Double Trade-off Curves with Different Cognitive Processing Combinations: Testing the Cancellation Axiom of Mental Workload Measurement Theory

Herbert A. Colle and Gary B. Reid

Human Factors: The Journal of the Human Factors and Ergonomics Society 1999 41: 35
DOI: 10.1518/001872099779577327

Double Trade-off Curves with Different Cognitive Processing Combinations: Testing the Cancellation Axiom of Mental Workload Measurement Theory

Herbert A. Colle, Wright State University, Dayton, Ohio, and **Gary B. Reid,** Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio

The accomplishment model of average mental workload – a formal axiomatic measurement theory – was used as a basis for developing and testing secondary task indices of mental workload (H. A. Colle & G. B. Reid, 1997). Its cancellation axiom implies global sensitivity, which is an important theoretical and practical criterion for mental workload indices. Performance levels of different secondary tasks were empirically equated in mental workload and then used to test the cancellation axiom. Cognitive processing similarity – including orthographic, phonemic, and semantic processing of pairs of operator and secondary tasks – was manipulated in three experiments. Equivalencies between secondary tasks were independent of secondary-operator task similarity, consistent with the cancellation axiom and the global sensitivity of these secondary tasks. The results suggest that standardized secondary task techniques can be developed for the practical measurement of mental workload. Actual or potential applications of this research include the development of functionally useful and realistic secondary task measures of mental workload.

INTRODUCTION

The conceptual bases of mental workload measurement have been supported by a three-legged foundation provided by subjective, performance, and physiological indices (Gopher & Donchin, 1986; O'Donnell & Eggemeier, 1986). Together, these three legs have the potential for providing convergent validity for mental workload measurement procedures. In current practice, however, mental workload assessments are primarily balanced on only one of the legs (Wierwille & Eggemeier, 1993). Subjective indices have been recommended for use in practical evaluations because they have been found to be globally sensitive and easy to use. In contrast, secondary task performance indices, although of conceptual and theoretical interest, were deemed less practical because they have not been found to be globally sensitive and are difficult to use. If secondary task indices are to provide convergent validity for subjective indices, globally sensitive secondary task indices need to be developed.

Globally sensitive indices are those that detect mental workload changes in a wide variety of tasks and situations. In practical assessments of mental workload, only a single index may be used, and evaluators need to have confidence that it will yield useful information. Not only must it be sensitive (that is, likely to detect differences), but it must also be sensitive to a wide variety of design manipulations. Global sensitivity of secondary task mental workload indices has been determined by manipulating primary task difficulty to determine if secondary task performance covaries with the manipulations (Wierwille & Eggemeier, 1993).

Secondary task indices have been of considerable theoretical interest, but this interest has

Address correspondence to Herbert A. Colle, Department of Psychology, Wright State University, Dayton, OH 45435-0001; herbert.colle@wright.edu. **HUMAN FACTORS,** Vol. 41, No. 1, March 1999, pp. 35–50.

been directed at their diagnosticity as opposed to their global sensitivity. Research has been guided by multiple resource theories, which assume that more than one type of resource or underlying capacity exist and that performance on tasks may demand differential amounts of these resources (Gopher & Donchin, 1986). Accordingly, pairs of tasks that demand the same resources should be more difficult to perform together than those that demand different resources and do not have to compete. The focus has been on trying to find laboratory tasks that show these interactions – that is, diagnostic tasks that demonstrate the differential sensitivity predicted by multiple resource theories.

Not only has the search for secondary task indices been focused on diagnostic laboratory tasks developed to test multiple resource theories, but it has been focused on instantaneous capacity, instantaneous resources, or single location bottlenecks consistent with an assumption that "workload at any specific instant of measurement" (Gopher & Donchin, 1986, p. 42) is the appropriate time unit. Thus, even slight disruptions of indices, such as reaction time, are interpreted as reductions in capacity. These performance indicators do not allow operators the freedom to redistribute activity slightly so as to accommodate both primary and secondary tasks, accomplishing both in a timely manner but with some shifts in task sequencing. Subjective judgments, in contrast, may have been found to be more globally sensitive, because operators base their judgments on longer time intervals rather than on one or a few moments in time. In order to be comparable to subjective indices, secondary task indices may have to be chosen so that they are representative of mental work over longer time intervals. Thus, rather than concluding that all secondary task indices inherently lack global sensitivity, we may need to develop alternative means of specifying tasks so that they are more globally sensitive.

Global sensitivity also may not have been found because secondary tasks are sensitive in different performance ranges, as generally would be expected from diverse performance-resource functions (Norman & Bobrow, 1975). Sensitivity depends on the slope of the performance resource function. Secondary tasks may dissociate on a test if they are sensitive in different performance ranges. The most extreme case would be when one of the tasks is performed in a data-limited range, even if both tasks depend on the same underlying resource. However, if secondary task indices of mental workload are calibrated by equating them, inconsistencies attributable to differential range sensitivity could be eliminated by specifying the indices in equivalent workload units.

Finally, secondary task procedures for assessing mental workload need to be simplified for practical measurement. Over the years, secondary task methodology has moved from the relative simplicity of early methods (Knowles, 1963) to recommendations of procedurally complex and costly methods (Gopher & Donchin, 1986). Our research has been directed at resolving this impasse by providing a simple, yet theoretically sound, methodology for developing secondary task indices that are globally sensitive, thus providing another leg for mental workload measurement.

Mental Work

Colle and Reid (1997) provided a theoretical basis for the development of secondary task measures of mental workload: the accomplishment model of average mental workload, a formal axiomatic measurement theory. The model addresses the issues of global sensitivity, instantaneous versus average intervals, and differential range sensitivity, and suggests more simplified practical measurement techniques. We argued that the concept of mental workload is an applied construct and, in particular, that it does not have a one-to-one relationship with attentional capacity or resources in information processing theories. Most important, the theoretical focus is on the amount of mental work that can be accomplished rather than on "the interaction between a person and a task that cause task demands to exceed the person's capacity to deliver" (Gopher & Donchin, 1986, p. 3). For example, if you write 5 pages of a paper that is expected to be 25 pages long, then 5 pages of prose is taken as a performance index of the amount of mental work accomplished

rather than as a failure to complete an expected task demand of 25 pages of prose. Time is introduced explicitly in order to develop a measure of average mental workload. Note that 5 pages could be completed in 1 h or in 10 h, which would change the average mental workload but not necessarily the total amount of mental work accomplished. Of course, in this example, other mental work may have also been performed during the time interval, and if so, it should be indexed also. The important factor is the total amount of mental work that can be completed during the time interval *T*, a specified unit of time. Thus, mental workload is considered to be the average rate of mental work that can be successfully completed or accomplished during this fixed time interval.

According to the accomplishment model, one class of tasks can be developed into secondary task indices most readily. These are cognitive classification tasks in which task performance levels are specified by the number of correct classifications completed during a fixed time interval, for example 1 or 2 min. Therefore, more mental work is accomplished when more correct classifications are made. In participant-paced serial classification tasks, a new stimulus is presented after each response. This performance indicator, called a system's working rate, has been described previously as an indicator of system performance, in contrast to reaction time, response latency, or system lag (Broadbent, 1971). Working rate indicators – the number of correct serial classifications – are used to index mental workload performance levels in this paper.

Mental Workload Equivalence Curves

As an extensive measurement theory, the accomplishment model empirically defines three fundamental elements: empirically identified mental workload entities, empirically specified concatenation operations, and empirically defined equivalence operations. Although the general definitions are broader (Colle & Reid, 1997), in this paper they are defined as follows:

1. Mental workload indices are specified as the number of correct classifications completed in 1 min on a task.
2. Mental work is concatenated over 1-min intervals by having participants complete *m* classifications on one task and *n* classifications on another task during the same time interval.
3. Equivalences are determined by using the method of double trade-off curves.

Double trade-off curves can be used to determine secondary task performance levels that are equivalent in mental workload (Colle, Amell, Ewry, & Jenkins, 1988; Colle & Reid, 1997). A double trade-off curve is generated by manipulating the difficulty of an operator task (i.e., primary task) and observing performance on two different secondary tasks. Table 1 shows hypothetical data generated by the method of double trade-off curves. Difficulty levels of an operator task, labeled with ordinal values 1, 2, 3, 4, and 5, are in column 1 in order of increasing difficulty. Columns 2 and 3 show performance on Secondary Tasks A and B, given by the number of correct participant-paced classifications made when each task was paired with the operator task. Performance on

TABLE 1: An Operator Task and Hypothetical Data on Two Secondary Tasks for Use in the Method of Double Trade-Off Curves

Operator Task Difficulty Level	Secondary Task Performance	
	Task A	Task B
1	40	35
2	35	25
3	30	15
4	20	10
5	10	5

the two secondary tasks can be directly compared at each loading level of the operator task. For example, at Difficulty Level 1, 40 classifications were made on Secondary Task A, but only 35 classifications were made on Secondary Task B. At Level 4, secondary task performance was 20 for Task A and 10 for Task B, respectively.

Assuming a limit on the amount of mental work that can be accomplished in the fixed time interval, less mental work can be completed on a secondary task because more mental work is performed on the operator task. Given that the primary operator task requires a portion, then each of the two secondary tasks is performed using the same remainder. Thus, performance indices for the two secondary tasks denote equivalent amounts of mental work.

Plotting Secondary Task B's performance index against Secondary Task A's performance index as operator loading level varied (column 3 against column 2; Table 1) generates a mental workload equivalence curve that equates performance indices at several levels. Thus, mental workload of Task B can be specified in terms of units of Task A, or vice versa. Eventually, many secondary tasks could be specified in common mental workload units. The method of double trade-offs, however, is robust; it always yields equivalencies if trade-offs are generated. In order to be meaningful, secondary task equivalencies must generalize across the operator tasks used to generate them.

Cancellation Axiom and Global Sensitivity

The capability of generalizing across operator tasks is what Wierwille and Eggemeier (1993) called "global sensitivity." To test for it, they recommended that multiple measures of mental workload be evaluated by manipulating loading of operator tasks. This evaluation procedure is related to the cancellation axiom (also called the monotonicity axiom) of extensive measurement (Krantz, Luce, Suppes, & Tversky, 1971). Informally, the axiom is a statement that "equals added to equals are equal," and is important as an extensive structure for the accomplishment model. It was formally described by Colle and Reid (1997) and can be tested by examining sets of secondary task equivalence curves. The axiom predicts that there should be no special interactions between tasks. Therefore, if two secondary task performance indices are equated, this equivalence should hold regardless of the operator task with which they are paired. Specifically, the cancellation axiom predicts that a mental workload equivalence curve for two secondary tasks should be independent of the operator tasks used to generate the double trade-off curves. Not only is this global sensitivity, but it is also even more specific.

Global sensitivity requires only that secondary tasks be monotonically related. The cancellation axiom requires the two secondary task equivalence curves to be superimposed as well. This property is important theoretically for interpreting equivalencies between secondary tasks, and it is important practically for generalizing beyond the specific tasks used in an evaluation.

Previous tests of the cancellation axiom have been consistent with the accomplishment model (Colle et al., 1988). In these tests, operator tasks were simulated cockpit data entry tasks using two different chord keyboards and the mathematical processing task from the Criterion Task Set (Shingledecker, 1984). The secondary tasks were auditory participant-paced classification tasks. These tests demonstrated the viability of the cancellation axiom in complex situations with multiple dimensions, although parameters relevant to multiple resource theories and predicted task-specific interactions were not systematically manipulated (Gopher & Donchin, 1986).

The aim of the present research was to systematically manipulate tasks requiring different types of cognitive processing to provide a more stringent evaluation of the cancellation axiom. In Experiment 1, operator and secondary tasks were chosen that minimize the amount of cognitive processing overlap between dual task pairs while manipulating memory load. Vocal responses were made to auditory operator tasks that required auditory digit transformations. Key-press responses were made to visual secondary tasks that required memory classifications of geometric stimuli. In Experiments 2 and 3, tasks requiring different types of processing were paired so that dual task pairs required similar or different types of cognitive

processing. In all three experiments, the cancellation axiom predicts that mental workload equivalence curves obtained with different operator tasks will be superimposed.

EXPERIMENT 1

Method

Participants. There were 16 male students recruited from the campus of Wright State University who were paid to participate in this experiment. An additional 4 male students were tested, but they were replaced by 4 other students, as described in the results.

Task descriptions. The two secondary tasks were visual varied-set memory classification tasks using geometric shapes as stimuli (Checkosky, 1971). One task had a memory set size of two (ms2) and the other one had a memory set size of four (ms4). Test stimuli were presented successively on a Techtronics 604 monitor under participant-paced control until a trial terminated. A stimulus remained displayed until a response was made, and it was then followed immediately by the next stimulus.

For the test trials of each secondary task, 24 test stimuli lists and 24 memory sets were generated, and each list was used only once by a participant. Test stimuli were randomly generated so that repetitions did not occur and positive and negative stimuli occurred equally often. Memory sets were generated by classifying the 10 geometric figures into 3 subsets – (a) circle, ellipse, half-circle, heart; (b) triangle, diamond, trapezoid; and (c) square, rectangle, parallelogram – and selecting items by randomly sampling first from different subsets until all subsets were represented at least once. Geometric figure use was balanced over the trials of each block. Negative sets always consisted of all of the items not used in the positive memory set.

Two auditory addition tasks, add1 and add3, were used as operator tasks. They were given priority, and task difficulty levels were systematically varied by manipulating stimulus presentation rate. On add1 task trials, a participant heard a digit and added 1 to it, saying the sum aloud; similarly, on add3 task trials, 3 was added to each digit. Five levels of difficulty were created for each auditory addition task by presenting digits at the rates of 15, 30, 42, 48, and 57 digits/min for the add1 task, and 9, 18, 30, 39, and 48 digits/min for the add3 task. The highest rate was close to the maximum single task rate at which participants could be 100% correct.

Digits were computer-generated by a Computalker speech synthesizer at the appropriate rates and were tape recorded. The tapes were played under computer control by a TEAC 7090 GSL tape deck, and its output was amplified and played to a participant's right ear via a TDH-39 earphone mounted in an MX41/AR cushion (Grason-Stadler, Inc.). All testing was conducted in a single-walled IAC chamber (Industrial Acoustics Corporation). Participants made vocal responses into a microphone placed directly in front of them. Timing tones, which coincided with the onset of the presented digit, were generated to assist scoring. If two vocal response onsets were made, only the first one was scored; that is, false starts were scored as incorrect. Lists of the digits 1–6 were randomly generated without digit repetitions so that each digit occurred the same number of times during the 2-min test period. Each stimulus list was presented only once to a participant and lists were counterbalanced across participants.

Procedure. Each participant completed five sessions, one session per day. Session 1 was a practice session of 20 trials in which each individual task was practiced twice and each of the 4 pairs of concurrent tasks was practiced 3 times. Dual task pairs were formed by having each of the 2 auditory operator tasks performed concurrently with both visual secondary tasks.

Sessions 2 through 5 were test sessions of 16 trials. During each session, one of the auditory operator tasks was paired with both visual memory classification secondary tasks. Participants performed the auditory tasks (add1, add3) in ABBA counterbalanced order across the 4 test sessions. Each session consisted of two blocks of trials, one using the ms2 task and one using the ms4 task. The order of testing the ms2 and ms4 conditions was counterbalanced over participants so that it was balanced over test sessions and over add1 and add3 tasks.

Each block of trials consisted of 8 trials; these trials consisted of 2 practice trials on which the pair of tasks that were tested on that block of trials was used, plus 6 test trials on which a visual secondary task was performed once singly and once with each of the five presentation rates of the auditory addition task. Participants were always told the task conditions for a block. Testing order for the six difficulty levels of the auditory task was randomly determined for the first blocks of Sessions 2 and 3, but was balanced across participants so that each rate was tested at each ordinal position once or twice. The second block of each session used the same rank order of rates with the other visual task. In Sessions 4 and 5, testing orders were the reverse of the orders used in Sessions 3 and 2, respectively, but the other set of auditory lists was used. Auditory tapes were balanced across sessions.

The sequence of events on each trial was as follows. After a participant initiated a trial, the message "Memory set" was displayed on the screen and was followed by the serial presentation of each figure in the memory set for 1.5 s each until the memory set had been presented twice. After the message "Test beginning" was presented, the first visual test figure was displayed at the same time that the first auditory digit was played. Trials were 140 s in length, but the first 20-s period was treated as warm-up and discarded; performance during only the following 2 min was recorded. Participants were told that the auditory operator task had priority and that they had to keep up with it; they were asked to perform the participant-paced visual secondary task as fast as possible while keeping errors low on both tasks. If more than 10% errors were made, participants were informed that they had made too many errors, and a makeup trial with the same rate and conditions was given at the end of the trial block.

Results and Discussion

Mental workload equivalence data from Days 2–3 are presented in the left panel of Figure 1 and the data from Days 4–5 are presented in the right panel of Figure 1. The mean number of correct classifications/min on the ms4 and ms2 secondary tasks are on the horizontal and vertical axes, respectively. Each data point was generated from two different test trials. For each test trial, the mean number of participant-paced correct responses/min made by each of the 16 participants was

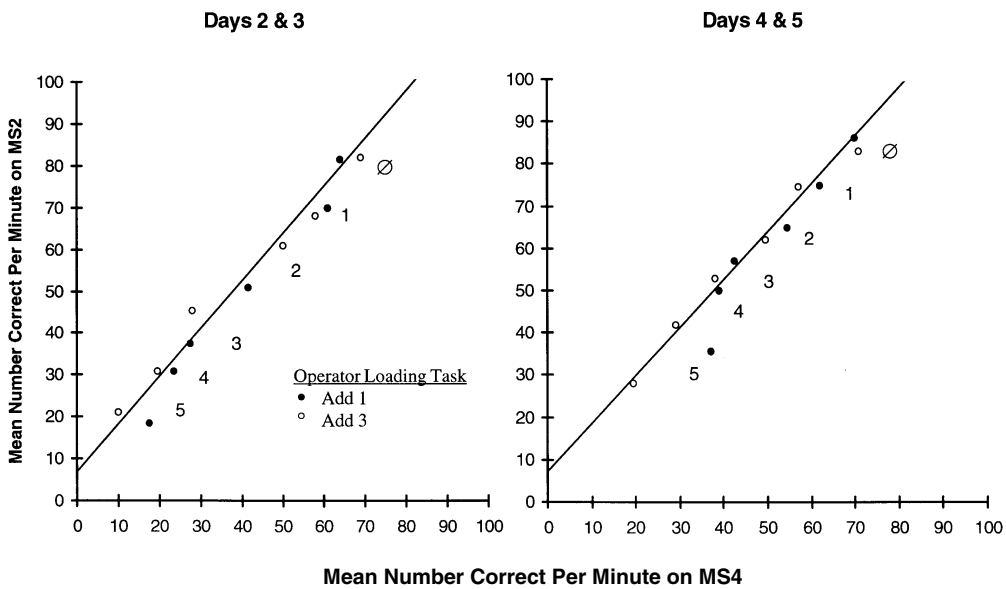

Figure 1. Mental workload equivalence curves for the two memory search tasks in Experiment 1. (MS2 refers to a memory set size of 2, and MS4 refers to a memory set size of 4.)

computed. For example, the data denoted by the filled circle labeled "1" were obtained from a trial when ms4 was paired with add1 at its lowest difficulty level (15 digits/min) and were obtained from another trial when ms2 also was paired with add1 at the same level. The data denoted by open circles were obtained in the same way, except that the ms2 and ms4 classification tasks were each paired with add3. The numbers next to each pair of data points identify the difficulty levels of the operator tasks with ∅ denoting the single task condition when a secondary task was performed singly.

The cancellation axiom implies that sets of filled and open data points should define an identical but unknown function. The ms2-ms4 mental workload equivalence curve, which was estimated by using the add1 task (filled circles), should be the same as the equivalence curve estimated by using the add3 task (open circles). As Figure 1 shows, these two sets of data points were similar on both test days. With more experience, performance seemed to improve (points shifted up from Days 2–3 to Days 4–5), but the data from both replications appear to lie along the same equivalence curve. Practice on the tasks may have reduced the amount of mental workload needed to achieve a given level of performance.

In order to statistically test whether the two sets of data came from the same mental workload equivalence curve, a general purpose procedure developed by Colle et al. (1988) was used, because equivalence curves are bivariate with an unknown function. First, the unknown function was specified by finding a function that provides a reasonable fit to all of the data. Second, the parameters of this function were estimated separately for each of the two equivalence curves. Finally, a multivariate analysis of variance (MANOVA) was used to test the equality of the estimated parameters, using a within-subjects MANOVA because both equivalence curves were collected on the same participants (Timm, 1975).

A linear function provided a good fit to all of the data from add1 and add3 pairs of scores on Days 2–5 in Figure 1. The Pearson correlation coefficient for these points was .98. The best-fitting straight line is presented together with the data in both panels of Figure 1, and it provided a good fit for both replications. The best-fitting slope and intercept are 1.11 and 6.70, respectively.

To statistically test the equivalence of the add1 versus add3 workload equivalence curves, best-fitting straight lines were obtained for each participant separately for add1 and add3 workload equivalence curves. The estimated slopes and intercepts of the add1 workload equivalence curves and the slopes and intercepts of the add3 workload equivalence curves were entered as data in a MANOVA. The multivariate F was not statistically significant, $F(2, 14) < 1.0$, $p > .05$. Separate univariate tests on the slopes and intercepts were also not statistically significant, $F(1, 15) < 1.0$, $F(1, 15) = 1.12$, respectively. Separate analyses of Days 2–3 and Days 4–5 yielded similar results. The multivariate, $F(2, 14)$, was 1.55 for Days 2–3 and was less than 1 for Days 4–5. A univariate analysis of variance (ANOVA) also was conducted on the intercept estimates, using days and addition tasks as factors. There were no differences between these factors or their interaction, but the overall mean intercept was found to be significantly greater than zero, $F(1, 15) = 15.8$, $p < .01$. Comparable results for the aforementioned analyses were obtained with alternate analyses that used only data sets with individual good-fitting lines ($r > .9$) or that did not use single task data (designated by ∅ in Figure 1) in estimating the lines.

Few errors were made on the tasks, as required by the procedure. The percentage error was 1.7% and 2.4%, respectively, for the ms2 and ms4 visual tasks on Days 2 and 3, and 1.6% and 2.0% on Days 4 and 5. Percentage error was 2.8% and 3.4%, respectively, for the add1 and add3 tasks on Days 2 and 3, and 2.7% and 2.4%, respectively, on Days 4 and 5. Trials with more than 10% error on either task were made up at the end of the trial block, and makeup trial performance was used in all of the aforementioned analyses. Alternate treatments of these trials, such as using the original data instead of makeup trial data or dropping both the original and makeup data, did not change the conclusions or appreciably change the means.

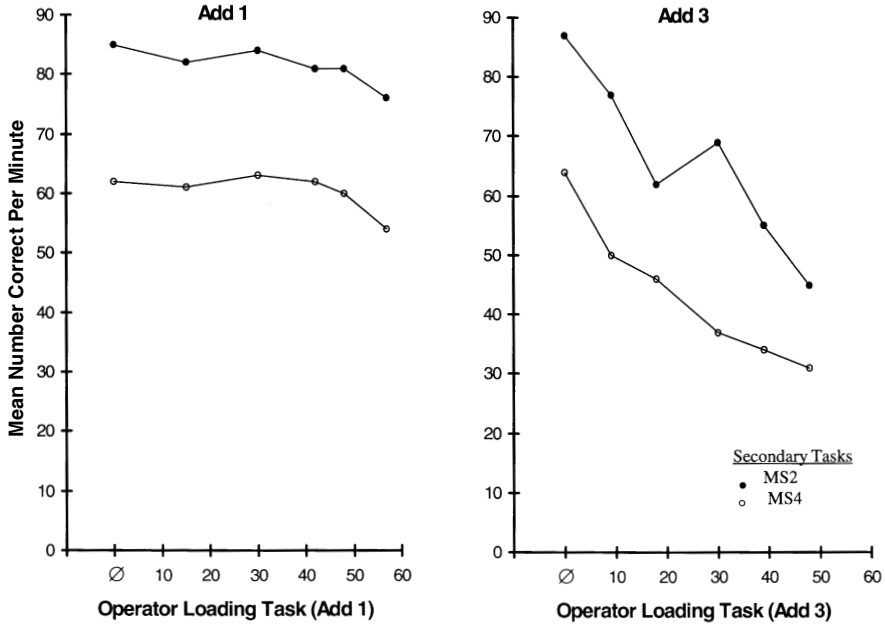

Figure 2. Double trade-off curves from 4 participants who showed independence between the two secondary tasks and the add1 operator task in Sessions 4 and 5.

Of the participants, 4 were replaced by 4 others and were not used in the above analyses. The original 4 participants were replaced because their data did not exhibit a performance trade-off for the add1 operator task during the second replication on Days 4–5. Figure 2 shows the double trade-off curves for the original 4 participants on Days 4–5. At the lowest 4 difficulty levels (1–4) of the add1 operator task, secondary task performance did not appear to decrease and was about the same as single task performance (∅) on the secondary task. Performance trade-off curves for the other 16 participants clearly were not flat on these trials, which is important because, theoretically, equivalence could be determined only if performance trade-offs were demonstrated for all visual-auditory task combinations (Colle et al., 1988; Colle & Reid, 1997). However, the absence of a trade-off on these trials was not diagnostic (i.e., was not a violation of global sensitivity) because both secondary tasks acted similarly. Neither secondary task showed a substantial trade-off for the add1 task, but both did for the add3 task. It is possible that the add1 task required little or no mental work because the cognitive processes needed for the transformation had become automatic for these participants (Shiffrin & Schneider, 1977). Add1 transformations are commonly used in vocal counting and should be highly practiced.

Results of Experiment 1 were consistent with the cancellation axiom. This test, however, was not very demanding because both operator tasks and both secondary tasks required similar processing. More stringent tests would systematically manipulate the overlap between the types of processing used by operator tasks and secondary tasks.

EXPERIMENT 2

In Experiment 2, the cancellation axiom of the accomplishment model was tested by systematically manipulating the cognitive processes needed to perform secondary and operator tasks. Workload equivalence curves were obtained using pairs of tasks that required similar or different cognitive processes. Multiple resource theories predict that the amount of interference between dual tasks should depend on the amount of processing overlap (Gopher & Donchin, 1986). There is no agreed-upon set of multiple resources, and at least three

different sets have been proposed (Polson, Wickens, Klapp, & Colle, 1989), including Wickens's (1980) multidimensional multiple resource theory, hemispheric differences theory (Friedman & Polson, 1981), and working memory theory (Baddeley, 1992).

Three different types of cognitive processing (orthographic, phonemic, and semantic processing) were manipulated in Experiment 2. Besides spanning a wide range of cognitive processing, there is structural evidence that the three types are distinct. Using functional magnetic resonance imaging, these cognitive processes produced differential activity at separate cortical sites (Pugh et al., 1996). Hemispheric differences between orthographic processes and the other two types were also reported. According to multiple resource theories, if an operator task and a secondary task both require similar processing, there should be greater interference between the two dual tasks. Therefore, the cancellation axiom should be violated. If one of the secondary tasks in a workload equivalence curve requires cognitive processing that is similar to the cognitive processing required by the operator task, it should selectively show lower performance compared with performance with an operator task that requires different cognitive processing.

Orthographic and semantic secondary tasks were combined factorially with orthographic and semantic operator tasks to generate dual task trade-off curves. A phonemic secondary task also was paired with both operator tasks so that equivalence curves could relate performance on the orthographic and semantic secondary tasks to performance on this third task. According to the cancellation axiom, mental workload equivalence curves are independent of the operator task used to generate them. Therefore, it predicts that the equivalence curve for orthographic-phonemic tasks will be the same whether it is generated by an orthographic or a semantic operator task. Likewise, it predicts that the equivalence curve for the semantic-phonemic tasks will be the same whether it is generated by a semantic or an orthographic operator task.

Method

Participants. There were 24 students recruited from the campus of Wright State University who were paid to participate. None of the students participated in the other experiments.

Task descriptions. Three visual secondary tasks were used: the third letter task, the rhyming word task, and the categories task. Dot matrix stimulus characters (white on a black background) were presented on a black and white monitor under microprocessor control. Each character was 3 mm wide by 4 mm high and was viewed from about 50 cm. Pairs of words were presented horizontally and simultaneously, and a same-different judgment was required for each pair. Responses were made using the right hand on a specially constructed momentary contact keypad. The forefinger rested on the left key and was used to make "same" responses; the middle finger rested on the right key and was used to make "different" responses.

Each visual secondary task emphasized a different type of processing; the third letter task emphasized orthographic processing. Participants decided whether or not both words had the same third letter. Words were displayed in upper case, were 4 to 6 letters in length, and came from Thorndike and Lorge's (1944) AA count. Only words in which the third letter was c, e, i, l, m, n, o, r, t, or u were used. Lists of word pairs, using a set of 400 "same" and 400 "different" pairs, were generated by sampling without replacement so that every set of 30 pairs had an equal number of same and different stimuli.

The rhyming word task emphasized phonemic processing. Participants decided whether or not both words rhymed. Words were monosyllabic, 3 to 7 letters in length, and came from the A or AA counts in Thorndike and Lorge (1944). Same and different pairs were created so that spelling patterns could not be reliably used as a basis for a decision, and words with 2 common alternative pronunciations were not used. Once again, sets of 30 pairs were generated by sampling without replacement from 200 same and 200 different pairs.

The categories task emphasized semantic processing. Participants decided whether or not 2 words came from the same category. High-frequency words from 10 Battig and Montague (1969) categories were used: musical

instrument, vegetable, article of clothing, metal, four-footed animal, furniture, human body part, weather phenomenon, building part, and flowers. These 100 words were used to create 400 "same" and 400 "different" stimuli, and lists were again generated in sets of 30 pairs.

Two auditory operator tasks were used: the vertical line task and the letter-word task. Auditory stimuli were single letters spoken by a Computalker speech synthesizer under computer control and presented to the left TDH-39 earphone of a headset. Stimuli were equated in intensity and duration. Yes-or-no judgments were made by the left hand on a specially constructed pistol grip with momentary contact keys. The forefinger rested on the trigger switch for "yes" responses, and the thumb rested on the top switch for "no" responses. Assembly language programming allowed response acquisition from either the auditory or visual task or the stimulus presentation from the alternate task to occur while an auditory or visual stimulus was being presented.

The two auditory operator tasks emphasized a different type of processing. The vertical line task emphasized orthographic processing. Participants decided whether or not a letter they heard had a vertical line in it when it was printed in upper case. Of the letters, 8 had vertical lines in them (B, H, K, L, N, P, R, T), and another 8 did not (A, C, O, Q, S, U, X, Z). The letter-word task emphasized semantic processing. Participants decided if the names of the letters they heard were also common nouns, pronouns, or verbs. Nonwords, proper names, and exclamations were to be responded to negatively. Of the letters, 8 satisfied the criterion – b (bee), c (see), i (eye), j (jay), p (pea), r (are), t (tea), u (you) – and 8 did not satisfy the criterion – f, g, h, k, l, m, v, z. There was little or no relationship between the classification of letters in the letter-word task and the classification of letters in the vertical line task. Vowel sounds also did not predict letter classifications well. For both tasks, stimulus lists were constructed so that positive and negative stimuli occurred equally often in sets of 30 stimuli without repetition of letters.

Procedure. Each participant completed three sessions: a practice session of 22 trials, in which they received each of the 5 single tasks twice and each of the 6 dual task combinations twice, and 2 test sessions.

Auditory operator tasks were experimenter-paced and visual secondary tasks were participant-paced. Each auditory operator task was paired with each of the visual secondary tasks. The experimental design was the same as the one used for Sessions 2 and 3 of Experiment 1, except that there were six blocks of trials in each session and six trials per block. In each block, one of the three visual secondary tasks was paired with one of the auditory operator tasks. On each trial, one of the auditory task difficulty levels was used; stimuli were presented at one of the rates, including the null rate when the visual secondary task was performed alone. After all three visual tasks were performed, they were performed again in counterbalanced order, producing six blocks. Only one of the auditory operator tasks was used in a session. Order of testing auditory operator tasks was counterbalanced across participants, and within each auditory task condition the six sequences of visual secondary tasks were used equally often. Overall, tasks and difficulty levels were balanced to control for order of testing effects. Before the start of testing, participants also performed two dual task practice trials.

Trials started with the simultaneous presentation of both the first visual and auditory stimuli when the ready switch was pressed. Auditory stimuli were presented at the prescribed rate, and participants were instructed to keep up with them. Visual stimuli were participant-paced, and the next stimulus appeared immediately after a response was made. Trials were 140 s in duration, but the first 20-s period was not scored. Participants were instructed to respond as fast as they could on the visual secondary task while keeping their errors below 10% on both tasks. If errors were greater than 10%, participants were informed and the trial was run again at the end of the block.

Results and Discussion

Figure 3 presents the mental workload equivalence curves. The left panel shows third letter task performance as a function of

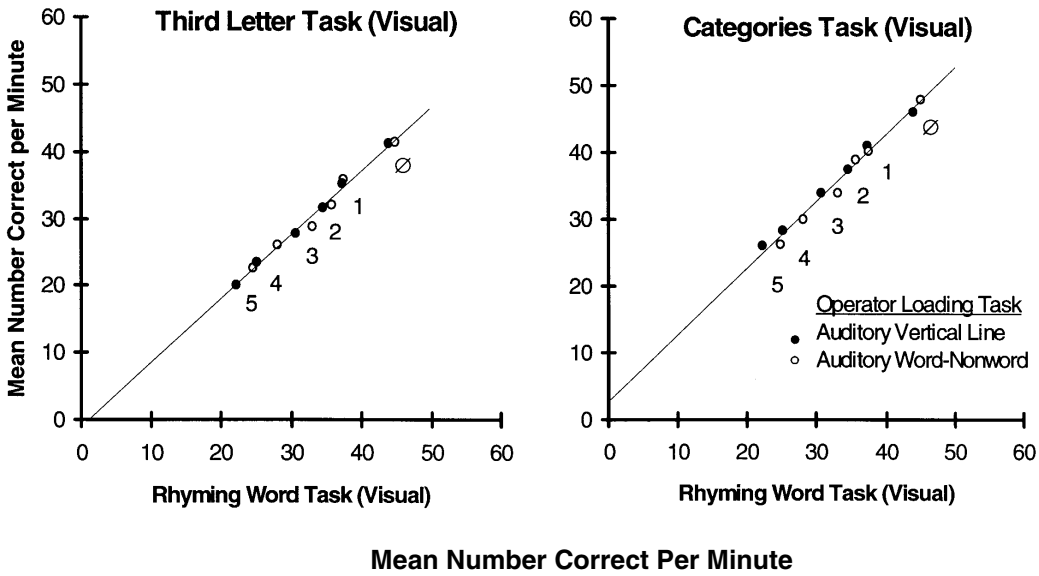

Figure 3. Mental workload equivalence curves for combinations of orthographic, phonological, and semantic tasks in Experiment 2.

rhyming word task performance, and the right panel shows the categories task performance as a function of rhyming word task performance. Filled circles represent data obtained from using the auditory vertical line task to generate trade-offs with the three visual tasks, and open circles represent data obtained from using the auditory letter-word task.

The cancellation axiom predicts that the equivalence curves should be independent of the operator task used. Thus, the opened and closed circles in each panel should fall along the same function. However, if similar processing produces extra interference, the visual third letter task (orthographic) and the auditory vertical line task (orthographic) should be harder to perform together. Thus, the filled circles in the left panel of Figure 3 should lie below the open circles. Likewise, the open circles should lie beneath the filled circles in the right panel if there is extra interference because the auditory and visual tasks both require semantic processing.

Workload equivalence curves were analyzed statistically, as they were in Experiment 1. Again, the workload equivalence curves were described well by linear functions. The correlation coefficient was .99 for the 12 points in the left panel and .99 for the 12 points in the right panel of Figure 3. The best-fitting slope and intercept were 0.96 and –1.18, respectively, for the data in the left panel, and 0.99 and 2.82, respectively, for the data in the right panel. Few errors were made on the tasks. Mean percentages of error on the two auditory tasks were 1.52% and 1.38%. Error percentages for the visual tasks were 2.30%, 3.40%, and 2.60%.

Straight-line functions were again fit separately for each participant and auditory task. Slope-intercept pairs were entered as dependent variables in a MANOVA program to test whether the same function was obtained from each auditory operator task. One MANOVA was used to analyze the third letter task and vowel sound task data, and a separate MANOVA was used to analyze the category task and the vowel sound task. Neither multivariate F was statistically significant – for the data in the left panel, $F(2, 22) < 1.0$, and for the data in the right panel, $F(2, 22) < 1.0$. Separate univariate ANOVAs on the slopes and intercepts were also not statistically significant. Alternate analyses yielded similar results.

The results of Experiment 2, like Experiment 1, were consistent with the cancellation axiom and did not demonstrate diagnosticity. Dual tasks with similar cognitive processes

did not appear to produce more interference than those with less similar cognitive processes.

EXPERIMENT 3

In Experiment 3, the cancellation axiom was tested using another set of secondary tasks – ones that were more similar to the operator tasks and that differed in sensitivity. Tasks again stressed orthographic, phonemic, and semantic processing, but this time judgments were made on letters in both the visual secondary tasks and the auditory operator tasks. Three new visual secondary tasks were paired with the two auditory tasks that were used in Experiment 2. Secondary tasks also were chosen so that they differed in sensitivity over the performance range tested. In Experiment 2, secondary tasks had similar sensitivity; the slopes of the mental workload equivalence curves were close to 1.

Method

The experimental design and procedure were the same as in Experiment 2 except that three new visual secondary tasks were used. Secondary tasks were the same – letter task, the vowel sound task, and the lexical decision task, which emphasized orthographic, phonemic, and semantic processing, respectively. Another 24 students participated in Experiment 3.

The same-letter task emphasized orthographic processing. Participants decided whether or not pairs of uppercase letters were the same; this was a serial, participant-paced version of Posner's (1969) physical identity task. All 20 consonants were used. A set of 200 "same" pairs was created by pairing each letter with itself 10 times. A set of 200 "different" pairs was created by selecting pairs with a graphic distinctiveness between 2.26 and 4.93 on Pavur and Kausler's (1974) uppercase scale. Sets of 30 pairs were generated without repetition, as in Experiment 2.

The vowel sound task emphasized phonemic processing. Participants decided if pairs of uppercase letters had the same vowel sound or not. Sixteen letters that could be grouped into 3 different vowel sound groups were used: (a) j, k; (b) b, c, d, g, p, t, v, z; and (c) f, l, m, n, s, x. With reversals, a total of 88 "same" stimuli were created; a letter was never paired with itself. An equal number of "different" pairs was created by selecting letters from 2 different groups. Sets of 30 pairs without repetitions were again generated.

The lexical decision task emphasized semantic processing. Participants decided if two strings of letters were both words or both nonwords. If both letter strings were words or both were nonwords, the correct judgment was "same." If one letter string was a word and one was a nonword, the correct judgment was "different." Nonwords were created by changing one letter in a word string. This was a serial, participant-paced version of a same-different lexical decision task (Meyer & Schvaneveldt, 1971). A set of 400 same and 400 different pairs was created and used to generate sets of 30 pairs.

Results and Discussion

Figure 4 presents the mental workload equivalence curves. The left panel shows same-letter task performance as a function of vowel sound task performance, and the right panel shows lexical decision task performance as a function of vowel sound performance. Filled circles represent data obtained from dual task trials that used the auditory vertical line task, and open circles represent data obtained from trials that used the auditory letter-word task.

Once again, the cancellation axiom predicts that open and closed circles will yield identical mental workload equivalence curves. Equivalence curves were described well by linear functions; data in the left and right panels both had correlation coefficients of .99. Data in the left panel yielded a slope of 2.10 and an intercept of –0.37; data in the right panel yielded a slope of 0.78 and an intercept of 5.49. Few errors were made on the tasks. Mean error percentages were 1.68% and 1.39% for auditory tasks, and 1.36%, 1.49%, and 3.12% for visual tasks.

Slopes and intercepts from each participant and auditory task combination were obtained, and the slope-intercept pairs were the dependent variables in MANOVA analyses, as in Experiments 1 and 2. Again, neither multivariate F was statistically significant – for the data in

the left panel of Figure 4, $F(2, 22) = 2.36$, and for the data in the right panel of Figure 4, $F(2, 22) = 1.01$. Alternate analyses and univariate F-ratios were also not statistically significant.

The results of Experiment 3 were similar to those found in Experiments 1 and 2 and were consistent with the cancellation axiom. These data suggested that workload equivalence curves did not depend upon operator tasks. The results were still consistent with the axiom, even though secondary tasks differed considerably in sensitivity. The slope of the same-letter/vowel sound equivalence function was 2.7 times steeper than the slope of the lexical decision/vowel sound equivalence function. Thus, similar sensitivity is not needed to obtain global sensitivity. In addition, these results rule out using simple response competition to account for double trade-offs.

GENERAL DISCUSSION

The results from all three experiments were consistent with the cancellation axiom of the accomplishment model. Over a wide range of performance, there was no evidence that mental workload equivalence curves for pairs of secondary tasks depended on the operator task used to generate them. These results and those of Colle et al. (1988) were obtained using a total of 10 different secondary tasks and 6 different operator tasks, requiring a variety of cognitive processing activities. Thus, these data suggest that it may be possible to develop globally sensitive secondary task measures for use in operational test and evaluations or in other applied evaluations. In particular, a battery of secondary tasks with known mental workload equivalencies could be developed by using the

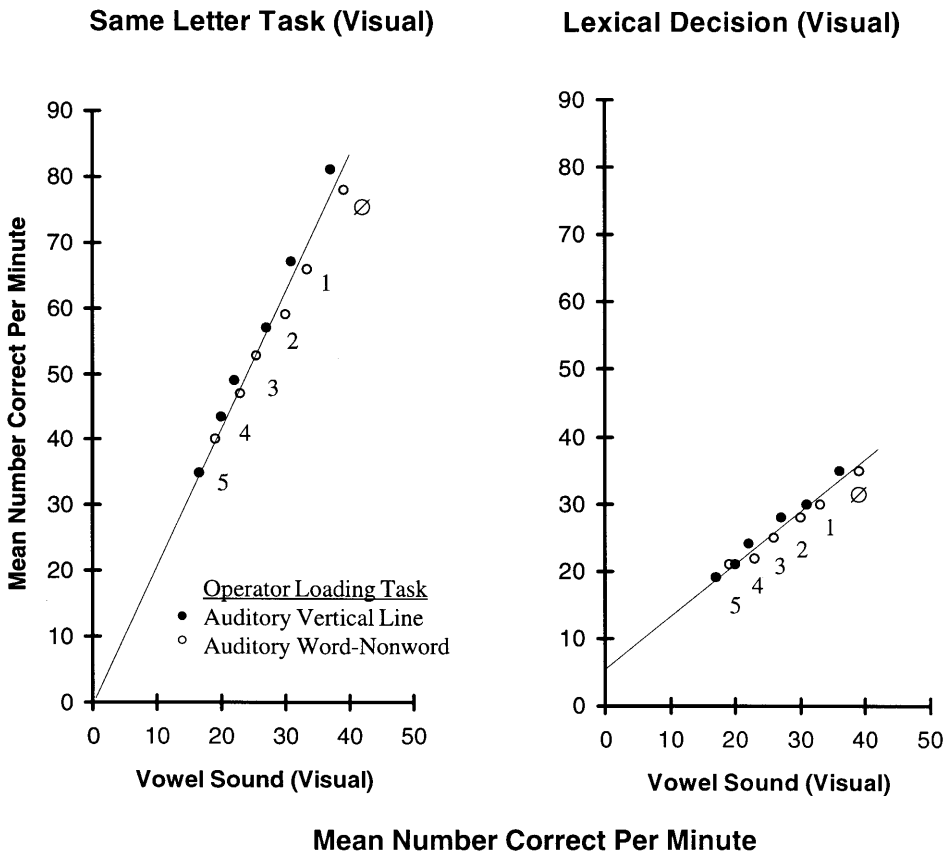

Figure 4. Mental workload equivalence curves for combinations of orthographic, phonological, and semantic tasks in Experiment 3.

double trade-off curve method to estimate equivalence curves. This step alone would greatly facilitate practical mental workload measurement because data from different evaluations could be compared, whereas in any particular assessment, the choice of a secondary task could depend upon the restrictions of that situation. In addition, if the accomplishment model's assumptions are tested and supported, numbers can be assigned to secondary task performance levels so that they represent mental workload units that are unique up to ratio scale transformations (i.e., unique except for multiplication by a constant positive real number). A standardized secondary task testbattery would then have ratio scale mental workload units. The procedures for developing these scales were described by Colle and Reid (1997).

Evidence for global sensitivity of secondary tasks in the current experiments contrasts with repeated failures to find global sensitivity (see Gopher & Donchin, 1986; O'Donnell & Eggemeier, 1986; Wierwille & Eggemeier, 1993). We argued that these inconsistencies may arise because of the way secondary task performance indices are specified. To find globally sensitive secondary task indices, indices should reflect the total mental work accomplished during a fixed time interval, consistent with the accomplishment model. The time distribution of the responses to a secondary task was not deemed critical; thus, reaction time indices in particular, which are sensitive to these local disturbances, were not used.

Townsend and Ashby (1978) made a similar distinction between instantaneous capacity (called power) and the integral of power (called energy). Average capacities may be additive even when instantaneous profiles are not additive. This has been called the integral capacity restriction versus the instantaneous capacity restriction (Colle & Reid, 1997). The accomplishment model assumes only an integral capacity restriction and thus does not place strong constraints on information processing activities that are consistent with it. Processing stages may have a mixture of serial and parallel processes (Schweichert & Boggs, 1984). Dual tasks may be switched and swapped as well as performed concurrently without violating the axioms of the accomplishment model. Other recent mental workload models have also introduced time interval as an explicit factor (Hancock & Caird, 1993; Hendy, Liao, & Milgram, 1997).

On the other hand, tests of the accomplishment model to date have used only cognitive classification or transformation tasks; therefore, it is difficult to generalize to the wider range of operational tasks, such as driving or flying. It is also not clear whether they could be used to detect short periods of elevated mental workload (Verwey & Veltman, 1996). It is notable that mental workload equivalence curves have been well-fit by straight lines. The linear relationship of these secondary task mental workload performance indices needs to be explored to determine the conditions under which linearity is found. Linearity is a desirable property, but not all task indices should be expected to be linear. A nonlinear equivalence relationship was found for the mathematical processing task and the categories task at lower performance levels (Colle et al., 1988).

The data suggest that it may be possible for secondary tasks to join subjective techniques as globally sensitive measures of mental workload. Eventually, it may be possible to demonstrate the comparability of these two types of mental workload measurement techniques. The Subjective Workload Assessment Technique was also developed using a formal axiomatic measurement theory – conjoint measurement (Reid & Nygren, 1988). Colle and Reid (1997) have indicated how formal measurement theory may be used to integrate subjective and secondary task techniques, as well as physiological indices of mental workload.

Secondary Task Assessment Methodology

Issues such as transferability, implementation requirements, and intrusion must be addressed before secondary task indices can be used for practical assessment. Transferability refers to the portability of a technique, and implementation requirements refer to needs related to data collection, instrumentation, and training. Methods directed at using a single universal standard secondary task for all situations are limited in addressing these issues. For example, the visual secondary tasks

used in the present experiments may be inappropriate for evaluating mental workload when the primary task is to find camouflaged visual targets. In the present experiments, pains were taken to avoid possible sensory interactions, such as masking, by presenting operator and secondary task stimuli to separate sensory modalities. In previous experiments, auditory secondary tasks were also used together with visual operator tasks in order to avoid sensory interactions (Colle et al., 1988). By developing a battery of equivalent tasks instead of a single standard task, a secondary task's characteristics can be matched to the restrictions imposed by an operator's environment and mission.

Secondary tasks create problems when they intrude on and impair performance on operators' primary tasks (Wierwille & Eggemeier, 1993). Intrusion may result when operators have expectations that they should perform both tasks even if it is impossible (Gopher & Donchin, 1986). Although it remains to be seriously evaluated, serial, participant-paced secondary tasks may be less likely to be intrusive, especially if participants are clearly instructed not to perform the secondary task whenever it might interfere. There is less demand pressure because, in a sense, a self-paced, serial secondary task waits for the operator. It does not repeatedly poll the operator for a response. Intrusion might also occur because there are task interactions such as masking, which might be alleviated by task selection, as previously described.

It should be realized that the complicated procedures used to develop equivalency curves would not be used in practical evaluations. The calibration procedures that were used in the current experiments are precursors to developing practical secondary task evaluation procedures. Mental workload evaluations can be conducted in two ways: the gauge technique and the criterion technique. The gauge technique yields a point estimate of mental workload using subsidiary task procedures (O'Donnell & Eggemeier, 1986). With this technique, estimates of the level of workload are obtained and can be compared for alternative systems, displays, or conditions. In many cases, however, this detailed information is not necessary; the real question is whether mental workload is too high, exceeding a red-line level. In these cases, the criterion technique could be used, in which an operator tries to perform a secondary task at a specified performance level concurrent with his or her primary task. The criterion technique can be very efficient, but it yields only a pass-fail decision. Either the operator can perform both tasks acceptably or not. In order to develop a criterion procedure, however, a red-line criterion must be defined and a scaled battery of secondary task indices must be available.

REFERENCES

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556–559.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs, 80* (3, Pt. 2).

Broadbent, D. E. (1971). *Decision and stress*. New York: Academic.

Checkosky, S. F. (1971). Speeded classification of multidimensional stimuli. *Journal of Experimental Psychology, 87*, 383–388.

Colle, H. A., Amell, J. R., Ewry, M. E., & Jenkins, M. L. (1988). Capacity equivalence curves: A double trade-off curve method for equating task performance. *Human Factors, 30*, 645–656.

Colle, H. A., & Reid, G. B. (1997). A framework for mental workload research and applications using formal measurement theory. *International Journal of Cognitive Ergonomics, 1,* 303–313.

Friedman, A., & Polson, M. C. (1981). The hemispheres of an independent resource-system: Limited capacity processing and cerebral organization. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1031–1058.

Gopher, D., & Donchin, E. (1986). Workload – An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. 2. Cognitive processes and performance* (pp. 41/1–41/49). New York: Wiley.

Hancock, P. A., & Caird, J. K. (1993). Experimental evaluation of a model of mental workload. *Human Factors, 35*, 413–429.

Hendy, K. C., Liao, J., & Milgram, P. (1997). Combining time and intensity effects in assessing operator information-processing load. *Human Factors, 39*, 30–47.

Knowles, W. B. (1963). Operator loading tasks. *Human Factors, 5,* 155–161.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–234.

Norman, D. A., & Bobrow, D. J. (1975). On data limited and resources limited processes. *Cognitive Psychology, 7*, 44–64.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. 2. Cognitive processes and performance* (pp. 42/1–42/49). New York: Wiley.

Pavur, E. J., Jr., & Kausler, D. H. (1974). Rated graphics distinctiveness of consonants. *Behavior Research Methods and Instrumentation, 5*, 23–26.

Polson, M. C., Wickens, C. D., Klapp, S. T., & Colle, H. A. (1989). Human interactive informational processes. In P. A. Hancock & M. H. Chignell (Eds.), *Intelligent interfaces: Theory, research and design* (pp. 129–164). Amsterdam: Elsevier.

Posner, M. I. (1969). Abstraction and the process of recognition. In G. B. Bower (Ed.), *The psychology of learning and motivation III* (pp. 43–100). New York: Academic.

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Constable, R. T., Skudlarski, P., Fulbright, R. K., Bronen, R. A., Shankweiler, D. P., Katz, L., Fletcher, J. M., & Gore, J. C. (1996). Cerebral organization of component processes in reading. *Brain, 119,* 1221–1238.

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

Schweichert, R., & Boggs, G. J. (1984). Models of central capacity and concurrency. *Journal of Mathematical Psychology, 28*, 223–281.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Shingledecker, C. A. (1984). *A task battery for applied human performance assessment research* (Tech. Report AFAMRL-TR-84-071). Wright-Patterson Air Force Base, OH: Armstrong Aerospace Medical Research Laboratory.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Columbia University, Teachers College, Bureau of Publication.

Timm, N. H. (1975). *Multivariate analysis with application in education and psychology.* Monterey, CA: Brooks/Cole.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 199–239). Mahwah, NJ: Erlbaum.

Verwey, W. B., & Veltman, H. A. (1996). Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of Experimental Psychology: Applied, 2,* 270–285.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson & R. Pew (Eds.), *Attention and performance VIII* (pp. 239–257). Mahwah, NJ: Erlbaum.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35,* 263–281.

Herbert A. Colle is a faculty member in the Department of Psychology at Wright State University in Dayton, Ohio and is also a visiting scientist at the U.S. Air Force Research Laboratory at Wright-Patterson Air Force Base, Ohio. He received a Ph.D. in psychology from the University of Washington in 1969.

Gary B. Reid is an engineering research psychologist at the U.S. Air Force Research Laboratory at Wright-Patterson Air Force Base, Ohio. He received an M.A. in educational technology from Arizona State University in 1974.