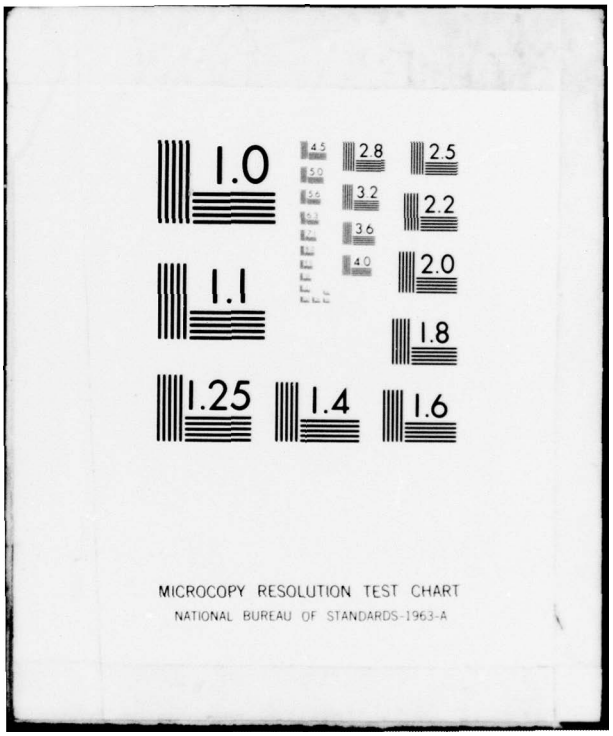
AD-A078 132 WRIGHT STATE UNIV DAYTON OHIO DEPT OF PSYCHOLOGY F/G 5/10
A CAPACITY-THEORETIC APPROACH TO WORKLOAD ASSESSMENT.(U)
AUG 79 H A COLLE AFOSR-78-3595
UNCLASSIFIED WSU-AFOSR-TR-79-1 AFOSR-TR-79-1267 NL

1 OF 1
AD
A078132

END
DATE
FILMED
1-80
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LEVEL

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR-79-1267	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): A Capacity-theoretic Approach to Workload Assessment	5. TYPE OF REPORT & PERIOD COVERED: Final Report. May 1978 — August, 1979.
6. PERFORMING ORG. REPORT NUMBER: WSU-AFOSR-TR-79-1	
7. AUTHOR(s): Herbert A. Colle	8. CONTRACT OR GRANT NUMBER(s): AFOSR — 78-3595
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, Wright State University, Dayton, Ohio 45435	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, Project Task 2313/09
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research, AFOSR/NL, Bolling, AFB, D. C. 20332	12. REPORT DATE: August, 1979
	13. NUMBER OF PAGES: 38
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

[redacted]
Approved for public release;
Distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DDC RECEIVED DEC 12 1979 A

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Workload
Attention
Dual task
Secondary task
Timesharing
Measurement theory
Human Performance

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Secondary task measurement of mental workload was explored both theoretically and empirically. By analyzing sets of secondary tasks over broad performance ranges it was found that an additive scale of workload was feasible. Therefore rational and empirical bases exist for using secondary tasks to measure the workload during flying tasks or other complex skills. A formal measurement theory for testing additivity and a scheduling theory of workload were described.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD A078132

DDC FILE COPY

408872

AFOSR-TR- 79-1267

A Capacity-Theoretic Approach to Workload Assessment

Herbert A. Colle

Department of Psychology
Wright State University
Dayton, Ohio 45435

AFOSR-78-3595

Acknowledgements

I thank Thomas Eggemeier and Joseph DeMaio for their comments on previous versions of this report. I thank William Acton, Kimball Curry and Frances Rubick for their assistance with data collection and analyses. Special Thanks go to Ronald Spicuzza, Systems Research Laboratory, who made the Computalker tapes for me.

The assessment of mental workload is a problem which, if solved, would be important both theoretically and practically. As a practical problem a need exists to develop standards that specify human cognitive or information processing limitations (e.g., Rohmert, 1971). Similar standards are already available for perceptual capabilities and for the physical arrangement and utilization of space (e.g. Kryter, 1970; Parker & West, 1973; Van Cott & Kincaid, 1972). With workload standards man-machine systems and their operating procedures could be designed for more optional performance.

Modern systems have a tremendous complexity and capability and therefore there exists a real possibility that present technology overloads pilots and others in command and control situations. A weapons director, for example, in the airborne warning and control system (AWACS) faces an enormously complex task. Helm (1978) has found that in technologically similar systems the tactical edge rests with the operator who can employ his weapon system in a maximum fashion.

Of course, a number of steps have been taken to aleviate workload problems. Heads up displays and multifunction switching are being utilized to simplify monitoring tasks and computers are used to augment other operator tasks such as in airborne decision-aiding.

Nevertheless, there exists a need to adequately assess the workload demands of these new systems as well as other potential systems to determine the extent to which they help overcome human operator limitations. In addition, providing workload standards would foster the rapid development of new technology by making the design stage more efficient and efficacious.

Also, workload assessment is important for training. Workload appears to decrease with training (Logan, 1979). Thus, an assessment technique would be useful to determine if a pilot has sufficient excess capacity to

meet the expected range of operational demands, or to determine if sufficient training had been given to introduce more demanding tasks safely. This is particularly important in high risk training situations. For example, because A-10 pilots train without an instructor pilot in the aircraft, advancement to more demanding tasks, such as low level flight or weapons delivery in low level flight, critically depends upon not exceeding their workload capacity.

The assessment of the workload that is utilized to perform a task is not straightforward. Performance on the task may not reveal the underlying effort that must be expended to produce that performance. For example, performance on recently acquired skills requires more effort than it does after automatization following extended practice (La Berge & Samuels, 1974; Logan, 1979; Spelke, Hirst & Neisser, 1976).

The *secondary task paradigm* has been used to deal with situations where primary task performance might inadequately reflect the workload. Experimentally, subjects are asked to perform a secondary task during the time that they are performing a primary task. Often instructions request that subjects perform the secondary task without interfering with the primary task. Conceptually, the general notion underlying the use of secondary tasks as measures of workload is simple. Performance on the secondary task is taken as an index of the amount of spare capacity not used by the primary task. Thus, secondary performance should increase when spare capacity increases and decrease when it decreases.

Secondary task paradigms have become increasingly popular with the increasing utilization of capacity theories of attentional limitations (Kahneman, 1973; Norman & Bobrow, 1975). Much of the research has been directed at testing specific processing theories, but an increasing amount

has been directed at the development of practical workload scales (Brown & Poulton, 1961; Hicks & Wierwille, 1979; Kalsbeek, 1968; Kalsbeek & Sykes, 1967; Kerr, 1973; Michon, 1964, 1966; Wierwille & Williges, 1978). Unfortunately, while these practical workload scales make theoretical assumptions they are not tested to determine the validity of the measure. To date, only properties such as sensitivity and monotonicity have been evaluated for practical workload measures (Michon, 1964, 1966; Hicks & Wierwille, 1979). An approach to relating practical measures to theory is described in the next section.

Measurement Theory: A titration model

In the present paper a measurement theory approach is taken to the development of workload scales. A measurement theory approach appears desirable because it can be used to develop practical scales with stronger properties while at the same time it constrains processing models of information processing limitations. To be testable, processing models must make very explicit assumptions about the type of processes that exist and how they interact. As a class they will be called microtheories. A measurement theory approach is a macrotheory.

First, the typical approach to using secondary task measures will be formalized. Let (A, w) denote that task A was performed yielding performance w on a dependent variable. Task B, C and D would be denoted similarly as (B, x), (C, y) and (D, z). Let Θ denote the act of performing two tasks concurrently, so that (A, w) Θ (C, x) denotes that task A was performed together with task C, yielding performances w and x respectively. The measurement model assumes that if

$$(A, w)\ \Theta\ (C, y) \text{ and if}$$

$$(B, x)\ \Theta\ (C, y') \text{ then}$$

$$(A, w)\ \angle\ (B, x) \text{ IFF } y > y',$$

where $\angle$ is an order relation indicating task A's workload is less than task B's at the given performance levels. Task C is the secondary task. It is used to order A and B because y and y' are in the same units and, therefore, can be ordered. Generally, w, x, and y are in different units and cannot be compared.

Typically the model is tested by assuming that workload within a single task increases monotonically with performance. Therefore primary task performance is an inverse monotone function of secondary task performance. In terms of the above formalism, if

$$(A, w)\ \theta\ (C, y) \text{ and if}$$

$$(A, w')\ \theta\ (C, y') \text{ then}$$

$$w > w' \text{ IFF } y < y'.$$

Thus Task C's performance should improve as Task A's declines.

Monotonicity has been tested several times (Hicks & Wierwille, 1979; Michon, 1964, 1966). It should be a basic requirement of any workload measure. It was to insure that this condition was being met in an experimentally interesting way that led Norman and Bobrow (1975) to develop the concept of a performance operating curve (POC). POCs that exhibit a trading relationship satisfy the above monotonicity condition. Unfortunately, there is considerable confusion in the literature about the role of POCs. They will be discussed more fully later.

The monotonicity test is a starting point for the development of a scale, but by itself it is weak. Although workloads can be ordered, the measure does not indicate whether each step in the order is a small or large one. Fortunately, methods exist for developing more powerful scales, if the operation θ can be considered as a concatenation operation.

If an equivalence relation could be devised to determine when two workloads are equal, then extensive measurement could be possible if the basic assumption of additivity were met (Campbell, 1957; Suppes & Zinnes, 1963). Additive measures based upon a concatenation operation exist in only a few areas of psychology, mainly sensory psychophysics. The two most familiar are metameric color matching (Krantz, 1974) and binaural loudness summation (Marks, 1979). Unlike these cases, the present development attempts to scale response performance, not stimulus intensity.

A matching relation can be determined for the secondary task paradigm. The secondary task thus can be used to establish equivalence classes. Assume that the monotone trading relation holds between task A and C and between task B and C. Now when the performance on C for the A - C pair equals the performance on C for the B- C pair then the task A and B performances produce an equivalent workload. As before this can be formalized. If

$$(A, w) \ominus (C, y) \text{ and if}$$

$$(B, x) \ominus (C, y') \text{ then}$$

$$(A, w) \ominus (C, y) \sim (B, x) \ominus (C, y') \text{ IFF } y' = y.$$

An additivity assumption in this context would produce:

$$(A, w) \sim (B, x) \text{ IFF } (A, w) \ominus (C, y) \sim (B, x) \ominus (C, y).$$

Thus, the equivalence of tasks A and B follows directly.

Although additivity cannot be checked directly because tasks A and B cannot be matched singly, a strong test of the additivity assumption follows directly from it. The equivalence between (A, w) and (B, x) should be found when a different secondary task D is used. That is, there exists a (D, z) such that:

$$(A, w) \ominus (D, z) \sim (B, x) \ominus (D, z).$$

A physical analogy for this measurement theory can be contrived. Assume that the volume of a number of odd-shape vials is to be measured. The vials can be filled with water, but because of their openings, as unknown amount of water is lost. However, water can be poured from them without loss. Unfortunately, the only available container they can be poured into is an opague jug so that only when the water level is at the brim is the water level measurable. In this situation two vials, A and B, can be equated in volume if another vial, C, can be found such that water from A plus water from C just reaches the brim and water from B plus water from C just reaches the brim.

When (A, w) and (B, x) are equated in workload over a broad performance range, a curve describing the relationship will be described. This curve will be called a performance equivalence curve (PEC). It's use will be described more fully below.

The macrotheoretic measurement system described above could be developed further in two different directions. (a) One direction would be to describe the operations that are necessary to recover the workload scale if additivity is found. These operations are straightforward and should await comprehensive testing of the additivity assumption. (b) A second direction would be to explore measurement theories for multidimensional workloads. Metameric color matching shows what these would be like (Krantz, 1974). Once again, this development would only be taken if the additivity assumption failed badly. Alternate empirical strategies, such as limiting the set of tasks to subsets within which additivity is met also, could be tried.

The next section will apply the measurement theory approach, developed above, to Norman and Bobrow's (1975) fixed capacity theory, a specific microtheory with strong and straightforward assumptions. The fixed capacity theory is a good illustration of the predictions made by the macrotheory; nevertheless, the fixed capacity theory and the macrotheory are not identical. Although the fixed capacity theory is consistant with the macrotheory, other microtheories also are consistant with it. For example, both Kahneman's (1973) variable capacity theory and Welford's (1952) single channel theory are consistant with the macrotheory. Each of these microtheories put restrictions upon the empirical tests of the macrotheory. The macrotheory claims only that empirical measurement procedures can be devised to satisfy the assumptions. It does not state what the procedures are. In fact, although these microtheories define capacity or workload restrictions that operate instantaneously, the macrotheory may only hold for average workload. In the experiment reported below an average workload rather than an instantaneous one was used.

Fixed Capacity Theory

Norman and Bobrow's (1975) fixed capacity theory is consistant with the formal measurement theory. The implications of this macrotheory can be clarified within the context of the microtheory because it lends itself to a simple mathematical description. The fixed capacity theory assumes that each person has a fixed amount of processing resources, r_m, which can be distributed arbitrarily to the tasks that are to be performed. It is assumed that the allocation of processing resources, r, to task one allows it to be executed at some performance level, P_1. P_1 is assumed to be a weakly monotonic function of r. Likewise, if task two is performed, P_2 should be a weakly monotonic function of the resources devoted to it, r_2. If the two tasks are performed together then the resources needed to attain the previous performances would

be $r_1 + r_2$. In general, $r_1 + r_2 \leq r_m$. Two tasks will be called capacity-compatible if the sum of the resources demanded by task one and those demanded by task two is less than the maximum available resources, that is $r_1 + r_2 < r_m$. If $r_1 + r_2 > r_m$. Then both tasks cannot be performed together at the same performance levels that they were performed individually. Thus, one task's performance must be degraded so that $r_2 = r_m - r_1$.

A performance operating curve (POC) describes the trade-off in performance that occurs when $r_2 = r_m - r_1$ and r_1 is varied. No tradeoff would be expected if two tasks are capacity-compatible over their entire performance range. Under these conditions performance is said to be data-limited rather than resource-limited. Performance is thought to be limited because the incoming data are insufficient for better performance, regardless of the attentional resources devoted to the task.

Performance operating curves can be described as follows: Let P_i denote performance on task i and let

$$P_1 = f(r)$$

$$P_2 = g(r)$$

$$P_3 = u(r)$$

$$P_4 = w(r)$$

when f, g, u, and w are arbitrary monotonic functions on r, the resources devoted to that task. If tasks one and three are performed together and if task performance is in a resource-limited range, then $r_1 = r_m - r_3$. Therefore, the performance operating curve can be specified by the following parametric equation in this range:

$$(1) \qquad P_1 = f(r_m - r),\ P_3 = u(r).$$

Similarly, the POC for tasks two and three is given by:

$$(2) \qquad P_2 = g(r_m - r),\ P_3 = u(r),$$

where r denotes the resources devoted to task three.

By varying the value of r, a POC is swept out for the task one and three pair and another one for the task two and three pair. By algebraic manipulation, these POCs also are given by:

$$P_1 = f(r_m - u^{-1}(P_3)) \text{ and} \quad (3)$$

$$P_2 = g(r_m - u^{-1}(P_3)). \quad (4)$$

POCs can be used to derive a more important fundamental relationship, the performance equivalence curve (PEC). Figure 1 shows two POCs. Task 1 was paired with task 3 to produce one POC and task 2 was paired with task 3 to produce the other POC. The vertical line shows a point on the abscissa where $P_3 = u(r)$ or $r = u^{-1}(P_3)$. Therefore, both the P_1 and the P_2 performances at that point are produced with the same resources, $r_m - r$. These two performance points are points that satisfy the macrotheory's measurement procedure. They both match by just using up all remaining resources when task 3 at the given performance level is added to each one. Similar equivalent points can be found for each point on the axis.

A performance equivalence curve is a curve relating task 1 to task 2 performance. It specifies the performance on task one, P_1, that could be expected given that the same resources were devoted to it as were devoted to task two in order to produce performance, P_2. Mathematically, a PEC can be derived from the two POCs in Eq.(1) and (2) or (3) and (4). By algebraic manipulation, the PEC is given by:

$$P_1 = f(r_m - r),\ P_2 = g(r_m - r) \quad (5)$$

in parametric form, or by:

$$P_1 = f(g^{-1}(P_2)). \quad (6)$$

Although a PEC shows the equivalence classes that result from following the macrotheory's measurement procedure, the additivity assumption is not tested. The macrotheory's additivity test is implemented by determining the invariance of Eq.(5) when it is generated by using a new task, four, to generate

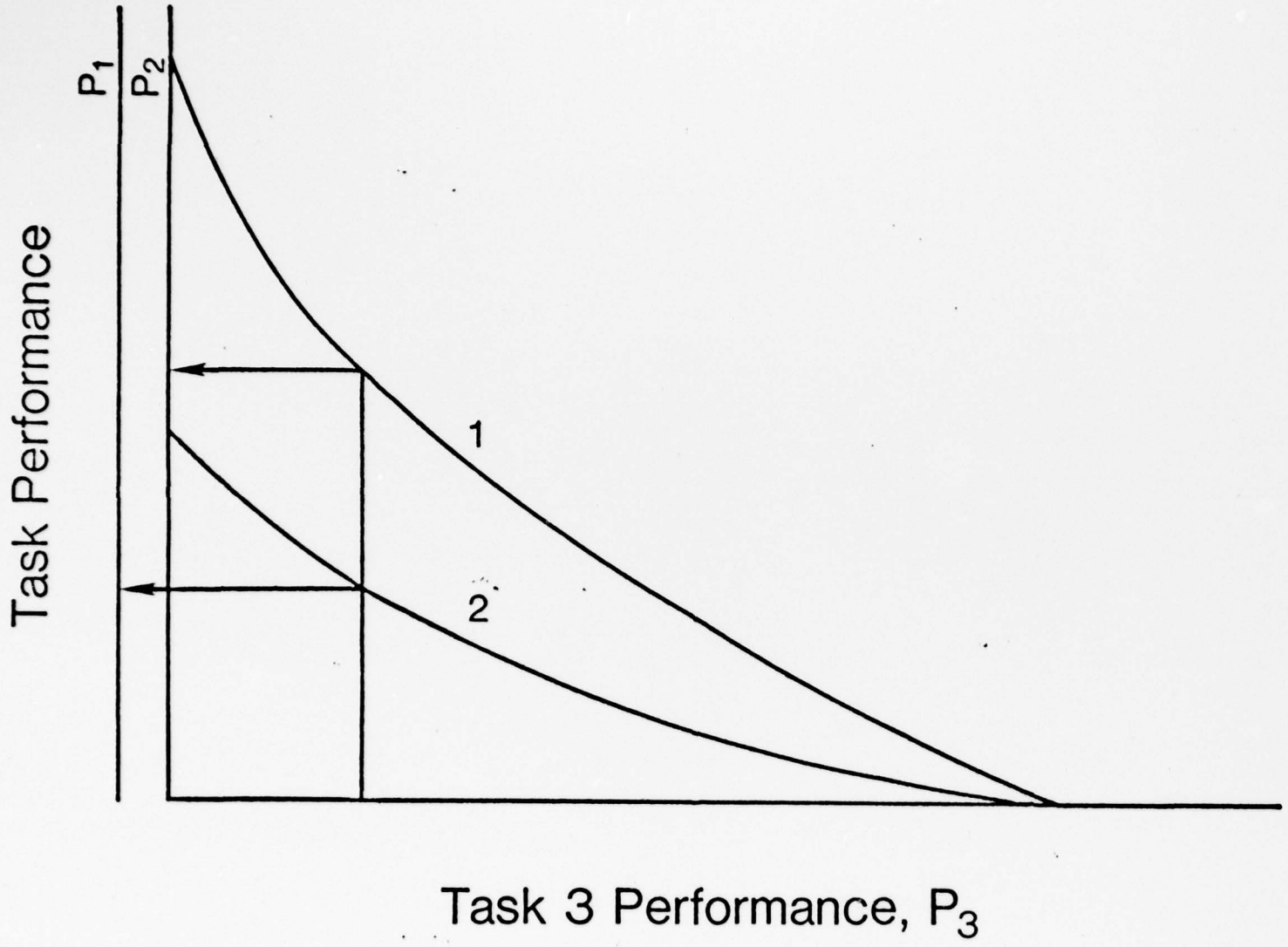


Figure 1. Hypothetical performance operating curves illustrating how a performance equivalence point can be obtained.

the PEC. It can be shown that Eq.(5) again should be obtained. Once again, the POCs would be:

$$(7) \qquad P_1 = f(r_m - r),\ P_4 = w(r) \text{ and}$$

$$(8) \qquad P_2 = g(r_m - r),\ P_4 = w(r).$$

The PEC from Eq.(7) and (8) obtained as before is:

$$(9) \qquad P_1 = f(r_m - r),\ P_2 = g(r_m - r).$$

Notice that Eq.(5) and Eq.(9) are identical.

Therefore, the fixed capacity theory is testable because it implies that the PEC for tasks one and two that is obtained by using task three as a co-task should be the same as the PEC that is obtained by using task four as a co-task. If the resources used for some level of performance on task two can be specified in terms of an equivalent level of performance on task one, then this relationship should be independent of the arbitrary task that is paired with the tasks. If the relationship depends heavily on the nature of the third task, then the additivity assumption of the fixed capacity theory is violated.

Performance Equivalence Curves

The macrotheory helps to put a previous controversy over POCs (Norman & Bobrow, 1975, 1976; Kantowitz & Knight, 1976; Navon & Gopher, 1979) into proper perspective. Conflict arose over the conditions necessary for generating a POC. Kantowitz & Knight (1976) have taken the extreme liberal position of "lets wait and see." Navon and Gopher (1979) have taken an extremely restrictive position that "there is only one way." Norman and Bobrow (1976) appear to be moderates, rejecting the position that all conditions are candidates but allowing conditions that Navon and Gopher proscribe.

It is difficult to resolve this controversy because there has been no way to determine if a POC was influenced only by resource demands. The present

paper shows that POCs are of secondary interest to PECs and that PECs provide a mechanism that can be used to check the appropriateness of resource manipulations, the additivity test. In addition, PECs should be much easier to generate than pure POCs. To generate a PEC the resources used by the co-task to generate performance P_3 must be manipulated. These manipulations, however, could include some that affect data quality, besides affecting resources. Valid PECs would still be generated because matching is done at the same fixed P_3 level for both P_1 and P_2. For example, both Navon and Gopher (1979) and Norman and Bobrow (1975) have argued that Kantowitz and Knight (1976) did not generate a POC when they varied the difficulty of their tapping task. Nevertheless, a PEC can be generated if it is assumed that at least some resources were varied, because each of the tapping tasks was paired with two different digit transformation tasks, adding three to digits (DT+3) or shadowing digits (DTO).

The circles in Figure 2 show estimated PEC points that can be generated. The solid curve is the best-fitting power function. This PEC can be compared with the PEC between DT+3 and DTO that was generated by using a two-tone classification task (Colle & DeMaio, 1978). Although Colle & DeMaio's study was only a pilot study not designed to test additivity, the results were suggestive. The dotted curve shows the best-fitting power function for their PEC. No data points are shown because the PEC had to be derived from POCs. The experiment was not designed to yield a PEC directly. The curves are drawn to represent the range of variation in Colle & Demaio's POCs. The two PECs are surprizingly similar given that the two experiments differed substantially both in the estimation procedures that were used and in the co-task that was used to generate the PECs.

POCs can be useful. They can be used to help avoid capacity-compatible task combinations from which PECs cannot be generated. They also can be used to

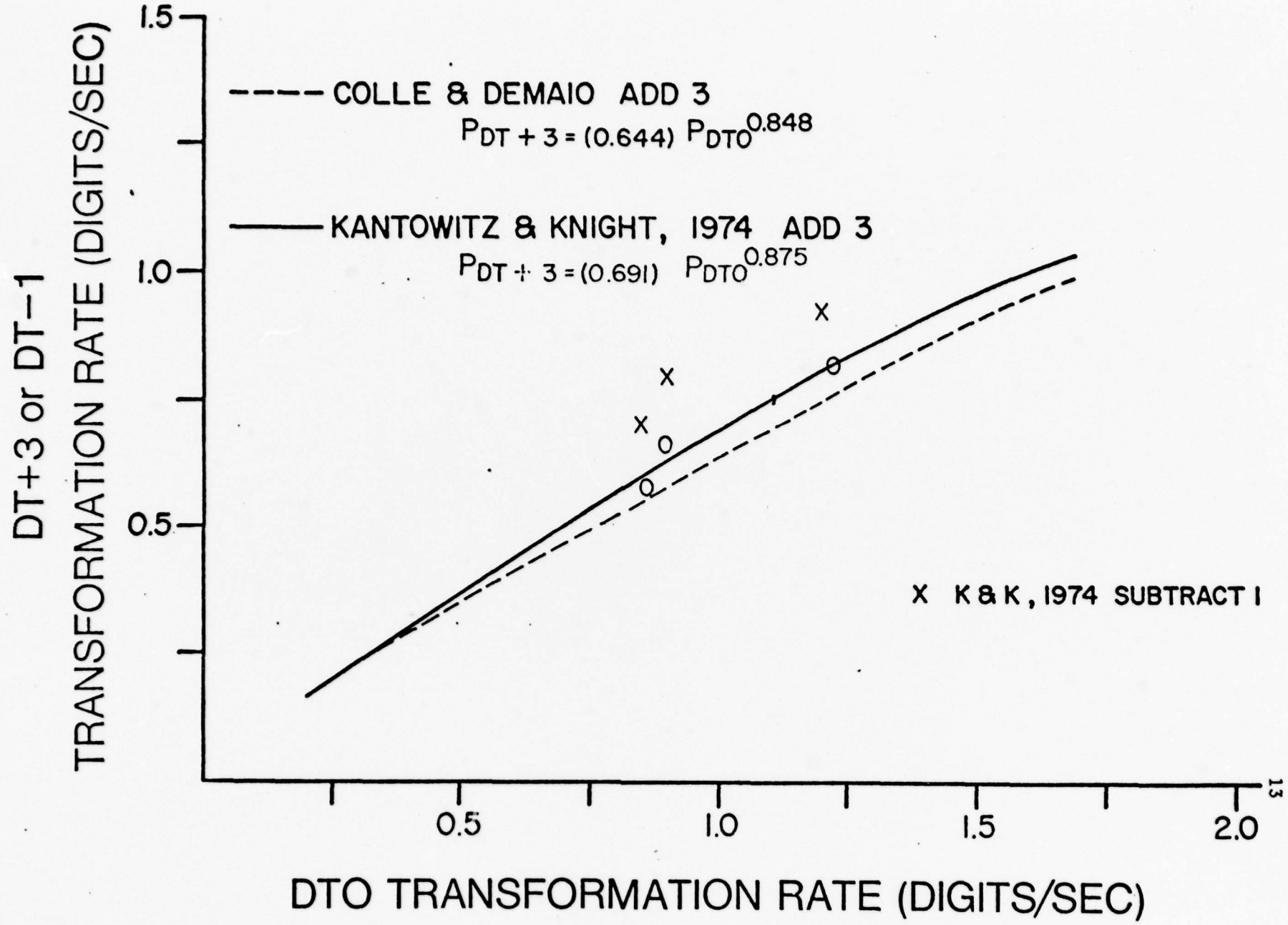


Figure 2. Performance equivalence curves for the same tasks from two different studies.

test the processing assumptions of some microtheories. Nevertheless, it is clear that a concern for POCs should be relegated to a secondary position to a concern for PECs.

An Empirical Test of Additivity

The experiment described below was designed to test the additivity assumption. Two digit transformation tasks each were paired with two memory classification tasks. Subjects were given two minute trials during which a digit transformation task was performed together with a memory classification task. On each trial the rate at which digits had to be transformed was fixed and the rate at which the memory classification task could be performed was determined.

The two digit transformation tasks were Add one (DT+1) and Add three (DT+3). The memory classification tasks were continuous versions of the discrete Sternberg paradigm (Nickerson, 1973). One used a memory set of size two and the other one used a memory set of size four. A PEC between the two memory classification tasks was estimated using DT+1 as a co-task; another PEC was estimated using DT+3 as a co-task. The two PECs should be identical if additivity holds.

Method

Subjects

The subjects were 16 males between the ages of 18 and 30 who were paid $2.90 per hour for their participation. They were recruited from the campus of Wright State University.

Apparatus and Task Descriptions

Digit Transformation Tasks. Two different digit transformation tasks were used. In the digit add one task (DT+1) the subject was asked to add one to the digit presented and to say aloud the name of the sum. In the digit add three task (DT+3) the subject was asked to add three to the digit presented and to report the sum.

Only the digits 1-6 were presented and each had an equal probability of occuring. Blocks of digits were recorded at the rates of 0.25, 0.50, 0.70, .80, and .95 digits/sec for the DT+1 task and 0.15, 0.30, 0.50, 0.65 and 0.80 digits/sec for the DT+3 task. To increase list homogeneity, the random device generating each list was constrained so that no digit repetitions occured and so that during the 2-min test period each digit was used exactly the same number of times in each list. This was accomplished by sampling without replacement from a finite population consisting of an equal number of exemplars of each digit, unless a repetition occurred. A repeated digit was returned to the sample and another sample was taken until a different digit was selected. For the DT+1 lists there were 5, 10, 15, 16, and 19 exemplars of each digit in the population for each of the respective rates listed above. The corresponding frequencies for the DT+3 lists were 3, 6, 10, 13, and 16 exemplars, respectively. The presentation rates were selected so that each digit could be presented an equal number of times in each list and so that the presentation rates were about equally spaced across the entire performance range from the maximum possible rate to zero (Colle & DeMaio, 1977).

Two sets of lists were produced for each DT task. Each set consisted of two separate blocks of trials and each block consisted of one list at each of the five presentation rates. Each set was used on one of the two test days for each DT task in counterbalanced order across all subjects. The lists were re-recorded to vary the order of testing the DT rates in each set across subjects. These testing orders are described in the procedure section. In addition to the test trial stimuli, enough additional stimuli were recorded for the practice trials and warmup periods so that each stimulus list was presented only once to a subject.

The digits were recorded by using a Computalker speech synthesizer to present the digits at the appropriate rates. The tapes were played to the subjects by a TEAC 7090 GSL tape deck which was under computer control. Its output was amplified and sent to the right earphone of a headset (TDH39 earphone mounted in a MX41/AR cushion). The subject's responses were monitored by a microphone mounted within six inches of his mouth. Timing tones which coincided with the onset of the presented digit were generated to assist the scoring. If two response onsets occurred the first one was scored, that is "false starts" were scored as incorrect. All testing was conducted in a single-walled IAC chamber.

Memory Classification Tasks

Sternberg varied-set memory classification tasks using either memory set sizes of two (MC2) or four (MC4) were used. Ten geometric shapes were used as stimuli (circle, ellipse, half-circle, heart, triangle, diamond, trapezoid, square, rectangle, and parallelogram). The shapes were displayed visually on a Techtronics 604 monitor under computer control. The memory set was displayed twice prior to a trial and the test stimuli were presented successively until the 2-min trial terminated. The presentation rate was self-paced. The test

stimulus remained displayed until the subject responded by pressing a momentary contact switch to classify it as in the memory set or not. The next test stimulus was presented immediately after the response switch was released.

Each list of test stimuli were constrained by the following rules. (a) No repetitions were allowed. (b) Within sets of 24 stimuli, positive and negative instances occurred equally often. (c) Each member of the positive set was used equally often and each member of the negative set was used equally often in each set of 24 stimuli. The (b) and (c) constraints were accomplished, again, by sampling without replacement from a total set of 24 stimuli. The 24 stimuli lists were strung together to produce the complete list with enough test stimuli for the entire 2-min trial.

Twenty-four test stimuli lists and 24 memory sets were generated for the test trials of the MC2 task and another 24 test stimuli and memory sets were generated for the test trials of the MC4 task. Each list was used only once by a subject. In order to make the memory sets more comparable across trials, restrictions were placed on their composition. The stimulus set was divided into three subsets (circle, ellipse, half-circle, heart), (triangle, diamond, trapezoid), and (square, rectangle, parallelogram). MC2 memory sets consisted of no more than one stimulus from each subset and MC4 memory sets consisted of at least one from each subset. In addition, the memory sets for each block of six trials were balanced so that each stimulus was used at least once but not more than twice in MC2 memory sets and that each stimulus was used at least twice but not more than three times in MC4 memory sets. The memory sets also were balanced across all 24 trials so that each stimulus was used at least four times but not more than five times in MC2 memory sets and that each stimulus was used at least nine times, but not more than ten times in MC4 memory sets.

Varied-set classification tasks were chosen because there is evidence that performance is stable even with extensive practice (Schneider & Shiffrin, 1976; Shiffrin & Schneider, 1976). Also, increasing the memory set size from two to four reliably changes the time to respond (Nickerson, 1973), without, perhaps, changing the type of processing utilized (Sternberg, 1969). Checkosky (1971) previously has used geometric stimuli in the Sternberg paradigm.

Procedure

Each subject was tested for five different sessions. During session 1 each of the four tasks were practiced individually for two trials. In addition, each of the four pairs of tasks were practiced for three trials for a total of 20 trials. Each subject received the DT+1 task at the 0.70 and 0.8 rates and the DT+3 task at the 0.50 and 0.80 rates. During the MC2 and MC4 tasks, subjects were encouraged to respond as fast as possible while keeping the error rate low. For the paired tasks the DT+1 was presented at the 0.25 and 0.8 rates for the DT+1 - MC4 pairs and at the 0.50 and 0.95 rates for the DT+1 - MC2 pairs. The rate of presentation for the DT+3 task was 0.15 and 0.80 when it was paired with MC4 and 0.30 and 0.65 when it was paired with MC2.

Sessions two through five were test sessions. During each session the subject had to perform one of the digit transformation tasks in combination with the memory classification tasks. Half of the subjects received the DT tasks in each counterbalanced order across the four sessions (either DT+1, DT+3, DT+3, DT+1, or DT+3, DT+1, DT+1, DT+3). Each session consisted of two blocks of trials, a block with the DT task being paired with the MC2 task and a block with it paired with the MC4 task. Of the eight subjects receiving the DTO task in the first test session (session two), four were

given the MC2 task in the first block and the MC4 task in the second block and four were given the opposite order. The blocks were counterbalanced in the same way for the eight subjects receiving the DT+3 in the first test sessions. For each subject the order of testing MC2 and MC4 alternated across days so that their order was counterbalanced within subjects with respect to the two sessions during which the same DT task was used (days 2, 5 and 3, 4).

Each block of trials consisted of two 2-min practice trials using the pair of tasks that was tested on that block of trials plus six 2-min test trials in which the MC task was performed once singly and once with each of the five rates of presentation of the DT task. The testing order for the different DT task presentation rates (including the zero rate condition when MC was performed alone) was determined as follows. For the first block of trials in session two (the first test session) the DT presentation rates were presented in a different randomly determined order for each of the first eight subjects that were tested (four with DT+1 and four with DT+3). The orders were constrained so that across the subjects each of the six rates was tested at each position not more than twice. Each subject also used the same rank order of rates during the second block of trials when the other MC task was used. The testing order in session three was determined in the same way. In sessions four and five, however, the testing orders were the reverse of the order used in sessions three and two, respectively, but the other set of DT lists that was recorded was used. The eight subjects who were tested second used the same tapes and therefore the same order of DT rates as the first eight, except that the tapes that were used on days two and three were reversed and those used on days four and five were reversed.

The sequence of events on each trial was as follows. The subject initiated the start of a trial by pressing a starting switch. Following the words "memory set", each figure in the memory set was presented individually for 1.5 sec. After the whole set was presented twice the words "test beginning" followed. The first visual test figure was displayed at the same time as the first digit was played. The first 20 sec of the trial was discarded as warmup. Performance on the following two minutes was recorded. Subjects were told the DT task and memory set size before each block of trials started.

The subjects were told to perform the MC task as fast as they could while minimizing errors. If a subject made more than 10% error on either task, they were given a makeup trial with the same rate and conditions at the end of the trial block for those conditions and before a new task combination was used.

Results

Performance Equivalence Curves

The number of correct responses that were made during each two minute test period was transformed into responses per minute. The means for all 16 subjects were computed. If a subject made more than 10% error on either task, he was given a makeup trial at the end of the trial block and his makeup performance was used in the analysis. If, in addition, the subject made more than 10% error on either task on the makeup trial, then the score for that condition was treated as missing and dropped from the analysis. Alternate treatments of these trials, such as substituting the original response rate as the score for high error makeup trials so that there were no missing values, or dropping all makeup data so that there were more missing values, did not change the means or the analyses substantively.

Figure 3 presents the data that is needed to test additivity. Days 2 & 3 and days 4 & 5 were plotted separately. Each test day pair completely replicated the important experimental conditions. The two axes of Fig. 3 present the mean classification rates for the memory classification tasks. Each data point was generated by recording the classification rate that was achieved when each classification task was paired with one of the digit adding tasks at a given rate of presentation. The PEC generated by using the Add 1 task as a co-task (filled circles) should be the same as the PEC generated by using the Add 3 task as a co-task (open circles). As Fig. 3 shows, these two PECs are very similar in both replications. Furthermore, although overall performance improved substantially from Days 2 & 3 to Days 4 & 5, the data from both replications appear to lie along the same PEC curve.

As with receiver operating curves and performance operating curves, performance equivalence curves are bivariate and, therefore, devising statistical tests to test their equivalence is not straightforward. This is particularly

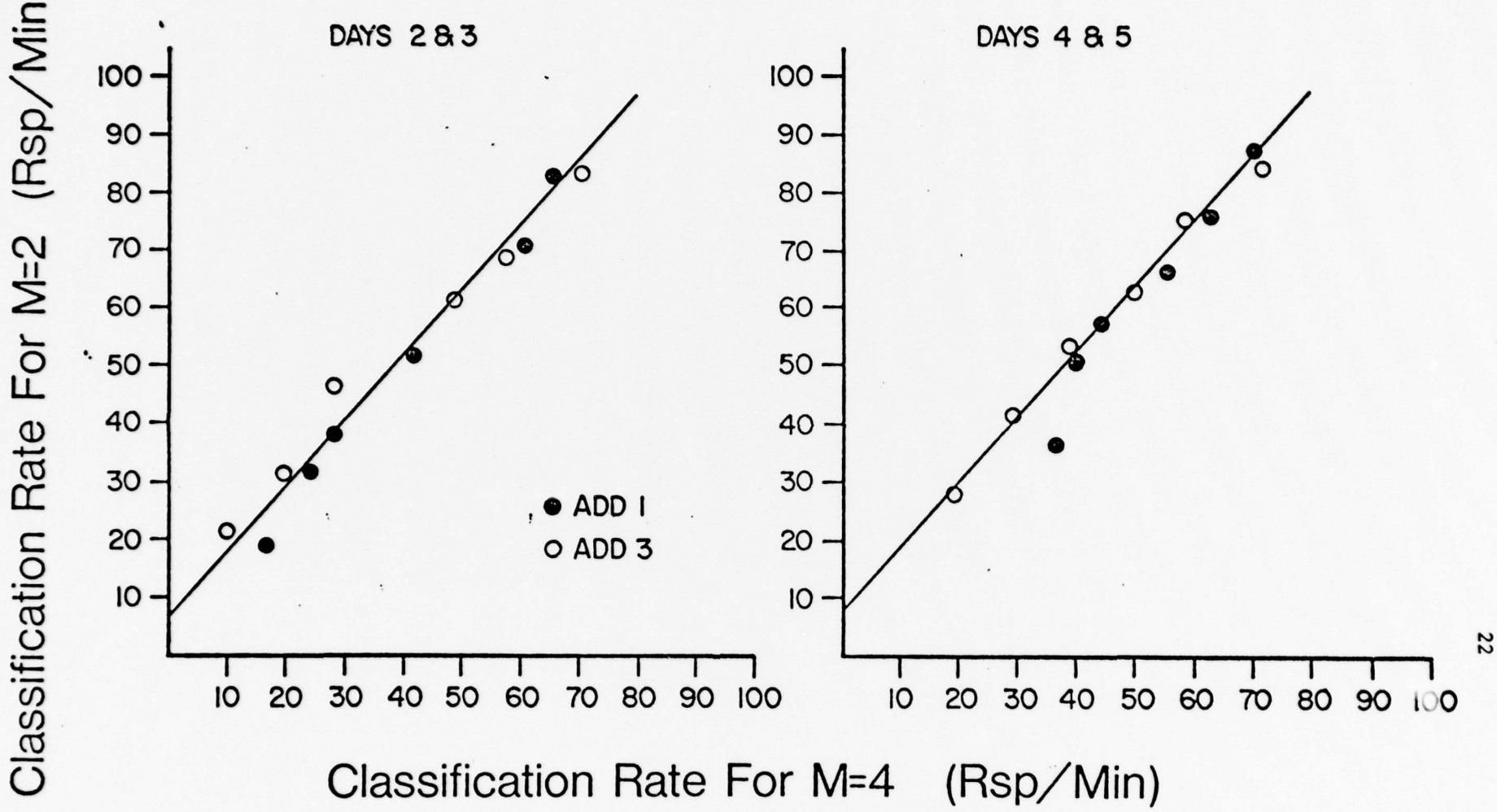


Figure 3. Performance equivalence curves (PECs) for two memory classification tasks (M = 2, 4) obtained by using different co-tasks (add 1, add 3).

true for PECs because the form of the function relating them is unknown. The additivity prediction holds for any function. In order to test the equivalence of the PECs, the following statistical test was planned. First, a single function was fit to all of the data. Secondly, the parameters were estimated separately for each of the two PECs. Finally, a multivariate analysis of variance (MANOVA) was used to test the equality of the parameters. Since both PECs were collected on the same subjects,the MANOVA was a within-subjects MANOVA (Timm, 1975).

A linear function provided a good fit to the PEC curve. The analysis was conducted on the means from Days 2-5, yielding six Add one and six Add three pairs of scores. The pearson correlation coefficient for these 12 points was 0.984. The best-fitting slope and intercept was 1.11 and 6.70, respectively. This straight line is presented in both panels of Fig. 3. The overall PEC appears to provide a good fit for both replications.

To statistically test the equivalence of the Add one versus the Add three PEC, best-fitting straight lines were obtained for each subject separately for the Add one and the Add three PECs. Differences between the slopes and intercepts of the Add one PECs and the slopes and intercepts of the Add three PECs were tested using the program Multivariance (Finn, 1976). The multivariate F was not statistically significant, $F(2, 14) < 1.0$, $p > .05$. Univariate tests on the slopes and intercepts separately also were not statistically significant, $F(1, 15) < 1.0$, $F(1, 15) = 1.12$, respectively. Analyses of Days 2 & 3 and Days 4 & 5 separately yielded similar results. The multivariate $F(2, 14)$ was 1.55 for Days 2 & 3 and was less than one for Days 4 & 5.

The feasibility of using a special case of the linear function to represent the PECs was investigated. The best least squares slope under the assumption that the intercept was zero was estimated as 1.24. This special case did not

fit as well as the general linear function. An analysis of variance was conducted on the intercept estimates from the general linear function, using days and DT tasks as factors. Although there were no differences between these factors or their interactions, the overall mean intercept was significantly greater than zero, $F(1, 15) = 15.8$, $p < .01$.

By estimating the zero-intercept slopes for each subject separately for the Add one and Add three PECs, the equivalence of the PECs can be tested for this special case. Since there is only one dependent variable, slope, univariate statistics were used. A dependent t-test between the Add one and the Add three PECs was not statistically significant, $t(15) = 0.69$, $p > .05$. Analysis of each replication separately also revealed no significant differences.

Figure 4 presents the POCs that were generated. As you can see there was a trade-off between the jointly performed tasks. Not all of the subjects displayed this trade-off throughout the test. Figure 5 shows POCs from four subjects who were tested and were replaced by four others because their data did not exhibit a trade-off for the Add one task during the second replication. It should be noted that they did show evidence of trade-offs for the Add three task during the second replication and for both tasks during the first replication. They were replaced because the lack of a trade-off suggests that they might be capacity-compatible over their entire performance range, and thus, the additivity test would be inappropriate. In addition, using them would have added unnecessary variability to the planned statistical tests.

Error Analyses

Table 1 presents the percentage of errors that were made on each of the tasks for the important experimental conditions. Error differences would affect the additivity test on the PECs only if there was an interaction between the two digit tasks and the two memory classification tasks. Although

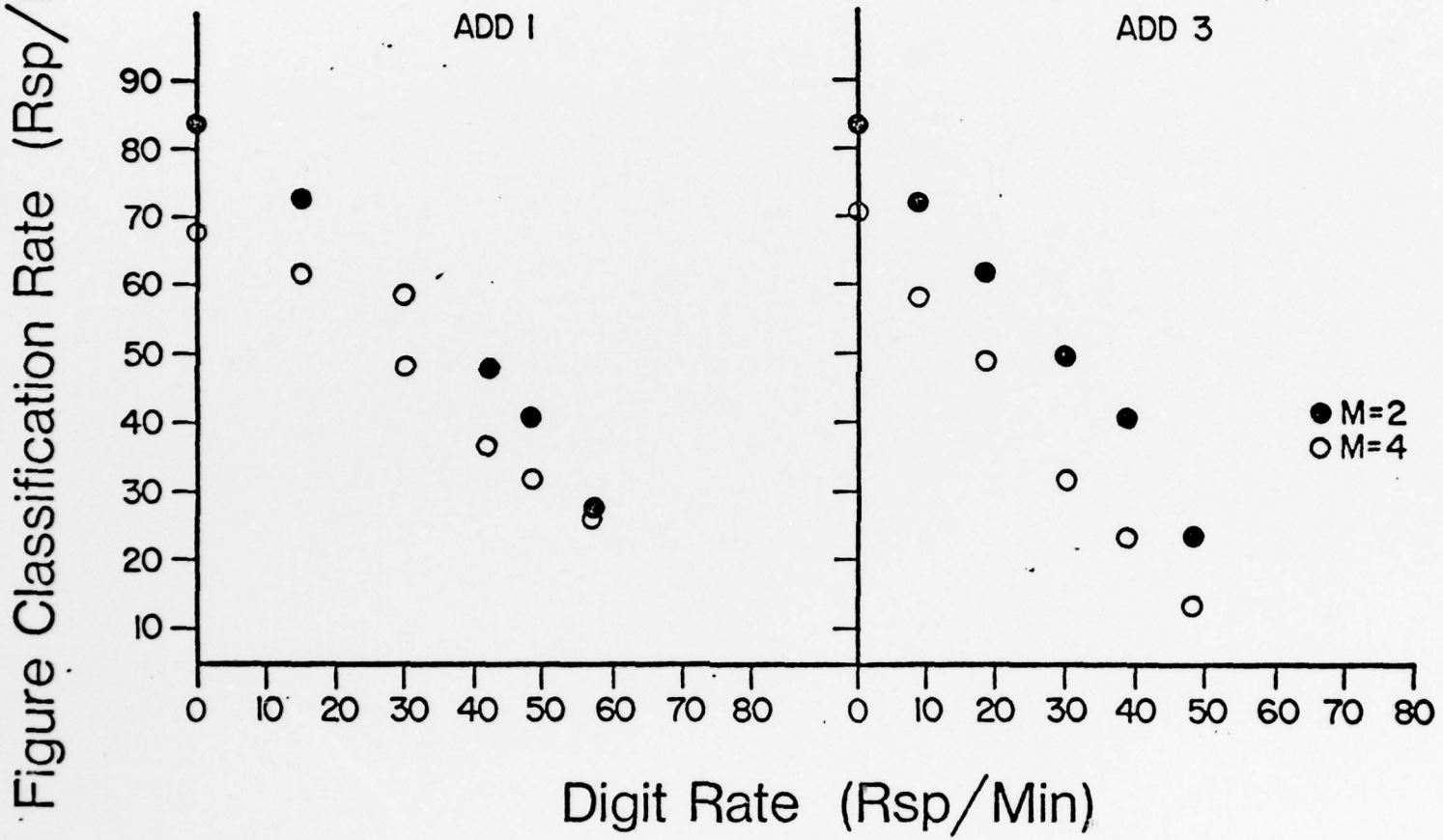


Figure 4. Performance operating curves (POCs), showing the trade-off between memory classification and digit transformation performance, averaged over days 2-5.

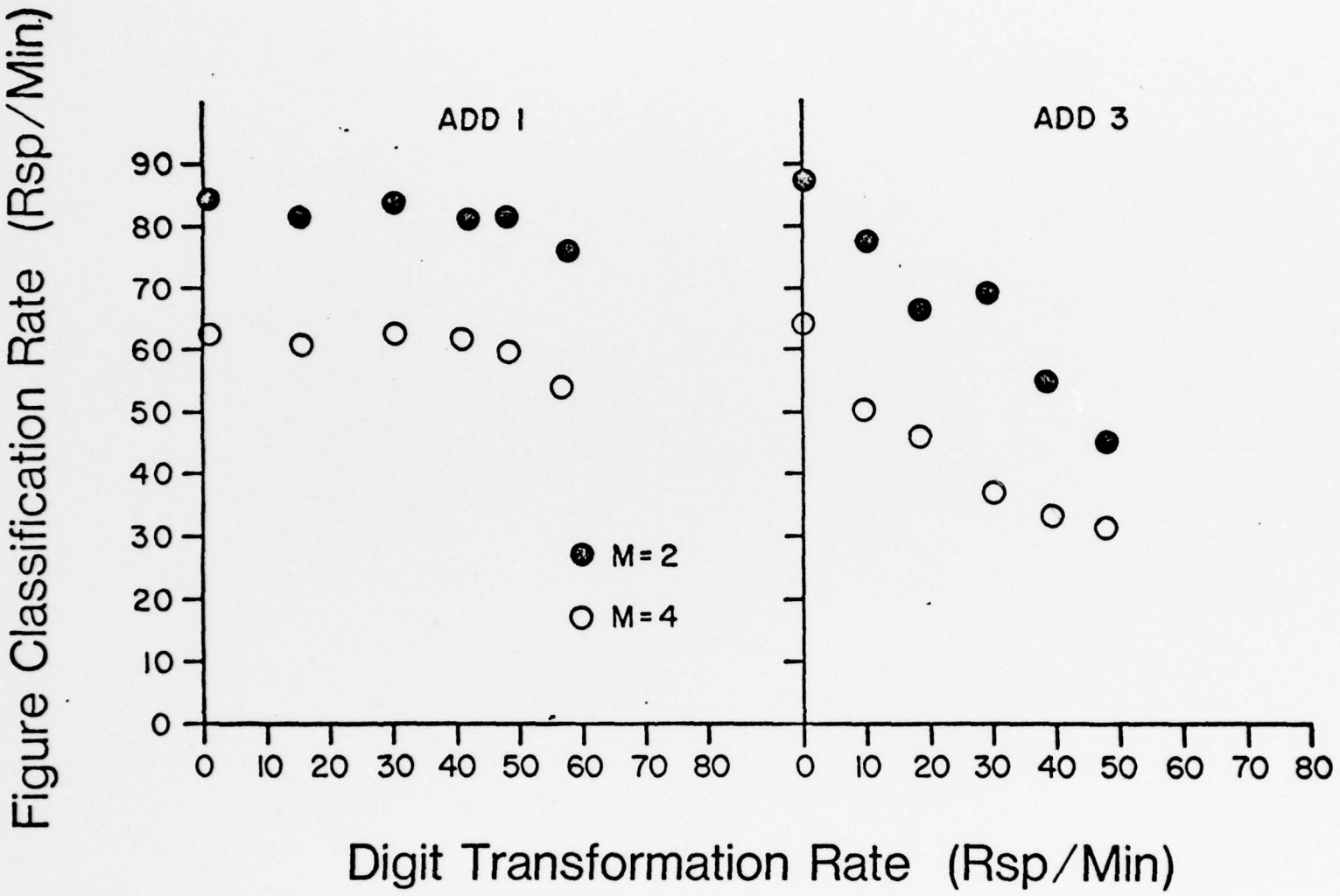


Figure 5. Performance operating curves (POCs) from days 4 & 5 for four subjects who were not used in the experiment. No trade-off appears in their Add 1 data.

a speed-accuracy trade-off could influence the form of the PEC function, the function was derived empirically. No assumptions about its form were made. The additivity test only assumes that the Add one and the Add three PECs can be described by the same function. No evidence for the interaction is present in the error data presented in Table 1, although errors appear to increase as the digit transformation rate increases.

Table 1

Percentage of Incorrect Responses on the Memory Classification Tasks and on the Digit Transformation Tasks

Test Conditions	DT Rate[a] 0	1	2	3	4	5	$\bar{x}$
Memory Classification Tasks							
Days 2 & 3							
Add 1							
m = 2	1.3	1.1	2.0	1.8	2.8	1.6	1.8
m = 4	2.1	1.0	1.8	1.6	3.0	1.3	1.8
Add 3							
m = 2	1.8	1.3	1.2	1.6	1.2	2.3	1.6
m = 4	1.9	1.8	2.3	4.4	4.8	2.2	2.9
Days 4 & 5							
Add 1							
m = 2	1.3	2.2	0.8	1.1	2.0	1.4	1.5
m = 4	1.6	2.1	2.1	2.9	2.4	1.8	2.2
Add 3							
m = 2	1.9	1.4	2.0	1.5	1.0	2.0	1.6
m = 4	1.2	1.7	1.9	1.7	2.1	2.1	1.8
Digit Transformation Tasks							
Days 2 & 3							
Add 1							
m = 2		0.9	1.2	4.1	4.4	5.6	3.1
m = 4		0.4	2.2	4.0	4.1	5.7	3.3
Add 3							
m = 2		0.7	0.9	2.2	3.3	5.8	2.6
m = 4		2.2	1.2	3.5	4.5	6.7	3.6
Days 4 & 5							
Add 1							
m = 2		0.9	1.4	2.1	3.7	4.8	2.6
m = 4		0.2	1.2	3.4	3.7	4.5	2.6
Add 3							
m = 2		0.8	1.5	2.9	3.8	5.2	2.8
m = 4		0.7	0.2	2.6	3.8	4.1	2.3

[a]Add 1 and Add 3 transformations had different rates. Numbers indicate the rank. The single task condition is indicated by 0.

Discussion

The results suggest that additive measures of workload can be constructed, for at least a limited subset of tasks. Both PECs were very similar over a wide range of performance. Although a somewhat different experiment is necessary to actually construct an additive scale, studies similar to the present one can be used to equate different types of secondary task measures and to test for the boundary conditions on the additivity assumption. It remains to be seen how robust additivity will be.

How can these results be reconciled with previous failures to find additivity (see Navon & Gopher, 1979)? Basically, previous studies have tested a less interesting type of additivity, although it may be important for some microtheories. Previous additivity failures were based upon discrepancies between single task performances and nearby dual task performance. Although these discontinuities do indicate non-additivity, they are not a serious problem for a workload measure. This type of non-additivity is analogous to the non-additivity of an amplifier that has insertion loss or gain. The amplifier may be linear over a considerable range, but the output with the amplifier set at unity gain may be different from the output that is obtained without the amplifier in the circuit. As long as the insertion loss or gain is taken into consideration, valid workload measures can be obtained.

Practically, if additivity is robust, it means that sets of secondary measures can be equated and their relationship to the underlying workload can be derived. In this way the workload of complex tasks can be measured with a battery of tests that have been equated. Thus an even more robust measure is obtainable because any deviation from equality during the

application of the battery suggests the presence of structural interference which must be taken into account. The additivity of the secondary measures can be checked directly in complex tasks, because pure resource changes are not needed to produce a PEC. Application of POCs in complex flying tasks appears complex, if not impossible. Getting pilots to degrade their flying performance for an arbitrary cognitive task is very difficult (Colle & DeMaio, 1978). In addition, if flying performance was degraded, what measure would be used in the POC? POCs do not lend themselves readily to multidimensional measures. Fortunately, the use of PECs bypasses these problems.

Although the present additivity test is constant with the fixed capacity theory, it is also consistant with other theories of attention. The test is a strong one for the theory to pass. On the other hand, a much stronger test could have been devised had a different testing procedure been used. In this regard, an important distinction should be made between instantaneous workload and average workload. This distinction can be made both theoretically and empirically. Although theories of attention are described in terms of instantaneous workload (Broadbent, 1958; Deutsch & Deutsch, 1963; Norman, 1968; Treisman & Getten, 1967), most tests of them have used measures that probably relate more to average workload because they average performance over a relatively long period of time. Some exceptions do exist (Kristofferson, 1967; LaBerge, 1973). The importance of this distinction can be brought out more clearly in a microtheory that incorporates it explicitly. All previous attentional theories are indistinguishable from special cases of this microtheory. This microtheory, the scheduling model, will be briefly outlined below.

The Scheduling Model: A Microtheory

According to the scheduling model, people have a considerable flexibility in how they use the processes that they have available. A number of investigators have commented on this flexibility (Atkinson & Shiffrin, 1968; Reitman, 1970; Spelke, et al., 1976; Sternberg, 1969) and the problems it produces. In the scheduling model, the flexibility of this executive system helps rather than hinders the analysis.

People can be described as follows: Upon being given a task, the Executive system considers various constraints such as those presented in Figure 6. In addition, the set of available subprocesses is considered. From these considerations a schedule of subprocess activity is constructed. It indicates which subprocesses will be used at each point in time. A sample subprocess schedule is shown in Figure 7.

As you can see, in order to accomplish any work a number of subprocesses that are distributed over time must be executed. The order of these subprocesses is constrained by the precedence and capacity constraints, etc. That some subprocesses must be performed before others is an assumption that has been made by almost all information processing theories.

The model assumes that each subprocess i requires some capacity, denoted C_i. At any given point in time, t, the instantaneous workload, C_t, or capacity used would be:

$$C_t = \sum_{i=1}^{K} C_{ti} \leq C_{max}$$

Most theories of attention are directed to limitations upon C_t. Nevertheless, the scheduling model assumes that it can fluctuate considerably over the time during which a task is performed.

In the face of such fluctuation an attempt to measure workload would be directed at the average workload, WL, which can be described as:

$$WL = \frac{1}{T}\int_0^T C_t \, d_t,$$

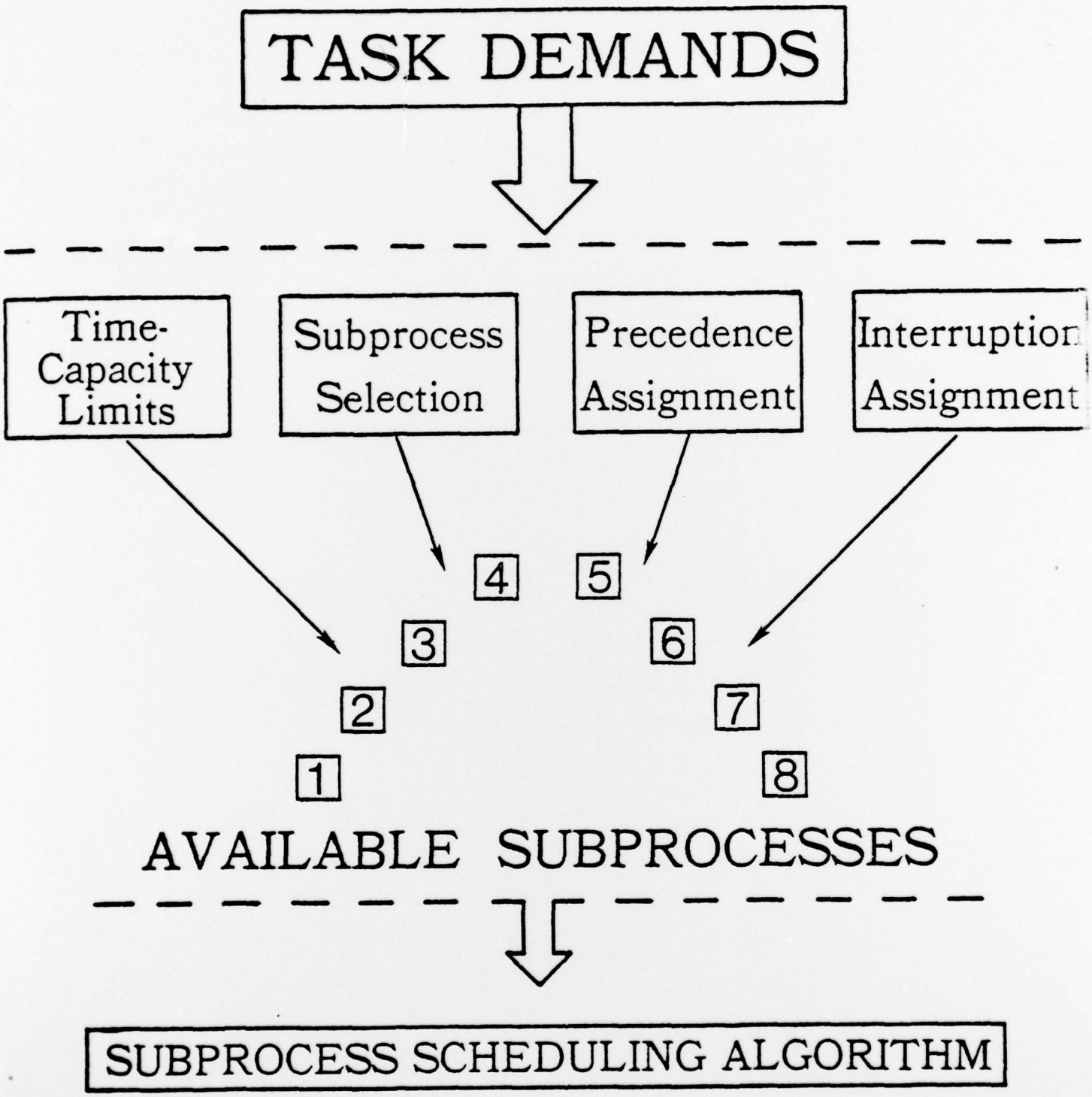


Figure 6. A scheduling model of workload.

Schedule of Subprocess Activity

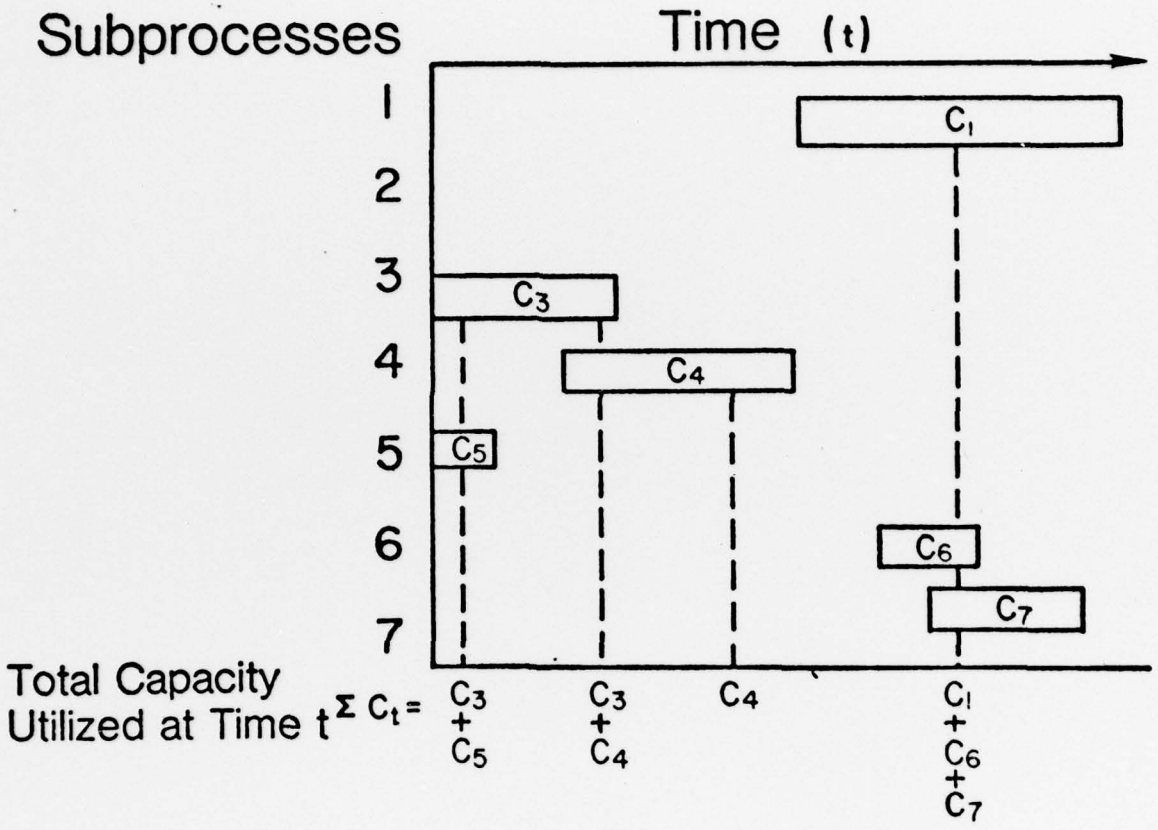


Figure 7. A hypothetical subprocess schedule exhibiting capacity utilization variations during task completion.

where T is the averaging or integration time. Generally, it would be expected that the variability of WL would decrease as T increases.

The additivity test reported in the present study and most secondary task measures appear to be more directly related to WL than they are to C_t. The flexibility of the scheduling models suggest why additivity of WL may be widely found and where non-additivity, usually attributed to structural interference, may be found.

In general, if task A requires a given set of subprocesses, the subprocess schedule for them when the task is performed alone would be different from when it is performed with task C. The Executive System would attempt to intermesh the two sets of subprocesses to perform then both efficiently. Likewise, pairing task A and D could change the distribution of subprocess use and of C_t. However, if the WL of task A is not changed, then WL would be additive. If the scheduling process achieves about the same efficiency for a variety of tasks then the WL of task A should be about the same. On the other hand, non-additivity, or structural interference, would be expected, if two tasks both require one or more subprocesses for a substantial proportion of the integration time.

In summary, WL provides a good description of the limitations on human processing. Future attempts should be directed explicitly at measuring it, coupled together with additivity tests on PECs to determine if status. The scheduling model provides a useful framework within which to interpret WL measures. Because other models are special cases of the scheduling model, the entire field of attentional studies may be integrated within it.

References

Atkinson, R. C., & Shiffrin, R. M. Human Memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), The Psychology of Learning and Motivation, Vol. 2. New York: Academic Press, 1968.

Broadbent, D. E. Perception and Communication. London: Pergamon Press, 1958.

Brown, I. D., & Poulton, E. C. Measuring the space "mental capacity" of car drivers by a subsidiary task. Ergonomics, 1961, 4, 35-40.

Campbell, N. R. Foundations of Science. New York: Dover, 1957.

Checkosky, S. F. Speeded classification of multidimensional stimuli. Journal of Experimental Psychology, 1971, 87, 383-388.

Colle, H. A., & DeMaio, J. Measurement of attentional capacity load using dual-task performance operating curves. AFHRL-TR-78-5. Williams AFB, AZ: Flying Training Division, Air Force Human Resources Laboratory, April, 1978.

Deutsch, J. A., & Deutsch, D. Attention: Some theoretical considerations. Psychological Review, 1963, 70, 80-90.

Finn, J. Multivariance: Univariate and multivariate analysis of variance, covariance and regression: Version V. Chicago: National Educational Resources, 1976.

Helm, W. R. Operator workload assessment in system test and evaluation. Sixth Symposium on Psychology in the DOD. U. S. Air Force Academy, April, 1978.

Hicks, T. G., & Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors, 1979, 21, 129-143.

Kahneman, D. Attention and effort. Englewood Cliffs, N. J.: Prentice-Hall, 1973.

Kalsbeek, J. W. H. Measurement of mental work load and of acceptable load: Possible applications in industry. The International Journal of Production Research, 1968, 7, 33-45.

Kalsbeek, J. W. H., & Sykes, R. N. Objective measurement of mental load. Acta Psychologica, 1967, 27, 253-261.

Kantowitz, B. H., & Knight, Jr., J. L. On experimenter-limited processes. Psychological Review, 1976, 83, 502-507.

Kerr, B. Processing demands during mental operations. Memory and Cognition, 1973, 1, 522-536.

Krantz, D. H. Color Measurement and Color Theory: 1. Representation theorem for Grassman structures. Journal of Mathematical Psychology, 1975, 12, 283-303.

Kristofferson, A. B. Attention and psychophysical time. Acta Psychologica, 1967, 27, 93-100.

Kryter, K. D. The effects of noise on men. New York: Academic Press, 1970.

La Berge, D. Identification of two components of the time required to switch attention: A test of a serial and a parallel model of attention. In S. Kornblum (Ed.) Attention and Performance IV. New York: Academic Press, 1973.

La Berge, D., & Samuels, S. J. Toward a theory of automatic information processing in reading. Cognitive Psychology, 1974, 6, 293-323.

Logan, G. D. On the use of a concurrent memory load to measure attention and automaticity. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 189-206.

Marks, L. E. A theory of loudness and loudness judgments. Psychological Review, 1979, 86, 256-285.

Michon, J. A. A note on the measurement of perceptual motor load. Ergonomics, 1964, 7, 461-463.

Michon, J. A. Tapping regularity as a measure of perceptual motor load. Ergonomics, 1966, 9, 401-412.

Navon, D., & Gopher, D. On the economy of the human-processing system. Psychological Review, 1979, 86, 214-255.

Nickerson, R. S. The use of binary-classification tasks in the study of human information processing: A tutorial survey. In S. Kornblum (Ed.) Attention and Performance IV. New York: Academic Press, 1973.

Norman, D. A. Toward a theory of memory and attention. Psychological Review. 1968, 75, 522-536.

Norman, D. A., & Bobrow, D. G. On data-limited and resource-limited processes. Cognitive Psychology, 1975, 7, 44-64.

Norman, D. A., & Bobrow, D. G. On the analysis of performance operating characteristics. Psychological Review, 1976, 83, 508-510.

Parker, J. F., & West, U. R. (Eds.) Bioastronautics data book. Washington, D. C.: GPO (NASA SP 3006), 1973.

Reitman, W. What does it take to remember? In D. Norman (Ed.) Models of Human Memory. New York: Academic Press, 1970.

Rohmert, W. (Special Ed.) An international symposium on objective assessment of workload in air traffic control tasks. Ergonomics, 1971, 545-672.

Schneider, W., & Shiffrin, R. M. Controlled and automatic human information processing: I Detection, Search, and Attention, Psychological Review, 1977, 84, 1-66.

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

Spelke, E. Hirst, W. & Neisser, U. Skills of divided attention. Cognitive Psychology, 1976, 4, 215-230.

Sternberg, S. Memory-scanning: Mental processes revealed by reation-time experiments. American Scientist, 1969, 57, 421-457.

Suppes, P., & Zinnes, J. L. Basic Measurement Theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.) Handbook of Mathematical Psychology, Vol. 1. New York: Wiley, 1963.

Timm, N. H. Multivariate analysis with application in education and psychology. Monterey, Calif.: Brooks/Cole, 1975.

Triesman, A. M., & Geffen, G. Selective attention: Perception or response? Quarterly Journal of Experimental Psychology, 1967, 19, 1-17.

Van Cott, H. P., & Kincaid, R. G. (Eds.) Human engineering guide to equipment design. Washington, D. C.: GPO, 1972.

Welford, A. T. The "psychological refractory period" and the timing of high speed performance: A review and a theory. British Journal of Psychology, 1952, 43, 2-19.

Wierwille, W. W., & Williges, R. C. Survey and analysis of operator workload assessment techniques. Blacksbury, Virginia: Systemaetrics, Inc., Report S-78-101, September, 1978 (AD-A059501).