Subsidiary Task Measures of Driver Mental Workload: A Long-Term Field Study

LAWRENCE R. ZEITLIN

Two auditory subsidiary task measures of driver mental workload, delayed digit recall and random digit generation, were evaluated in a 4-year field trial. Vanpool members performed the tasks for 2-min periods while traversing a mix of rural secondary roads, limited access expressways, high-density limited-access urban drives, and downtown city streets on a daily commute from upstate New York to New York City. The delayed digit recall task presented a sequence of random digits to the driver at 2-sec intervals for a 2-min period. The driver was required to recall the digit before the last one presented during the interdigit interval. Errors were scored for digits missed or omitted. The random digit generation task required the driver to verbalize a random sequence of digits from 1 to 9 for a 2-min period. The randomness of the generated sequence was determined by computer analysis. Data collected included the roadway being traversed, time of day, traffic conditions including density and estimated speed, weather, brake applications, and drivers' subjective difficulty ratings. Subsidiary task degradation was a conjoint function of traffic density, average speed, and uncertainty (estimated by the number of brake depressions). Weather conditions moderated these variables. Subjective difficulty ratings correlated with objective criteria of traffic density and speed. Unpredictability of traffic appeared to be the major determinant of perceived difficulty. The digit recall task correlated (r = .834) with a calculated driver workload index based on brake actuations divided by the square root of speed.

A major problem in the study of complex perceptual motor tasks such as driving an automobile is the direct determination of the mental workload imposed on the operator. The concept of mental workload is based on the dual assumptions that

- The act of responding to stimuli takes a finite amount of effort, and
- The total amount of effort that an individual can expend is limited.

For most vehicle control tasks, significant increases in task difficulty may produce few measurable changes in error rate as the operator attempts to hold performance constant by allocating more resources to the task. When all available resources are committed, any further increase in difficulty results in a sharply increased error rate.

The workload that can be handled before performance begins to deteriorate defines maximum capacity for a particular individual. At lower levels of workload, that portion of the individual's resources not committed to the task represents spare capacity. Because of the slow increase in error rate below the break point, the fraction of total capacity allocated to the task is difficult to determine from an examination of error rate alone. By requiring the operator to perform a subsidiary task that uses the unallocated resources, estimates of both spare capacity and primary task workload can be derived. Thus, spare capacity measures, although indirect, offer a means of evaluating operator workload in the less extreme range of operating conditions, before overt errors occur.

Subsidiary tasks have been used by researchers in the estimation of spare capacity and primary task workload for four decades (*1–5*). So many subsidiary tasks have been used, in such specific contexts, that the body of research offers insufficient guidance to the worker who desires general-purpose subsidiary tasks for comparing various implementations of man-machine systems across a variety of conditions (*6–8*).

To be effective, a subsidiary task should meet the following basic requirements:

- Interact minimally with the primary task within the range of conditions covered in the experiment,
- Have more performance degradation as a function of decreased capacity than the primary task, and
- Show performance changes as a monotonic or predictable function of spare capacity.

The ideal subsidiary task would have a number of practical attributes as well. It would

- Use alternative input-output modalities to the primary task (this is implied in the first basic requirement)—since most man-machine control tasks are visual-motor, good subsidiary tasks tend to use auditory inputs and verbal outputs;
- Require minimum learning on the part of the subjects;
- Be resistant to change as a function of repetition;
- Permit transient as well as continuous changes in spare capacity to be measured;
- Require minimum equipment; and
- Be easy to score.

In the course of a U.S. Department of Transportation study on automobile stability and handling, a variety of subsidiary tasks using an oral response mode were screened. They included digit span, mental arithmetic, self-paced generation of random digits, and delayed recall of random digits. Delayed digit recall and random digit generation met most of these

Bernard M. Baruch College, City University of New York, 17 Lexington Avenue, New York, N.Y. 10010.

criteria and were used in the ensuing investigation (5). After the conclusion of the study, other responsibilities intervened and the author discontinued active research on subsidiary task techniques of measuring mental workload.

Commuting to New York City from the surrounding suburbs is uncertain at best. In 1987, in anticipation of a transit strike, several large corporations subsidized employee car- and vanpools. One of the author's graduate students, in his capacity as director of transportation services for a large insurance company, made it possible for the author to conduct a continuing field evaluation of the two subsidiary tasks used in the previous experiment.

METHOD

Overview

Drivers in two commuter vanpools performed two subsidiary tasks (delayed digit recall and random digit generation) while traversing a mix of rural secondary roads, limited-access expressways, high-density urban drives, and downtown city streets on a daily commute from upstate New York to New York City. Data were gathered from 1987 through 1991. Four different vehicles were used during the study, a 1986 Dodge Caravan, a 1987 Plymouth Voyager, a 1988 Dodge Caravan SE, and a 1991 Chevrolet Astrovan LT.

Subjects

The vanpools consisted of six adult men at a time. Driving responsibility rotated on a daily basis. All members lived in the towns of Cortlandt and Yorktown in northern Westchester County and worked in mid-Manhattan. Over the 4-year duration of the study, the membership in each pool changed as the original members moved, changed jobs, changed working hours, or went on vacations or sabbaticals. By the end of 1991, 20 drivers had completed nine full trip trials of each task over the four roadway types, a total of 2,880 two-minute task samples.

Subsidiary Task

Delayed Digit Recall

A sequence of random digits from 1 to 9 was presented to the driver at 2-sec intervals for a 2-min period. The driver was required to say the digit before the last one presented during the interdigit interval. Errors were scored for digits missed or omitted. Each trial consisted of a set of 60 random digits. Before inclusion in the experiment subjects were trained on the task in a no load situation until a criterion of 98 percent accuracy (one or fewer missed digits) was achieved. Training took no more than two or three trials.

Random Digit Generation

During selected 2-min trials, the driver was requested to call out digits from 1 to 9 in random order at a self-paced rate. In past experiments, subjects generally selected a rate averaging two to three digits per second. In these experiments the rate of digit generation did not appear to vary despite major changes in loading on a primary task. Subsidiary task performance was evaluated by determining the randomness of the digit sequence. Evaluation criteria suggested by Wagenaar (9) were incorporated into a computer program. Digit frequency, shifts in the autocorrelation function, and comparison of the subject's digit digram frequencies with those of a mathematically random sequence were used in the analysis. The index of randomness used in this study was the χ^2–estimate of the probability that the observed distribution of digrams matched those of an expected Poisson distribution of a true random sequence of the same length.

Roadway Characteristics

Both commutes started at 7:00 a.m. in northern Westchester County, one in Yorktown Heights, the other in Peekskill. Members were picked up by 7:30, and the designated driver took his position. Each van traveled rural secondary roads (NY-129, Croton Avenue, Furnace Dock Road, and NY-9A) before entering the limited-access road system. Posted speeds on the rural roads ranged from 48 to 72 km/hr (30 to 45 mph). Traffic density was moderate, two to four cars visible. During the warm months, foliage screened driveways. Depending on starting location, the average one-way distance on secondary roads was 24 to 28 km (15 to 18 mi). After leaving the secondary roads, the vans entered the New York Thruway, a high-speed limited-access expressway. The posted speed limit was 89 km/hr (55 mph), but traffic exceeded the limit by 8 to 16 km/hr (5 to 10 mph). During the commuting hours, traffic density was moderate, four to eight cars visible but all moving in the same direction. Thruway distance was approximately 26 km (16 mi). At the city limits, the Thruway changed into a high-density limited-access urban drive, the Major Deegan Expressway. The vans crossed the East River to FDR Drive. Posted speeds on both roads are 80 km/hr (50 mph), but actual speed ranged from 100 km/hr (62 mph) to a bumper-to-bumper crawl depending on traffic conditions. Observed traffic density was extremely high during commuting hours, 20 or more cars visible. The limited-access urban drive distance was 23 km (14 mi). Downtown city streets comprised the final lap. Bidirectional traffic, cross traffic, pedestrian traffic, buses, taxicabs, and traffic lights required constant alertness and mandated low speeds and repeated brake applications. Depending on destination, city street distance was 5 to 8 km (3 to 5 mi). The roadway sequence was repeated in reverse order every evening.

Data Collection Method

The driver was requested to perform the assigned subsidiary task for a 2-min period during each of the four roadway conditions on the inbound commute. The process was repeated on the outbound commute. The same subsidiary task was used for all trials of a given commute. The tasks were alternated at each complete rotation of drivers, about every 2 weeks. Task stimuli were presented and responses recorded with paired cassette tape recorders. An event counter was wired into the brakelight circuit of each van, incrementing the count each time the brake was depressed.

At the conclusion of each trial, the driver was asked to rate the difficulty of the drive on a 10-point scale. In addition to initiating the task, the observer noted the roadway being traveled, time of day, traffic conditions including vehicle density and estimated speed, weather conditions, drivers' difficulty rating, and the starting and ending count on the event counter. Observer notations were made during and at the conclusion of each 2-min trial. No specific instructions were given to the observer to average density and speed observations for the trial period, but observers appeared to do so as a matter of course. As a backup, the observer snapped a picture of the roadway through the front window with a point-and-shoot camera equipped with a date and time recording back. This proved unnecessary and expensive and was dropped after several weeks. Initially, the observer was either the author or a graduate student. Several months into the experiment, vanpool members participated as observers using prepared data forms. Interobserver reliability was assessed by comparing the speed and traffic density estimates on known stretches of roadway. The estimates of all observers were similar. The author monitored data gathering periodically during the duration of the experiment.

Design Limitations

Roadway type is the only variable that can be considered independent in the classical experimental sense. Sequence of presentation, time of day, and traffic conditions (including speed and density) were largely determined by the invariant nature of the commute. A rough form of ABBA counterbalancing was provided by the reverse presentation of conditions on the homeward leg. Weather conditions, a factor that could be expected to influence driver workload, varied by season, but the long duration of the experiment had the effect of randomizing it over subjects. For safety reasons, data were not collected during extremely bad weather.

RESULTS

Observer data on roadway traffic density, average speed, and brake actuations for all drivers are given in Table 1. Even though the roadways clearly vary in all dimensions contributing to driver workload, no apparent differences in driver performance were noted during the course of the experiment. Data were collected on 180 days over 4 years. During that time the vans traveled more than 36,000 mi. Five serious driving errors were recorded: three resulting in moving violation summonses (two for failure to observe stop signs, one for speeding), and two in minor paint-scraping incidents in dense traffic. It is certainly true that an external observer might have noted occasional deviations from prudent driving practice; no consequences resulted from these errors. Such constancy implies that the driver was allocating resources or modifying driving behavior to stabilize performance.

Traffic density, as estimated by the number of vehicles visible through the front window, was dependent both on the nature of the roadway and the time of day. The commute, starting at the beginning of rush hour, encountered progressively greater traffic as the roadways converged on New York City. Peak density was observed on the urban limited-access drives. The number of vehicles observed varied greatly from moment to moment as traffic was slowed by bottlenecks and construction sites. City street density was largely regulated by traffic lights and pedestrian movement and varied substantially less than did the density on urban drives.

Speed limits were posted on all roadways, but actual driving speed appeared to be determined by the conjoint interaction of roadway characteristics, traffic density, weather, and the driver's comfort level. At the gross observational level, speed adjustment appeared to be the primary way of maintaining a comfortable workload. Drivers exceeded the posted limit on rural roads and expressways and dropped below the posted limits on urban drives and city streets.

The number of brake actuations was also determined by roadway characteristics, traffic density, and weather. In contrast to speed, the driver has little choice about braking. A high rate of brake actuation reflects a similarly high degree of uncertainty and unpredictability in the driving situation. Expressway driving, with good visibility, gentle curves, and constant speed allowed minimal use of the brake. City driving, by contrast, required brake actuation on an average of every 7 sec.

The subjective assessment of task difficulty confirms the direct observations. Table 2 presents a summary of driving difficulty ratings and subsidiary task measures for the different roadways. The driving difficulty rating is the mean value of all ratings for a given condition, taken just after the completion of the subsidiary task. The ratings are based on a scale of 1 to 10, where 1 is referenced to an extremely easy driving experience and 10 is referenced to an almost impossible, anxiety-provoking experience. The ratings are completely subjective, but the low standard deviations attest to considerable agreement. Rural driving and expressway driving were viewed as relatively unstressful, urban driving was near the midpoint of the range, and city driving was considered both stressful and unpleasant.

TABLE 1 Observed Roadway Characteristics

Roadway Type	Vehicle Density	Density Std. dev.	Av. Speed (kph)	Speed Std. dev.	Brake Actuations	Brake Std. dev.
Rural road	2.7	.92	53.3	3.96	5.5	1.46
Expressway	6.7	2.01	100.9	8.59	1.6	1.07
Urban drive	17.2	7.71	66.9	5.61	6.6	2.67
City street	10.1	2.45	19.0	1.76	17.1	4.52

TABLE 2 Driver Response Data

Roadway Type	Difficulty Rating	Rating Std. dev.	Digit recall Task errors	Recall Std. dev.	Random[1] Task Index	Workload[2] Index
Rural road	2.74	.73	2.60	.81	.165	.751
Expressway	2.76	.72	1.54	.64	.165	.154
Urban drive	4.22	1.32	5.29	.89	.082	.806
City street	6.15	1.84	9.48	1.54	.082	3.898

[1]Index = probability of "true" random sequence. Low load and high load conditions grouped.
[2]Workload index = (brake actuation rate) / $(\text{speed})^{0.5}$

The ratings are confounded with time of day. Rural and expressway driving took place both in the morning when the driver was fresh and near the end of the commute when an evening's relaxation could be anticipated. By contrast, urban and city street driving difficulty assessments might have been influenced by driving frustration and work fatigue. The present data analysis procedure does not provide a means of isolating morning and evening commutes to provide an estimate of fatigue effects. However, if the concept of mental workload is treated as an integrative construct, incorporating all psychological, emotional, and physical moderators of behavior, then such unconfounding is unnecessary. An analysis of variance (ANOVA) showed that the differences in difficulty ratings of the four roadway types were significant at $p < .001$ level.

Table 2 also gives a summary of errors made in each condition on the delayed digit recall subsidiary task. Delayed digit recall was sensitive to transient aberrations in the traffic flow, the driver forgetting or omitting responses during periods of heaviest load. Digit recall was quite accurate during rural and expressway driving; urban and city driving showed a higher error rate. The errors in these latter regimes appeared to be due to subjects' omitting or dropping digits during unanticipated avoidance maneuvers or brake applications. Indeed, the number of times the van was cut off by a vehicle abruptly changing lanes on city streets could be estimated by looking at the sequence of errors on this task. An ANOVA showing that the digit recall error differences for each roadway type were significant at the level $p < .001$.

Errors made on the delayed digit recall task correlate highly ($r = .784$) with subjective assessments of driving difficulty. Figure 1 plots the average difficulty rating across all conditions against the average delayed recall error rate for the 20 subjects.

The random digit generation task failed to discriminate between the four roadway conditions, although it was quite effective at separating the two conditions with low average difficulty (rural and expressway driving) from the two conditions with high average difficulty (urban and city driving). Direct observation of the driver's performance provided the reason for this insensitivity. During all of the driving regimes, the driving task appeared to be composed of two components: (a) a steady-state driving workload dictated by roadway conditions, speed, and traffic density and (b) a transient workload determined by the degree of uncertainty and unpredictability in the driving situation. The steady-state driving task may have been moderately difficult as a whole, but there were occasional moments of high difficulty. The integrative nature of the random digit generation subsidiary task tended to bury these moments of high difficulty, averaging them over the entire duration of the digit sequences. An ANOVA confirmed that the random digit generation task can discriminate between the low and high average workload situations at about the $p = .01$ level of significance.

A metric combining steady-state and transient conditions provides a reasonable estimate of driver workload. An empirical index of workload was established by dividing the number of brake actuations per 2-min test period by the square root of the observed speed. The results of this calculation averaged across all drivers for each roadway type are given in Table 2. An ANOVA showed that the workload index differed by roadway types at the $p < .001$ level of significance.

Figure 2 compares difficulty ratings, digit errors, and the workload index for the four roadway types. The workload index, derived from directly observed driver actions, tracks the subjective difficulty index quite well. The workload index correlates at a moderate level ($r = .389$) with the subjective driving difficulty ratings and at a higher level ($r = .834$) with errors on the digit recall task. Errors on the digit recall task prove to be a sensitive estimate of both subjective difficulty and the calculated workload index. This high degree of relationship lends confidence to the assumption that the digit recall subsidiary task provides a good measure of spare capacity and can be used to infer primary task workload.

DISCUSSION OF RESULTS

This experiment confirmed the fact that subsidiary tasks can provide a useful measure of spare capacity in the field and can, by inference, provide an estimate of driver mental workload. The two tasks used in this experiment have entirely different characteristics and should be matched to the specific problem at hand. The random digit generation task is of rel-

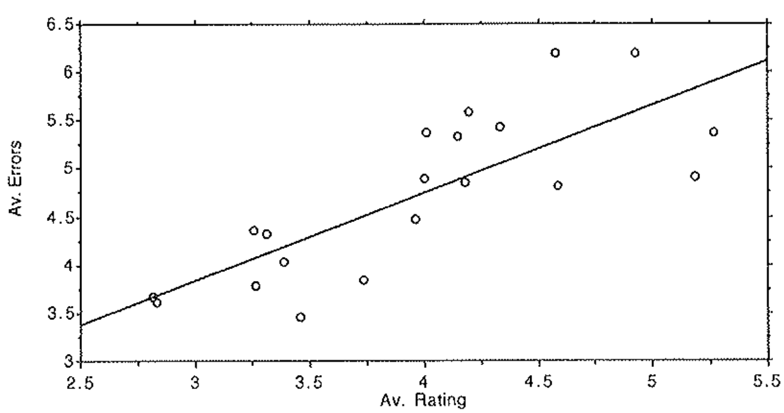

FIGURE 1 Delayed digit recall errors by ratings of subjective driving difficulty ($y = .914x + 1.099$, $R^2 = .615$).

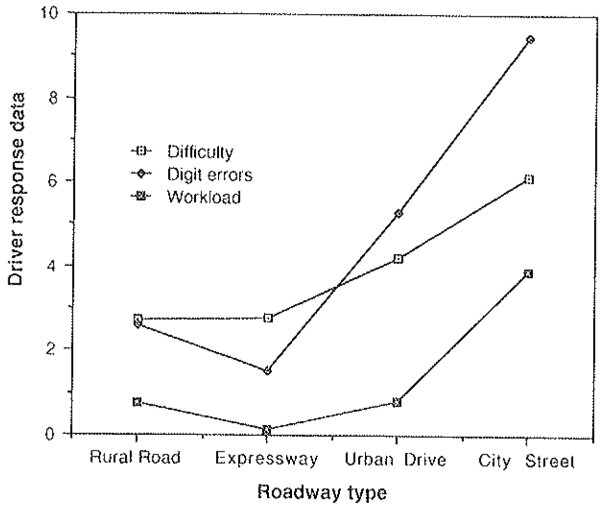

FIGURE 2 Driver response measures by roadway type.

atively little use in dealing with control situations in which primary task difficulty varies transiently over a significant range, such as is typical of "real world" automobile driving. It is of use, however, in evaluating the effect of general loading conditions or steady-state environmental stressors. The unstructured and self-paced nature of the task, coupled with its algorithmic method of scoring, suggests that it could also be useful as a simple measure of driver alertness and physical state.

In contrast, the delayed digit recall task is clearly sensitive to gradations of primary task loading. The high correlation between this task and the driving difficulty ratings suggests that a reasonable estimate of the driving difficulty and consequent mental workload imposed by a given vehicle-roadway combination can be derived by a practical subsidiary task. The experimental workload index,

$$\text{workload} = \text{brake actuation rate/speed}^{0.5}$$

is a parsimonious metric combining transient (brake actuations) and steady-state (square root of speed) conditions and offers a simple but directly observable estimate of subjective driving difficulty.

Interestingly, both subsidiary tasks appeared more sensitive to primary task loading in the field than on a simulator. The allocation of portions of total mental capacity to the tasks at hand requires a judgment call on the part of the individual. Since the consequences of a driving error on a simulator are considerably less serious than the consequences of an error on the roadway, the simulator driver may choose to allocate more capacity to the subsidiary task, decreasing simulated driving performance. No such tendency was noted in this experiment.

REFERENCES

1. H. P. Bahrick, M. Noble, and P. H. Fitts. Extra Task Performance as a Measure of Learning a Primary Task. *Journal of Experimental Psychology*, Vol. 48, No. 4, 1954, pp. 298–303.
2. I. D. Brown and E. C. Poulton. Measuring the "Spare Mental Capacity" of Car Drivers by a Subsidiary Task. *Ergonomics*, Vol. 4, No. 1, 1961, pp. 35–40.
3. I. D. Brown. Measuring the "Spare Mental Capacity" of Car Drivers by a Subsidiary Auditory Task. *Ergonomics*, Vol. 5, 1962, pp. 247–250.
4. L. R. Zeitlin and J. M. Finkelman. A Subsidiary Task Evaluation of the Information Processing Load in a Vehicular Simulation. *APA Exp. Pub. System*, Vol. 3, Ms 106B, 1969.
5. L. R. Zeitlin and J. M. Finkelman. Subsidiary Task Techniques of Digit Generation and Digit Recall as Indirect Measures of Operator Loading. *Human Factors*, Vol. 17, No. 2, 1975, pp. 218–220.
6. G. Johannsen, N. Moray, R. Pew, J. Rasmussen, A. Sanders, and C. D. Wickens. Final Report of Experimental Psychology Group. In *Mental Workload: Its Theory and Measurement* (N. Moray, ed.), Plenum, New York, 1979, pp. 101–104.
7. N. Moray. Subjective Mental Workload. *Human Factors*, Vol. 24, 1982, pp. 25–40.
8. W. W. Wierwille, M. Rahimi, and G. Casali. Evaluation of 16 Measures of Mental Workload Using a Simulated Flight Task Emphasizing Mediational Activity. *Human Factors*, Vol. 27, 1985, pp. 489–502.
9. W. A. Wagenaar. Generation of Random Sequences by Human Subjects: A Critical Survey of Literature. *Psychological Bulletin*, Vol. 77, No. 1, 1972, pp. 65–72.

Publication of this paper sponsored by Committee on Vehicle User Characteristics.