workload profiles. This is totally consistent with the a priori characterization that both the memory task and tracking (especially the second order task) place some demand on the perceptual/central dimension. Adding a tracking or memory task would logically add to the perceptual/central demand. The verbal demand and response demand also appeared to increase substantially from single to dual tasks. The first and second canonical variates each accounted for 16% of the variance among the eight single and dual task conditions.

In summary, the canonical discriminant analyses revealed that the subjective workload profiles for the different task conditions were distinguishable in ways that were meaningful and consistent with expectations derived from the multiple resource model. The workload profiles revealed that: (a) the tracking demand was primarily spatial in nature and the memory task was primarily verbal in nature, (b) the difficulty manipulation effectively increased the perceptual/central and response processing demands, (c) second order and first order tracking demands could be distinguished along the spatial, response, perceptual/central, and manual dimensions, and (d) performing an additional task effectively increased the perceptual/central, verbal, and response processing demand with the present experimental tasks. More important, these findings suggested that, when asked explicitly for the information, subjects were able to provide meaningful information about the nature of the demand of the various task conditions.

3.2.2. *Distinguishing between performances:* Two canonical correlation analyses were performed on the dual-task data to examine the relationship between the workload dimensions and the two performance measures. In the first analysis, the two dual-task performance measures formed a set of variables and the dual-task subjective ratings on the six workload dimensions (workload profiles) served as the other set of variables. The linear combination of the two performance measures is referred to as the performance variate and the linear combination of the six workload ratings is referred to as the workload variate. The canonical analysis was performed to determine if the various workload dimensions differentially explained RT and RMSE. The second analysis was the same as the first except that the sum of the single task workload profiles (i.e. ratings on each workload dimension for single tracking and memory tasks were added together to generate the workload profile) was used to distinguish dual RT and dual RMSE. Table 4 presents the principal statistics from these two analyses.

From the analysis with the dual-task workload profile, two canonical variates were significant (canonical variate 1: $F(12, 240) = 4{\cdot}01$, $p < {\cdot}0001$, $R_c^2 = {\cdot}23$; canonical variate 2: $F(5, 121) = 2{\cdot}51$, $p < {\cdot}05$, $R_c^2 = {\cdot}09$). However, since the squared canonical correlation for the second variate was not even ·10, only the first variate is interpreted here. The structure coefficients (first column in table 4) showed that tracking RMSE (structure coefficient = ·95) was highly correlated with the first performance variate. In the workload variate, the perceptual/central dimension had the highest loading, followed by the response, manual, and spatial dimensions. About 15% of the performance variance was explained by the first workload variate (redundancy coefficient = ·15). The squared multiple correlations (R^2) between the performance variables and the workload variate indicated that the first workload variate was not particularly useful in predicting RT ($R^2 = {\cdot}10$), but slightly better in predicting RMSE ($R^2 = {\cdot}21$).

From the analysis with the summation of the two single-task workload profiles, only the first canonical variate was significant ($F(12, 240) = 3{\cdot}22, p < {\cdot}001, R_c^2 = {\cdot}20$). RT was highly correlated with the performance variate (structure coefficient = ·82). The corresponding workload variate had a small loading on the verbal dimension (·26). RMSE was only slightly correlated with the performance variate and the corresponding workload variate had a moderate loading on the spatial dimension. The first workload variate explained about 7% of the variance in performance (redundancy coefficient = ·07). The squared multiple correlations indicated that the linear combination of the sum of the single memory and single tracking ratings could not predict RMSE ($R^2 = {\cdot}01$) and could predict RT only poorly ($R^2 = {\cdot}14$).

The last two analyses showed that changes in tracking RMSE were associated with changes in the subjective difficulty of perceptual/central, response, spatial, and manual demands. How the workload demand changed with the RT changes was less clear. Second, the dual-task workload profiles appeared to have slightly higher predictive value than the sum of the single-task profiles. This suggested that the subjects' report of their mental integration of the difficulty of two tasks were more accurate than the estimates derived from single-task ratings. So far, results from the canonical analysis suggested that the workload profiles had rather impressive diagnostic value and the multiple regression analysis suggested that the workload profiles had rather low predictive value. The remaining analyses compared the workload profile ratings with the Psychophysical and Bedford ratings.

3.3. *Sensitivity to task demands*

A Block × Task ANOVA (analysis of variance) was performed on each of the performance and overall subjective measures. All eight tasks (four single and four dual tasks) were included in the overall subjective ratings ANOVAs. Only six tasks were included in the ANOVAs for the performance measures because there was no tracking performance measure for the two single memory tasks and no memory performance measure for the two single tracking tasks. The block main effect was reliable (at least $p < \cdot 05$) for all measures showing a practice effect in performance that is reflected also in the subjective difficulty of the tasks. Invariably, the mean RMSE, RT, and subjective workload ratings decreased from Block 1 to Block 2.

The task main effect was statistically significant ($p < \cdot 001$) for all of the subjective and performance measures, demonstrating that they were all sensitive to changes in task demands. *Post-hoc* comparisons on the performance measures showed a clear difference between the two levels of difficulty in both the tracking and memory tasks. Second order dual-tracking tasks had significantly ($p < \cdot 01$) higher error than first order dual-tracking and single-tracking tasks. Single task RT from SB2 was faster ($p < \cdot 05$) and dual task RT from TR2SB4 was slower ($p < \cdot 01$) than any other task conditions. Dual-task RT was significantly slower than single-task RT with the same memory set size ($p < \cdot 05$). As predicted by the multiple resource model, TR2SB4 had the highest tracking error and slowest RT.

In addition, one Block × Task MANOVA (multivariate analysis of variance) was performed on the performance measures (RT and RMSE) and one MANOVA was performed on the six Workload Profile ratings obtained from the four dual tasks. The task main effect for both performance ($F(6, 88) = 69 \cdot 08$) and ratings ($F(18, 113 \cdot 62) = 3 \cdot 70$) was significant at ·0001 level. The multivariate effect size

Table 5. Comparison of the workload assessment instruments.

	Performance		Subjective		
	RMSE	RT	Bedford	Psychophysical	Workload profile
Subjective approaches					
Dimensionality			Unidimensional	Unidimensional	Multidimensional
Absolute versus relative			Absolute	Relative	Absolute
Immediate versus retrospective			Immediate	Retrospective	Retrospective
Time to complete			Several seconds	Few minutes	15–30 minutes
Sensitivity					
Univariate effect size	75	·30	·37	·57	·28
Multivariate effect size		·97			·73
Concurrent validity					
r with RMSE			·50*	·66**	·57*
r with RT			·65**	·62**	·66**
Reliability					
r, tracking tasks	·98**		·81**	·93**	·94**
r, memory tasks		·83**	·86**	·95**	·92**

* = correlation significantly different from zero at $p < \cdot 05$, ** = $p < \cdot 01$.

(1−Wilk's Lambda, Huberty (1972)) was ·97 for the performance measures and ·73 for the workload ratings.

Table 5 lists the effect size or magnitude of treatment effects (Dodd and Schultz 1973) for each of the measures. Of the two performance measures, tracking error had larger effect size than RT. However, this could be due to a difference in the magnitude of memory size manipulation and tracking order manipulation. The effect size of the three overall subjective measures rivalled that of one of the performance measures—RT. Among the overall subjective measures, the Psychophysical ratings had the largest (·57), and the overall Workload Profile had the smallest (·28), effect size. However, the multivariate effect size for the Workload Profile ratings was an impressive ·73. In sum, while all the performance and subjective measures were sensitive to task demands, some had larger effect size than others.

3.4. *Concurrent validity with task performance*

Pearson correlations between each performance and each subjective measure across the six tasks (two single and four dual) were calculated for each subject. The Fisher z-transforms (z_r) of these correlations (Silver and Hollingsworth 1989) were then averaged over subjects and backtransformed into a mean correlation. These correlations were tested to determine if they were significantly different from zero. Concurrent validity was demonstrated by significant correlations between each of the performance measures and the three overall subjective measures (table 5).

To compare the three subjective instruments, the z_r of the correlations between performance and the overall ratings were subjected to one-way ANOVAs. Correlations between RMSE and the three instruments were significantly different ($F(2,30) = 3{\cdot}73$, $p < {\cdot}05$). *Post-hoc* comparisons showed that RMSE had significantly higher correlation with the Psychophysical ratings than with the Bedford ratings ($p < {\cdot}01$). Correlations with RT were not different among the three overall ratings. In short, each of the overall ratings, and especially the Psychophysical ratings, changed systematically with an objective measure of task demands (performance).

3.5. *Test-retest reliability*

Test-retest correlations between Blocks 1 and 2 were obtained for RMSE and overall subjective ratings from the two single tracking and four dual tasks. Test-retest correlations for RT and overall subjective ratings were obtained from the two single memory and four dual tasks. The Fisher z_r transforms of the test-retest correlations were averaged over subjects and backtransformed to a mean correlation. These mean correlations were tested for significance. The test-retest correlations were significantly different from zero for both performance measures and all three overall workload ratings (table 5).

The magnitudes of the z_r of the test-retest correlations for the various measures were compared with ANOVAs. One ANOVA compared the z_r of the RMSE and the three overall ratings from the tracking tests. *Post-hoc* comparisons showed that RMSE (·98) had significantly higher test-retest correlations than the Bedford ratings (·81, $p < {\cdot}01$), but the three instruments were not different from each other. Another ANOVA compared the z_r of the RT and the three overall subjective ratings from the memory tasks. Test-retest correlation for RT was not different from the three instruments.

To summarize, both performance measures and all three overall subjective ratings

appeared to be reliable measures. Tracking RMSE had the highest test-retest correlation that was significantly higher than that of the Bedford ratings. The test-retest reliability of RT, however, was not different from that of the three overall subjective ratings.

4. Discussion

A new multidimensional subjective workload assessment instrument was introduced and evaluated against two common unidimensional subjective instruments. Table 5 summarizes the comparison in terms of their sensitivity to changes in task demand, concurrent validity with performance, and test-retest reliability. Theoretical and practical implications of the comparison are discussed below.

4.1. *Objective versus subjective measures*

That tracking error had large effect size and high test-retest reliability has also been observed in several other studies (e.g. Tsang and Vidulich 1994). In contrast, although the RT measure also proved to be a sensitive and reliable measure, it did no better than the overall subjective ratings obtained in the present study. In addition, memory set size—an objective difficulty parameter that had demonstrative effect on performance—did not account for any greater amount of variance in performance than the Workload Profile ratings. Examining the RT models in table 1, the amount of variance accounted for by memory set size could be estimated by the difference in R^2 between the second and the fourth column. Notice that this difference is quite comparable to the R^2 of the subjective ratings alone (fourth column). These observations emphasise that subjective measures are not necessarily inferior to objective measures. There is little support for making a categorical distinction in utility between objective and subjective measures (see also Muckler and Seven 1992).

4.2. *Dimensionality*

Compared with the two unidimensional Bedford and Psychophysical procedures, the multidimensional Workload Profile procedure exhibited similar concurrent validity and test-retest reliability. Although the overall Workload Profile rating (sum of the ratings on the individual dimensions) also proved to be sensitive to task demands, its effect size was considerably smaller than that of the two unidimensional procedures (table 5). On the other hand, the multivariate effect size for the six Workload Profile ratings demonstrated superior sensitivity.

Multiple resource models in fact would predict that a simple sum of the workload dimensions would not be an optimum way of combining multidimensional ratings in generating a composite overall rating. The weighted composites derived from multiple regression analysis and canonical analysis were also not impressive. This was evidenced by the rather small proportion of the variance in performance accounted for by the weighted composites. The small but significant proportion of variance accounted for in turn showed that although informative, the weighted composites only had limited predictive value for performance. These results suggest that the relationship between performance and subjective difficulty is not a simple linear one. It is not clear from the present results how the different dimensions should be combined to provide an overall workload index.

Second, further analysis revealed a significant contribution unique to the multidimensional approach. The canonical analysis demonstrated that subjects could

meaningfully discriminate among the workload dimensions and provided ratings for the different dimensions according to the a priori task characterization. This strongly supports the notion that mental workload is multidimensional and that subjects are capable of reporting the demands on different workload dimensions. In fact, results of the last two canonical correlation analyses on the dual-task data suggested that subjects could also integrate the demands from two tasks, and provide a joint rating that had more predictive value than the summation of the single task estimates.

To elaborate, while the performance measures indicated differential demand levels between the difficulty levels, and between single and dual tasks, the canonical analysis on the Workload Profile ratings revealed similar distinction. In addition to the levels of demand, the canonical analysis revealed distinctive workload profiles for the differential task conditions. The present results suggest that the primary dimensions by which the distinction was made could be extracted with quantitative analysis. Knowing more precisely the way in which a task is difficult, as opposed to having only the overall difficulty information, would be invaluable for task and training designs. For example, if the task is difficult primarily because of high demand on visual processing, then the display may need augmentation. If the task is difficult primarily because of high demand on response processing, then the procedure or the complexity of the response should be simplified. One potential use of the subjective workload profile is to obtain more precise information concerning the way in which a system can be improved.

Third, the secondary task technique is traditionally the workload assessment technique that one would use to obtain diagnostic information (Wierwille and Eggemeier 1993). The present study demonstrates that subjective rating is a viable and less intrusive alternative to the secondary task technique.

4.3. *Implications for the multiple resource model*

4.3.1. *Exigency of the workload dimensions:* The workload dimensions, as defined by the resource dimensions in the multiple resource model, appeared to be intelligible to the subjects and evidently elicited meaningful ratings. Although the various task conditions were characterized a priori according to the multiple resource model, the characterization was mainly a qualitative one and only in general relative terms. For example, the memory set size 4 task was characterized to place a heavier demand on the perceptual/central dimension than the memory set size 2 task. The subjective ratings turned out not only to be consistent with the qualitative a priori characterization, but also supplemented the characterization with quantitative estimates.

The most distinguishing dimension between the tracking and memory tasks in the present data was the spatial and verbal dimensions or the codes of processing dimension (Wickens 1987). A number of performance studies corroborated the role that the codes of processing play in predicting performance (Brooks 1968, Wickens and Liu 1988). However, Sarno and Wickens (1991), in their evaluation of the role of multiple resources in predicting time sharing performance, failed to support a distinction between spatial and verbal processing. In fact, removing the spatial/verbal distinction from the W/INDEX model improved the correlation between the model prediction and tracking error decrements. That the tasks used did not place continuous demand on the subject was one possible reason proposed for the lack of distinction between the codes of processing. The present results, partly based on a continuous tracking task, strongly suggest that the spatial/verbal dimensions are distinct workload dimensions.

Clearly, the present experiment did not test all dimensions proposed in the multiple resource model. Future experiments will be needed to systematically examine all the dimensions. In general, the present results are supportive of the multiple resource model, but also offer additional insights that are addressed below.

4.3.2. *Multicollinearity among workload dimensions:* Statistically, multicollinearity among the independent variables in multiple regression analysis suggests that the variables are not independent of each other. There could be several interpretations of the multicollinearity in the present context. First, subjects could not evaluate the various workload dimensions independently. Second, the workload dimensions were not independent of each other. Results of the canonical analysis did not support either one of these interpretations however. Subjects appeared to provide subjective ratings discriminately for the different dimensions in a manner that was consistent with expectations. While the current study was not designed to test the independence of the resource dimensions in the Wickens model, the canonical results suggested that the various dimensions were distinguishable in meaningful ways.

Third, a plausible explanation for the high multicollinearity among the workload dimensions is that it may be inevitable that manipulation of any difficulty parameter would affect more than one specific workload or processing resource dimension. For example, the canonical analysis suggested that increasing the order of tracking control increased the demand on: perceptual/central, response, spatial, and manual processing. Vidulich and Bortolussi (1988) also observed high correlation among the six workload dimensions in the NASA-TLX scale. Ratings on all six NASA-TLX dimensions were found to rise and fall in the same direction across different phases of flights.

Fourth, given that the six dimensions from the Workload Profile and NASA-TLX are quite different, the correlations among the workload dimensions from the two instruments could be incidental, or it may have a common cause. Possibly, in addition to affecting specific workload dimensions or specific processing resources, the difficulty parameters also affected some general undifferentiated resource that is in charge of the executive and housekeeping activities. A demand change in this general resource could be reflected in all the dimensions.

4.4. *Implications for future research*

A possible, if only partial, explanation for the small proportion of variance in performance accounted for by either the subjective ratings or the memory set size was the large variability attributed to the subjects (compare the first and second columns or the third and fourth columns of R^2 in table 1). Future research will need to examine this possibility by having a larger number of observations. The persistence of large subject variability would be troublesome as it would seriously diminish predictability of performance.

Although the Bedford and Psychophysical procedures have now been used in a number of studies and settings, the Workload Profile procedure is new and is therefore in particular need of replication. In particular, the Workload Profile procedure needs to be applied to a variety of tasks to better evaluate subjects' ability to provide meaningful diagnostic information concerning the nature of the task demand. Further, it would be instructive to compare the Workload Profile to other more established multidimensional techniques. A comparison with the projective W/INDEX would be particularly interesting.

Last and most important is the need for a better understanding of human information processing and continual development of better human performance models upon which subjective or objective workload metrics can be based. As indicated in the present findings and as Gopher and Kimchi (1989) have pointed out, it is not yet clear how the different manifestations of workload relate to one another, or how the different influences of workload interact. While subjects appeared to be able to provide useful information, it is not clear what information should be elicited and how the various information should be combined or utilized for predicting performance.

4.5. *Practical applications*

The present data demonstrate that the different approaches to assessing subjective workload each have something different to offer. Among the three subjective procedures, the Psychophysical ratings were the most sensitive to task demand manipulations and had the highest concurrent validity with tracking error. The overall ratings of the Workload Profile were the least sensitive to task demand manipulations; but the multivariate effect size was more impressive. The workload profiles also offered meaningful diagnostic information that unidimensional procedures could not. Finally, although the Bedford procedure was not outstanding in any of the criteria examined, the Bedford ratings did not do poorly. As others have shown (Roscoe and Ellis 1990), the Bedford procedure produced sensitive and reliable ratings, but its most attractive feature is probably its ease of use. Subjects found it intuitively easy to understand and required little time to provide the ratings (table 5). The Workload Profile took the longest to collect and required more effort from the subjects to generate the ratings.

Further, the Bedford procedure is easy to administer and can be applied to a variety of tasks without alteration. With the Psychophysical approach, the selection of the reference task would require careful consideration of all the tasks to be rated (Gopher and Braune 1984). In the present context, there was only a choice of two tasks and the tracking task was arbitrarily chosen. With a larger heterogeneous set of tasks, the choice of a reference task may require more deliberation (Gopher *et al.* 1985).

Since there are already a number of univariate instruments with demonstrated effectiveness (e.g. the Bedford scale and the Psychophysical scale), the attractiveness of the multidimensional approach lies in its potential in affording diagnostic information. However, the determination of the relative merit of a unidimensional and multidimensional rating cannot be done independently of the merits of the dimensions that are being considered. If the workload dimensions to be rated do not reflect those of the true workload construct, then ratings on these dimensions would simply be uninformative, if not misleading. It is therefore imperative that the construction of the theoretically based workload dimensions can be subjected to empirical testing.

5. Conclusions

The main conclusion is that subjective ratings can be valid workload metrics. It is however important to recognize that different approaches used in the assessment process can influence the ratings. For example, the overall Workload Profile ratings exhibited lower sensitivity to task demands than the unidimensional procedures, but the workload profiles provided the diagnostic information on the nature of task

demands for the different task conditions. Understanding one's assessment goal and the approaches behind the different subjective assessment instruments should thus be the first step in selecting a subjective workload assessment instrument. As recommended by Wierwille and Eggemeier (1993), multiple measures should be obtained when feasible.

Acknowledgements

Part of this article was presented at the Seventh International Symposium on Aviation Psychology in Columbus, Ohio, in 1993.

The research was supported in part by Southeastern Center for Electrical Engineering Education, Task SCEEE-HER/89-0005, subcontract under Air Force Aerospace Medical Research Laboratory, Contract No. F33615-88-D-0532. Gary B. Reid was the technical monitor. The research was also supported in part by National Institute of Aging Grant AG08589.

The authors thank Michael Vidulich, Thomas Nygren, Jeffrey Brookings, Richard Backs, Andrew Life, and two anonymous reviewers for their helpful suggestions and comments. They also thank Daniel Voss for his statistical consultation, Bill Bates for his assistance in data collection, and Nicole Schoop-Wyatt for her assistance in data analysis.

Correspondence concerning this article should be addressed to Pamela Tsang, Department of Psychology, Wright State University, 3640 Col. Glenn HWY, Dayton, Ohio 45435-0002, USA. Electronic mail may be sent to ptsang@desire.wright.edu.

References

BROOKS, L. 1968, Spatial and verbal components of the act of recall, *Canadian Journal of Psychology*, **22**, 349–368.

CORWIN, W. H., SANDRY-GARZA, D. L., BIFERNO, M. H., BOUCEK, G. P., LOGAN, A. L., JONSSON, J. E. and METALIS, S. A. 1989, September, Assessment of crew workload measurement methods, techniques, and procedures. Technical Report no. WRCD-TR-89-7006, Vol. I, Wright Research and Development Center, Wright-Patterson AFB, OH.

DODD, D. H. and SCHULTZ, R. F. 1973, Computational procedures for estimating magnitude of effect for some analysis of variance designs, *Psychological Bulletin*, **79**, 391–395.

GOPHER, D. and BRAUNE, R. 1984, On the psychophysics of workload: why bother with subjective measures? *Human Factors*, **26**, 519–532.

GOPHER, D. and DONCHIN, E. 1986, Workload—An examination of the concept, in K. R. Boff, L. Kaufman, and J. P. Thomas (eds), *Handbook of Perception and Human Performance* (Wiley, New York), chap. 41.

GOPHER, D. and KIMCHI, R. 1989, Engineering psychology, *Annual Review of Psychology*, **40**, 431–455.

GOPHER, D., CHILLAG, N. and ARZI, N. 1985, The psychophysics of workload—A second look at the relationship between subjective measures and performance, *Proceedings of the Human Factors Society 29th Annual Meeting*, (Human Factors Society, Santa Monica, 640–644.

HART, S. G. and STAVELAND, L. E. 1988, Development of NASA-TLX (Task Load Index): results of experimental and theoretical research, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (North-Holland, Amsterdam), 139–183.

HUBERTY, C. J. 1972, Multivariate indices of strength of association, *Multivariate Behavioral Research*, **7**, 523–528.

ISREAL, J., CHESNEY, G., WICKENS, C. D. and DONCHIN, E. 1980, P300 and tracking difficulty: evidence for a multiple capacity view of attention, *Psychophysiology*, **17**, 259–273.

JAGACINSKI, R. J. 1977, A qualitative look at feedback control theory as a style of describing behavior, *Human Factors*, **19**, 331–347.

KANTOWITZ, B. H. 1992, Selecting measures for human factors research, *Human Factors*, **34**, 387–398.

LEVINE, M. S. 1977, Canonical analysis and factor comparison. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-006, Sage University, Beverly Hills.

MCCRACKEN, J. H. and ALDRICH, T. B. 1984, Analysis of selected LHX mission functions: implications for operator workload and system automation goals. Technical note ASI479-024084, Army Research Institute Aviation Research and Development Activity, Fort Rucker.

MUCKLER, F. A. and SEVEN, S. A. 1992, Selecting performance measures: 'objective' versus 'subjective' measurement, *Human Factors*, **34**, 441–456.

NORTH, R. A. and REILEY, V. A. 1988, W/INDEX: a predictive model of operator workload, in G. R. McMillan, D. Beevis, E. Salas, M. H. Strub, R. Sutton and L. Van Breda (eds), *Applications of Human Performance Models to System Design* (Plenum Press, New York), 81–89.

O'DONNELL, R. D. and EGGEMEIER, F. T. 1986, Workload assessment methodology, in K. R. Boff, L. Kaufman and J. P. Thomas (eds), *Handbook of Perception and Human Performance: Volume II. Cognitive Processes and Performance* (Wiley, New York), chap. 42.

PEDHAZUR, E. J. 1982, *Multiple Regression in Behavioral Research: Explanation and Prediction*, 2nd edn (Holt, Rinehart, and Winston, New York).

REID, G. B. and NYGREN, T. E. 1988, The Subjective Workload Assessment Technique: A scaling procedure for measuring workload, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (North-Holland, Amsterdam), 185–218.

ROSCOE, A. H. (ed.) 1987, The practical assessment of pilot workload. AGARDograph No. 282, Advisory Group for Aerospace Research and Development, Neuilly Sur Seine, France.

ROSCOE, A. H. and ELLIS, G. A. 1990, A subjective rating scale for assessing pilot workload in flight: a decade of practical use. Technical Report TR 90019, Royal Aerospace Establishment, Farnborough.

RUEB, J. D., VIDULICH, M. A. and HASSOUN, J. A. 1994, Use of workload redlines: a KC-135 crew-reduction application, *International Journal of Aviation Psychology*, **4**, 47–64.

SARNO, K. and WICKENS, C. D. 1991, The role of multiple resources in predicting time-sharing efficiency: an evaluation of three workload models in a multiple task setting. Technical Report ARL-91-3/NASA A3I-91-1, Institute of Aviation, University of Illinois at Urbana-Champaign, IL.

SILVER, N. C. and HOLLINGSWORTH, S. C. 1989, A FORTRAN 77 program for averaging correlation coefficients, *Behavior Research Methods, Instruments, Computers*, **21**, 647–650.

TSANG, P. S. and VIDULICH, M. A. 1994, The roles of immediacy and redundancy in relative subjective workload assessment, *Human Factors*, **36**, 503–513.

TSANG, P. S., SHANER, T. L. and VIDULICH, M. A. 1995, Resource scarcity and outcome conflict in time-sharing performance, *Perception & Psychophysics*, **36**, 365–378.

VIDULICH, M. A. and BORTOLUSSI, M. R. 1988, Speech recognition in advanced rotocraft: using speech controls to reduce manual control overload, in *Proceedings of the American Helicopter Society National Specialists' Meeting—Automation Applications in Rotocraft*, Atlanta Southeast Region AHS, 1–10.

VIDULICH, M. A. and TSANG, P. S. 1987, Absolute magnitude estimation and relative judgement approaches to subjective workload assessment, in *Proceedings of the Human Factors Society 31st Annual Meeting* (Human Factors Society, Santa Monica), 1057–1061.

VIDULICH, M. A., WARD, G. F. and SCHUEREN, J. 1991, Using subjective workload dominance (SWORD) technique for projective workload assessment, *Human Factors*, **33**, 677–692.

WICKENS, C. D. 1986, The effects of control dynamics on performance, in K. Boff, L. Kaufman and J. P. Thomas (eds) *Handbook of Perception and Performance:* Volume II. *Cognitive Processes and Performance* (Wiley, New York), 39/1–39/60.

WICKENS, C. D. 1987, Attention, in P. A. Hancock (ed.), *Human Factors Psychology* (North-Holland, New York), 29–80.

WICKENS, C. D. and LIU, Y. 1988, Codes and modalities in multiple resources: a success and a qualification, *Human Factors*, **30**, 599–616.

WICKENS, C. D., GILL, R., KRAMER, A., ROSS, W and DONCHIN, E. 1981, The processing demands of higher order of manual control, in J. Lyman and A. Bejczy (eds), *Proceedings of the 17th Annual Conference on Manual Control* (Jet Propulsion Lab, La Canada), CA81–95.

WIERWILLE, W. W. and EGGEMEIER, F. T. 1993, Recommendations for mental workload measurement in a test and evaluation environment, *Human Factors*, **35**, 263–282.

First received 16 December 1993.
Substantial revision received 22 December 1994.
Accepted 30 August 1995.

Appendix A. Workload dimensions in the Workload Profile.

1. *Stages of processing*
   (1) *Perceptual/central processing*. These are attentional resources required for activities like perceiving (detecting, recognizing, and identifying objects), remembering, problem-solving, and decision making.
   (2) *Response processing*. These are attentional resources required for response selection and execution. For example, there are three foot pedals in a standard shift automobile; to stop the automobile, we have to select the appropriate pedal and step on it.
2. *Processing codes*
   (1) *Spatial processing*. Some tasks are spatial in nature. Driving, for example, requires paying attention to the position of the car, the distance between the current position of the car and the next stop sign, the geographical direction that the car is heading, etc.
   (2) *Verbal processing*. Other tasks are verbal in nature. For example, reading involves primarily processing of verbal, linguistic materials.
3. *Input modality*
   (1) *Visual processing*. Some tasks are performed based on the visual information received. For example, playing basketball requires visual monitoring of the physical location and velocity of the ball. Watching TV is another example of a task that requires visual resources.
   (2) *Auditory processing*. Other tasks are performed based on auditory information. For example, listening to the person on the other end of the telephone is a task that requires auditory attention. Listening to music is another example.

Note that spatial information may be processed visually or auditorily. For example, you can get to a new restaurant by following a map (visual processing) or by following the directions spoken by your friend (auditory processing). Similarly, verbal information may be processed visually or auditorily. Listening to the news on the radio requires auditory processing of verbal materials; reading the news from the newspaper requires visual processing of verbal materials.

4. *Output modalities*
   (1) *Manual responses*. Some tasks require considerable attention for producing the manual response as in typing or playing a piano.
   (2) *Speech responses*. Other tasks require speech responses instead. For example, engaging in a conversation requires attention for producing the speech responses.