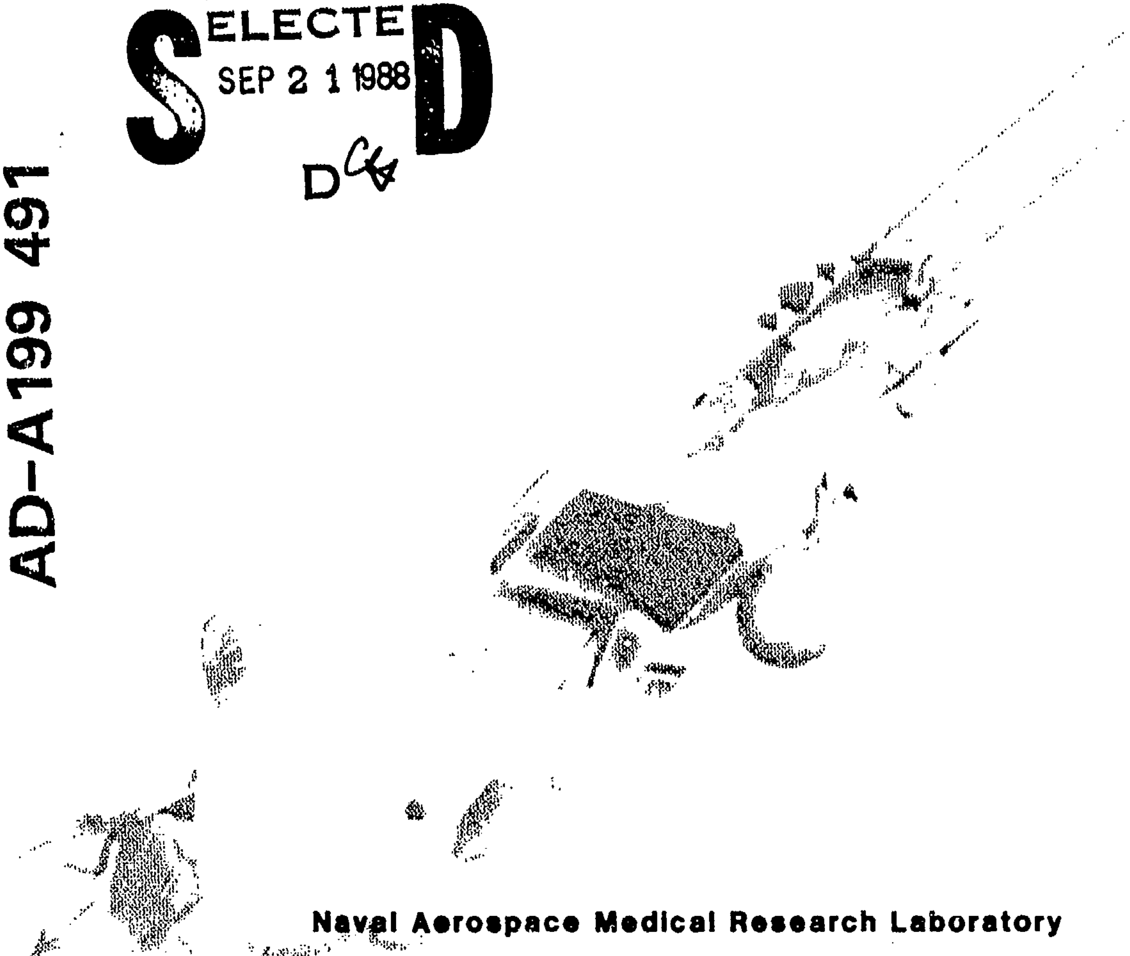
# SOME CONSIDERATIONS IN THE DESIGN OF A COMPUTERIZED HUMAN INFORMATION PROCESSING BATTERY

Diane L. Damos



Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, Florida 32508-5700

Reviewed and approved December 1987

J. O. Houghton

J. O. HOUGHTON, CAPT, MC USN
Commanding Officer

The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

AD-A 199491

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED | 1b RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3 DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>NAMRL MONOGRAPH 35 | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |

| | | |
|---|---|---|
| 6a. NAME OF PERFORMING ORGANIZATION<br>Naval Aerospace Medical Research Laboratory | 6b OFFICE SYMBOL (If applicable)<br>02 | 7a NAME OF MONITORING ORGANIZATION<br>Naval Medical Research & Development Command |
| 6c. ADDRESS (City, State, and ZIP Code)<br>Naval Air Station<br>Pensacola, FL 32508-5700 | | 7b. ADDRESS (City, State, and ZIP Code)<br>Naval Medical Command<br>National Capital Region<br>Bethesda, MD 20814-5044 |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Naval Medical Research and Development Command | 8b. OFFICE SYMBOL (If applicable)<br>Code 404 | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER |

8c. ADDRESS (City, State, and ZIP Code)
Commanding Officer, Naval Medical Research and Development Command, Naval Medical Command, National Capital Region, Bethesda, MD 20814-5044

10. SOURCE OF FUNDING NUMBERS

| PROGRAM ELEMENT NO | PROJECT NO. | TASK NO | WORK UNIT ACCESSION NO. |
|---|---|---|---|
| 62758N | MM58528 | 01.0004 | DN477519 |

11. TITLE (Include Security Classification)

(U) Some Considerations in the Design of a Computerized Human Information Processing Battery

12. PERSONAL AUTHOR(S)
Diane L. Damos

| 13a. TYPE OF REPORT | 13b TIME COVERED | 14. DATE OF REPORT (Year, Month, Day) | 15 PAGE COUNT |
|---|---|---|---|
| Final | FROM Aug 85 TO Feb 86 | 8712 | 56 |

16. SUPPLEMENTARY NOTATION
This monograph was partially supported by Contract N00014-86-K-0119 and was prepared during Intergovernmental Personnel Act assignment at NAMRL.

17. COSATI CODES

| FIELD | GROUP | SUB-GROUP |
|---|---|---|
| | | |
| [illegible] | | |

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Computerized testing, performance ~~battery~~ tests, skills, [illegible], methodology, methodological issues

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

This report discusses general issues and problems associated with the development of computerized human information processing test batteries. It is concerned primarily with batteries that will be administered in a repeated-measures paradigm although some of the information pertains to the construction of any battery. Among the issues discussed in the report are task selection, task ordering, the use of pacing, and hardware and software implementation problems. Implementation problems associated with specific information processing tasks--such as mental arithmetic tasks and vigilance tasks--are also described.

This report is intended for individuals who have a limited knowledge of human information processing and the pitfalls associated with computer-based testing. [illegible]

| | |
|---|---|
| 20 DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS | 21 ABSTRACT SECURITY CLASSIFICATION<br>UNCLASSIFIED |

| 22a NAME OF RESPONSIBLE INDIVIDUAL | 22b TELEPHONE (Include Area Code) | 22c OFFICE SYMBOL |
|---|---|---|
| J.O. HOUGHTON, CAPT MC USN, Commanding Officer | (904) 452-3286 | 00 |

83 APR edition may be used until exhausted
All other editions are obsolete

*U.S. Government Printing Office: 1985-507-047

## PREFACE

The purpose of this report is to describe some of the factors that should be taken into account in the construction of a battery of human information processing tests.[1] This report is concerned primarily with batteries that will be administered in a repeated-measures paradigm although some of the sections, such as Implementation Problems, pertain to the construction of any battery. This report is intended for individuals who have a limited knowledge of human information processing and the pitfalls associated with computerized testing. It is designed in part to supplement information previously published in professional journals and books on the properties of various information processing tests. For this reason, no data are presented, and detailed descriptions of each test are omitted since these are available elsewhere (e.g., the Unified Tri-service Cognitive Performance Assessment Battery (1) documentation). The strengths and weaknesses of each test are, however, described in some detail with specific implementation problems. The author has assumed that the reader will be assembling a battery from tests that are currently available. Therefore, no discussion of the development and verification of new performance tests is provided.

This report is divided into two major sections. The first section, Chapters 1-4, discusses general problems associated with the development of an information processing battery. These include test selection, methodological issues, and implementation problems. The second major section, Chapters 5-8, describes the two major types of tests that are included in information processing batteries: rate-of-information-processing tests and tests of higher processes. The most commonly used tests in both of these categories are described. Other types of tests, such as verbal reasoning or spatial visualization, were not included in this report because relatively few computerized versions of them have been constructed.

A glossary of terms is given at the end of this report. This glossary is intended primarily for readers with a limited knowledge of the terminology used in a human information processing context. The definitions in the glossary are specific to this monograph and should not be construed as general or exhaustive definitions of the terms.

[1] Throughout this report, the term "test" refers to the elements of a performance battery or to an instrument used to obtain data on an individual. The term "task" refers to specific instruments or classes of instruments, such as the Sternberg task or rate-of-information-processing tasks.

# ACKNOWLEDGMENT

The author thanks Mrs. Kathleen Mayer and Mrs. Neil Davis for their special efforts in the preparation of this manuscript.

## CONTENTS

# 1. ISSUES IN BATTERY CONSTRUCTION

A battery is nothing more than a set of tests that, as a whole, measure certain skills, abilities, and processes. The investigator's first job, therefore, is to select a subset of tests from those available that will measure the skills, abilities, or processes of interest. In some cases, it will be necessary to select tests that measure a broad spectrum of human information processing skills, abilities, and processes. Such 'broad spectrum' batteries usually are needed in two situations: (1) when little is known about the effect of the experimental factor on human information processing, or (2) when the investigator is interested in performance on some real-world activity that requires a broad range of information processing skills and abilities. Flying is an excellent example of the latter situation because almost every known skill, ability, and process is required by some aspect of flight.

Cont'd Pg. 1

'Narrow-spectrum' batteries are required when a given experimental factor is known to affect only certain skills, abilities, and processes. For example, alcohol has been found to disrupt response processes but to have no significant effect on memory retrieval processes (2). A scientist concerned with predicting the effects of alcohol on an activity requiring fine motor coordination would construct a battery with several tests of fine motor coordination and few, if any, tests of memory retrieval. Narrow-spectrum batteries are also used when the activity of interest requires relatively few skills and abilities. For example, most sonar operator's activities require visual perception and pattern comparison skills but relatively few fine motor skills. Thus, a battery for testing the effect of some experimental factor on a sonar operator's performance would include few, if any, tests of fine motor skills.

Currently, an investigator has at least three ways to select tests for either a narrow- or broad-spectrum battery that is concerned with a specific real-world activity. First, the investigator can abstract the skills and abilities needed to perform the activity successfully from the appropriate task analysis. The major problem with this approach is that most task analyses only describe observable events. As a result, determining which information processing skills and abilities could have resulted in the observed behavior is often almost impossible.

Second, the investigator can select tests that correlate with known predictors of the activity or use the predictors themselves in the test battery. For example, suppose scores on a general intelligence test correlate with scores on a paper-and-pencil test of spatial reasoning and that the spatial reasoning scores predict performance in flight training. To construct a test battery to predict success in flight training, the investigator could include either the general intelligence test or the spatial reasoning test. This approach has several problems. The major one is that performance on very few real-world activities is accurately predicted by either paper-and-pencil tests or computerized tests. Thus, this approach could be used only for a very few activities.

Third, the investigator can rely on personal knowledge of the activity to identify the required skills and abilities and select tests that measure these skills and abilities. Although many pitfalls are associated with this approach, it is often the one used in battery development because of unavailability of adequate task analyses and the lack of reliable predictors of performance.

If an investigator is interested in examining the effect of a given experimental factor on human information processing in general, selecting tests for a battery is much easier. In this situation, the investigator usually will construct a broad-spectrum battery and choose one or two tests from each major category of interest. So, for example, the investigator may include in the battery one verbal short-term memory test, one verbal reasoning test, one spatial short-term memory test, one spatial reasoning test, et cetera. The only common restriction in constructing such a battery is the total time available to test each subject.

## 2. TEST SELECTION

Once the investigator has decided on either a broad-spectrum or a narrow-spectrum battery, there are many questions to be considered in selecting specific tests for the battery.[2] This chapter presents and discusses 11 questions in approximately the order in which they should be considered. Some of the questions pertain primarily to a repeated-measures paradigm and may not be of concern to an investigator constructing, for instance; a neuropsychological battery. These 11 questions are suggestive of those the investigator should bear in mind when selecting tests for a battery; the list is not exhaustive. Additionally, these questions pertain only to tests of information processing skills, abilities, and processes. They are not directly applicable to tests of personality traits or mood scales.

**1. Does the test measure a specified skill, ability, or process?** What the test measures should be clearly identified and should be specific. Data should be available either from the developer or the scientific literature demonstrating that the test does indeed measure its purported skill, ability, or process (see Question 3). Tests of 'general information processing' or 'memory' are generally worthless.

**2. Does a test measure the same skills, abilities, or processes as other tests already included in the battery?** Multiple tests of a certain skill, ability, or process should be routinely included in narrow-spectrum batteries but excluded from broad-spectrum batteries unless the skill, ability, or process is extremely important for the successful completion of the activity.[3] For example, the aircrew selection battery under development at the Naval Aerospace Medical Research Laboratory (NAMRL) includes several tests of spatial processes. These tests are included because spatial processing plays a critical role in aircrew performance.

The major problem in selecting a test of a skill, ability, or process concerns the poor intercorrelation between tests purportedly measuring the same thing. The main point to bear in mind is that no 'pure' tests of most skills or abilities exist. For example, no 'pure' test of spatial ability has been developed because other skills and abilities affect performance on the 'spatial' test. Typically, if two tests that purportedly measure the same skill or ability correlate poorly, a detailed analysis of the tests will indicate that each requires a number of skills and abilities not required by the other test. In contrast, tests purportedly measuring the same processes typically correlate highly because processes are initially identified by a very rigorous and time-consuming series of experiments.

[2] The author assumes that all subjects will be screened for any basic physiological abilities required by the tests of a battery, for example, all subjects will be screened for color blindness if one of the tests of the battery requires color discrimination.

[3] Multiple tests of the same skill or ability are routinely included in narrow-spectrum batteries because, as discussed in the next paragraph, no pure tests of skills or abilities exist. Therefore, multiple tests of a given skill or ability are included to increase the probability of obtaining an accurate assessment of the desired skill or ability.

An investigator's objective in designing a battery is to select a test that correlates most highly with real-world performance or that is most representative of a general category of skills, abilities, or processes. As noted earlier, few correlations between information processing tests and performance on a real-world activity are available. Therefore, selecting tests that require the same skills, abilities, or processes as the activity under study must be based on other considerations, such as the amount of practice required to reach differential stability[4] (see also the appendix) or the amount of baseline data available. Similarly, selecting a test to represent a general category of skills, abilities, or processes should be based on a variety of considerations, including those discussed in Questions 1,3,5,7, and 8.

3. **How was the test validated?** To claim to have a test of a never-before-measured skill, ability, or process, the developer must complete an elaborate series of experiments. These experiments must demonstrate that the new test is affected by a variety of experimental factors in the way that would be predicted if it indeed measured the purported skill, ability, or process. This procedure is extremely time-consuming and costly and appears rarely to be attempted outside the university environment. The best example of this procedure probably is Sternberg's development of a test of memory scanning (3,4).

To demonstrate that a new test measures some skill, ability, or process that is measured by other validated tests, the test developer still must complete a fairly time-consuming procedure.[5] The developer must demonstrate that the new test correlates with the other validated paper-and-pencil or computerized tests of the same skill, ability, or process. Occasionally, the developer also must demonstrate that the new test predicts performance in a real-world situation where the skill, ability, or process is known to affect performance.

4. **Are baseline data available?** The test developer should at least furnish means, standard deviations, and ranges for asymptotic performance

[4] A differentially stable task has three characteristics: (1) the mean of the group's performance is constant or increasing in a slow, linear fashion; (2) the standard deviation of the group's performance is constant; and (3) the rank order of subjects is constant. Statistical tests are used to determine differential stability; thus, some level of error ($\underline{p}$ = .05, .01, etc.) is involved in asserting that the mean, standard deviation, or the rank order of subjects is "constant." All tests of differential stability are performed on the intertrial correlations, not on the raw data.

[5] This paragraph raises the question of why anyone would develop a test of some skill, ability, or process when a validated test already existed. The primary answer is convenience; many information processing tests require extensive practice before differential stability is reached (see Question 5) or require large amounts of data for analysis. Thus, all the validated tests of some skill, ability, or process may be impractical to use in an applied situation.

for a clearly defined population.[6] Learning data (time to asymptote, time to differential stability, or learning curve parameters) should also be given. These statistics are often not available for the population of interest. The investigator then must decide either to extrapolate from the existing data to the population of interest or to collect baseline data on individuals from the population of interest.

Baseline data can help the investigator implement the test. If summary statistics obtained in the investigator's laboratory do not correspond to those obtained from the test developer for a comparable population, the test may have been incorrectly implemented. That is, the differences may be caused either by programming errors or hardware problems. Another possibility is that the 'comparable' populations are different. In any case, the existence of baseline data is a valuable aid in the development of the battery.

A complete lack of baseline data is an extremely serious indicant that the test has not been adequately developed. If baseline data are not available, the test is either in an extremely early stage of development or has not been subjected to rigorous examination, and the nature of what is being tested should be questioned.

**5. How much practice is required for the test to obtain differential stability?** Stability is not a now-or-never state of affairs; it is determined by practice. Therefore, if a test has not reached differential stability at some point, a small additional amount of practice may make it stable. The major issue then concerns the amount of practice necessary to reach differential stability. The investigator must determine how much time is available for practice before the experiment begins and select tests that obtain stability during this period.

Currently, the only publication available describing the time to stability for a variety of tests is Bittner et al. (6). This report summarizes several years of work on the Performance Evaluation Tests for Environmental Research (PETER) project. The reader should remember that all of the tests were evaluated using one testing schedule and one type of subject--volunteer, enlisted personnel. In some cases, the same subjects performed many of the same tests. If the schedule affects the time to stability, the values given in Bittner et al. may differ from those obtained using either more massed or more distributed testing schedules. Similarly, if the

[6] The reader should note that asymptotic performance and differential stability are not identical. Differential stability is described earlier in the monograph and in the appendix and is mathematically determined. Asymptotic performance has two meanings. The first, the more uncommon, occurs when a learning curve has been fit and an asymptote has been identified mathematically. This use of the term "asymptotic performance" means "the terminal level of performance after an infinite amount of practice." The more common use of the term asymptotic performance (also called stable performance) implies that the mean performance on several consecutive trials did not change or changed very little. Most uses of this term imply a judgment by the investigator that performance would not improve further with practice. The riskiness of this assumption is demonstrated by Bradley (5).

subject population affects the time to stability, the figures in Bittner et al. may not be accurate, and the data should be regarded as positively biased by the experimental sophistication of the subjects. Currently, few data exist on the relation between time to stability and the testing schedule, and those data conflict. No data examine the time to stability as a function of the characteristics of the subjects taking the test. Therefore, the values given in Bittner et al. should be regarded as estimates of the time to reach differential stability, not as absolute figures.

Several other facts about stability should be discussed. One of these concerns tests, such as the Sternberg memory search task, that use the slope between the average correct reaction time and a task variable as a dependent measure. These tests are among the most carefully developed and theoretically important ones in cognitive psychology (see Chapter 6). Generally, Bittner et al. (6) found that the slopes of these tests did not stabilize during the testing period or stabilized very late in testing. Thus, these tests may require a great deal of practice before they can be used effectively.

A second fact concerns tests with multiple dependent measures, such as mental arithmetic tasks that use average correct reaction time and percentage correct. Commonly, one dependent measure of a test obtains stability before another. In selecting tests for a battery, a scientist should determine which dependent measures are of interest and ensure, as much as possible, that those measures will become stable in the time period allotted for training.

A third fact concerns the stability of task combinations. Many batteries designed to address applied issues include one or more task combinations. Although few data are available, performance under dual-task conditions appears to stabilize very slowly. At this time, almost no data show the relation between the stability of performance of each task singly and the stability of performance of each task under dual-task conditions. Performance on each of the tasks apparently does not have to be stable for the combination to be stable; a tracking-mental arithmetic combination investigated at NAMRL had some stable dependent measures although neither task was stable when performed alone. Because of the lack of knowledge about multiple-task stability, investigators should consider the value of including task combinations in a battery.

Finally, the investigator should consider the issue of post-stability test definition. Theoretically, differential stability does not depend on the magnitude of the intertrial correlations; only their consistency determines stability. Test definition is concerned with the magnitude of the intertrial correlations. Generally, if the average correlation is less than 0.7, the test is said to have poor definition and is considered to contain too much (50%) unpredictable variance to provide usable data.

**6. Is the test sensitive to the experimental factor under consideration?** Sensitivity implies that at least one experiment has demonstrated a statistically significant change in test performance from the experimental factor in question. Tests vary greatly in their sensitivity to experimental factors. A given test may be sensitive, for instance, to heat but not to vibration. Using a sensitive test greatly improves the probability t[illegible] investigator will find a statistically significant effect of the [illegible]ntal factor.

7. **How many data points can be collected per unit time?** Some tests, such as vigilance tests, generate, at best, one datum per 10-min period. In contrast, reaction time tests may generate 60 or 70 responses per minute. The investigator must consider the time scale of interest and select tests that generate a sufficient amount of data for analysis purposes during the experiment.[7]

8. **How much data will be unusable?** Equipment failure and operator error sometimes result in the loss of large amounts of data, but these events cannot be predicted. The dependent variables and the experimental factors strongly influence the proportion of unusable data, and to some extent, the proportion of unusable data can be predicted for a given experimental situation. For example, many experimental tasks require only a yes/no or a true/false response. For the majority of these tasks, 50% correct represents chance performance. If an experimental factor makes the task difficult, the percentage correct could fall to chance levels. The effects of any subsequent experimental manipulations would be almost impossible to detect, and the data are usually discarded. Some computerized mental arithmetic tasks seem particularly susceptible to this type of problem; the subject's response is often simply scored as correct or incorrect. Some subjects have so much trouble with this type of task that they make few correct answers, and their data are normally discarded.

9. **How many dependent measures does the test have?** Tests with one dependent measure must be analyzed using univariate analysis techniques, such as _T_ tests and _F_ tests. These are familiar to most investigators and are generally easy to execute and interpret. Tests with multiple dependent measures, such as percentage correct and average correct reaction time, may be analyzed using either univariate or multivariate techniques, depending on the characteristics of the obtained data and the inclinations of the investigator.

One school of thought maintains that multiple dependent measures obtained from a given task should be analyzed routinely using multivariate techniques. Then, if the measures are uncorrelated, univariate analyses can be used. Because this approach is relatively new, it is somewhat controversial and has several drawbacks associated with it. One drawback is that multivariate analyses are less familiar to most investigators than univariate analyses and are more difficult to interpret. A second drawback is that the statistical power associated with a multivariate analysis is difficult to determine. Therefore, if a investigator fails to find an expected effect, it is difficult to determine if the effect 'really' was not there or if the power associated with the analysis was just too low to detect the effect.

[7] The amount of data necessary for analysis purposes depends, among other things, on the analyses to be conducted and the amount of statistical power the investigator desires. Readers uncertain about the amount of data to be collected should consult someone with statistical expertise.

A second school of thought maintains that, for most applied work, multivariate analyses are simply too difficult to interpret to be of use. This school attempts to collapse multiple dependent measures from a given test into one derived score, such as an information-transmitted score. The problems associated with this approach are discussed below. In any case, the investigator should be aware that using tests with multiple dependent measures may require relatively sophisticated analyses that can be both time consuming and expensive.

**10. Are the dependent measures of a given test raw scores or derived scores?** Typically, derived scores--such as Z scores or proportion scores--are more difficult to analyze than raw scores because their characteristics more frequently violate the assumptions of univariate analysis. The author has performed simple statistical analyses on both the derived scores and the raw scores from several tests and found that the results of the analyses performed on the derived scores were considerably different from those performed on the raw scores. Thus, tests that use derived scores as dependent measures should be regarded with some caution.

**11. Are there large individual differences in performance?** Few tests have been studied in sufficient depth to identify consistent individual differences in performance in a meaningful manner. Some tests show large individual differences with a relatively normal distribution of scores. These tests often are excluded from use because the large individual differences mask the effects of experimental variables. Other tests show large individual differences with a bimodal distribution of scores. An investigator should carefully weigh the advantages of including in a battery any test with a bimodal distribution of one or more dependent variables; varying the proportion of one type of subject over another can result in statistically different outcomes.

## 3. METHODOLOGICAL ISSUES

Several methodological (procedural) issues must be given at least some consideration before the investigator selects the tests of the battery. Three of the most important--test order, pacing, and knowledge of results--are discussed in this chapter.

### TEST ORDER

One problem in the development of any test battery concerns the order in which the tests will be presented. In selecting the order, the investigator must consider two major issues: content and carry-over (sequence) effects.

Content issues are concerned with sequencing the tests to avoid subject boredom or fatigue. For example, some tracking tasks are physically fatiguing. The investigator, therefore, may not want to schedule another physically fatiguing task immediately before or immediately after a tracking task. As another example, classical vigilance tasks generally result in very low arousal levels. Indeed, these tasks are often so monotonous that subjects fall asleep while performing them. Thus, investigators may want to schedule vigilance tasks at the beginning of a battery when the subjects are most likely to be alert rather than at the end when subjects may be both physically and mentally fatigued.

At this time, there appears to be no guidelines for taking content issues into account in test sequencing. The investigator must rely strictly on general knowledge about the characteristics of each test and combine this knowledge with the purposes of the experiment to determine a sequence that introduces the smallest possible number of artifacts into the data.

More is known about carry-over effects than about content-related dependencies. The presence of carry-over effects is serious because it is not possible to use certain experimental designs if carry-over effects exist. For example, the Latin square design can be used only when there are no carry-over effects between any of the levels of the experimental factors (7). Unfortunately, there is no way to predict when carry-over effects will occur. Therefore, an investigator may conduct an experiment using some design that precludes carry-over effects and find after all the data are collected that these effects have occurred. In such a situation, the data cannot be analyzed using the specified design, and the investigator may find no satisfactory way of analyzing the data. The most conservative approach to carry-over effects is to assume that they will occur and then to conduct pretests to determine their magnitude. Simon (8) provides a good discussion of statistical methods of identifying and controlling for carry-over effects.

Most of the current knowledge about carry-over effects in applied research has been obtained from examining the simplest possible situation: two tests administered in a counterbalanced fashion. The carry-over effects obtained from a two-test, counterbalanced experiment are often called 'asymmetric transfer effects.' Basically, asymmetric transfer occurs when the transfer from Test A to Test B is not the same as the transfer from Test B to Test A.

Poulton documented numerous instances of asymmetric transfer (9,10). Most of these instances demonstrate asymmetric transfer between tracking tasks using (1) quickened versus unquickened displays, (2) magnified versus unmagnified displays, and (3) pursuit versus compensatory displays. More recent work by Poulton has demonstrated asymmetric transfer between different dual-task combinations (11). Damos (12) and Damos and Lyall (13) have shown asymmetric transfer between versions of the same task combination that differ only in the response mode (manual or speech) used to control one of the tasks.

The major problem with asymmetric transfer, like other carry-over effects, is that it can seriously bias the data. Damos (12) and Damos and Lyall (13) demonstrated that asymmetric transfer effects can be so large that they cause spurious statistical effects or completely mask true effects. Currently, there is no way to correct mathematically for asymmetric transfer once it occurs. Therefore, all data affected by asymmetric transfer must be discarded; and, if the experimental design is still usable, the analyses must be recalculated on the remaining data with the subsequent loss of statistical power.

At this time, there is no reason to assume that other types of carry-over effects have less serious consequences for data analyses than asymmetric transfer effects. Since predicting when carry-over effects will occur is impossible, carry-over effects must be planned for in these experimental designs.

## PACING

For each discrete test in a battery, the investigator must consider if the test should be paced or unpaced. Although a great deal of literature examines the effect of machine pacing on industrial workers, few of these studies provide information pertinent to the development of a performance battery. One reason for using a paced rather than an unpaced version of a test is to simulate a real-world task more closely. Another reason is that pacing may have an alerting property. Thus, the judicious use of pacing may decrease the boredom caused by prolonged testing. Finally, paced tests can result in better performance than unpaced versions of the same test (14).

There are also a number of reasons for not using a paced version of a test. The first reason is that the use of a paced test often adds an additional dependent variable; many investigators analyze the number of missed stimuli in a trial separately from either the number of incorrect or the number of correct responses. The use of a third variable complicates the knowledge of results given to the subject and subsequent data analyses.

A second reason is that certain information processing stages may be affected more by the speed stress induced by pacing than others (15). Thus, the paced and unpaced versions of a test may differ in a number of subtle and unidentified ways. Currently, not enough data are available to identify the stages differentially affected by speed stress, and there is no way to predict which stages may be more affected than others.

The third reason is that many subjects' performance is disrupted by pacing, particularly under multiple-task conditions. This disruption may be present even when the pacing interval is objectively too long to affect performance, that is, some subjects may be so distracted by the knowledge that the test is paced that their performance is adversely affected. Under multiple-task conditions, paced combinations usually result in different response strategies and appear to be much more frustrating and tiring than the unpaced version of the same combination.

Fourth, paced tests may result in different excretion levels of various catecholamines (14) than unpaced versions of the same test. This finding cautions primarily against changing from an unpaced version of a test to a paced version during the course of the experiment although other physiological measures--such as heart rate, respiration rate, and blood pressure--often show no difference between the paced and unpaced versions of a test (see reference 16 for an example).

## KNOWLEDGE OF RESULTS

Another issue an investigator must resolve during the design phase of an experiment concerns knowledge of results (KR). If the subject is given KR, the investigator must decide what type of KR should be presented and how often it should be provided. Fortunately, the effect of KR on motor and simple cognitive tasks has been extensively studied. (See reference 17 for a very basic review of the terminology and general results and 18 for an extensive literature review.)

Generally, KR has two functions. It decreases the time to reach any performance criteria established by the investigator, and it maintains the subject's motivation. Because KR is beneficial, it is almost always provided in human performance[8] laboratory research. Knowledge of results is routinely omitted only for vigilance tasks and for tasks that provide a great deal of intrinsic feedback. The reason KR is not provided during vigilance tasks is because it usually eliminates the main phenomenon of interest, the vigilance decrement. It may, however, be presented at the end of a vigilance session to provide the subject with a performance summary. Tasks providing large amounts of intrinsic feedback, such as some tracking tasks or risk-taking tasks, arguably do not need KR for performance information. Nevertheless, KR may be provided for these types of tasks to maintain the subject's motivation.

Therefore, for most experiments, the investigator must decide between concurrent (presented during the performance of the task) and terminal (presented at the end of a trial or a session) KR and must determine the accuracy (precision) of the KR. Considerations pertinent to both these decisions are described below.

### Concurrent Versus Terminal KR

Normally, deciding between concurrent and terminal KR is easy; except in multiple-task experiments, human performance research uses almost

[8] One notable exception to this is studies of exotic environments. Usually, KR is not provided during the exotic environment because subjects may be able to develop strategies to compensate for the environment. The subjects, however, are still usually trained with KR.

exclusively terminal KR. In some laboratory experiments, terminal KR may be given after every response of a discrete task to provide the subject with immediate performance information. Such a presentation schedule is rarely used in applied contexts because it requires too much time and may prevent the subject from developing a response strategy or a response rhythm. Instead, investigators in applied situations tend to present KR after a trial (which may be defined either as a fixed time period or a fixed number of responses), after a block of trials, or after a session. Sometimes, good reasons preclude providing any KR to the subject for a given test.

As noted above, concurrent KR is used almost exclusively in multiple-task experiments to control the priorities that subjects assign to the tasks. Few techniques for presenting concurrent KR have been developed, and all of these have serious drawbacks.

Gopher and North (19) developed one of the few intermittent concurrent KR techniques. If the subject's performance dropped below a certain level on one of the tasks, a brief tone sounded. The subject then attempted to improve performance on the associated task. Unfortunately, no feedback was provided to the subject indicating that performance on the task had once again reached an acceptable level. Thus, although this technique appears to be straightforward, subjects were frequently confused about the acceptability of their immediate level of performance.

This intermittent technique was superseded by the moving bars technique, a more complicated method for presenting concurrent KR (e.g., references 20-23). This method displays one bar graph and one desired performance line for each task. The height of the bar graph changes during a trial; its height reflects the subject's average performance calculated over some period of time, typically 5 or 10 s. The taller the bar graph, the better the subject's performance on that task. The subject is usually instructed to perform so that the moving bar graphs reach or exceed the desired performance lines for their respective tasks. The experimenter can adjust the height of the desired performance lines to any level to control the relative priorities of the two tasks.

Although this technique sounds impressive, it also has several drawbacks. The most obvious is that it requires a considerable amount of the processing capacity of the computer to calculate and adjust the height of the bar graphs. The resolution of the graphics system also must be sufficient to portray smooth movement rather than discrete jumps in bar graph height. Another problem is that the presence of the bar graphs may act as a third task or a distraction, depressing performance on the two tasks of interest. Additionally, subjects may be more inclined to regard the experiment as a game when the bars are present; some subjects appear to be much more interested in manipulating the height of the bar graphs than performing as instructed. Finally, the investigator must develop an algorithm or at least a rationale for calculating the momentary height of the moving bars and for setting the value of the desired performance lines. No guidelines exist for establishing these values. Determining the values of the various parameters is a time-consuming process. and the investigator should allow an adequate amount of pretest time to experiment with the display.

Another method for presenting concurrent KR was developed by S. Harris at NAMRL (24). To use this method, each trial must be divided into two parts. Performance data on each task are collected during the first part of the trial. The trial is stopped at the end of the first part, and the data are then analyzed according to an algorithm determined by the investigator. The results of the analyses are displayed to the subject using a circle with one pointer. If the pointer points towards the 12 o'clock position, the subject has assigned the correct priorities to the two tasks (or is distributing attention as intended). If the pointer is displaced to the left of the 12 o'clock position, then the subject is favoring the left-hand task by an amount proportional to the displacement of the pointer from vertical. Similarly, if the pointer is displaced to the right of vertical, then the subject has been favoring the right-hand task. After the subject has seen the KR display for a short time, it is erased, the trial resumes, and the subject changes his performance to correct for any displacements of the pointer from the vertical position.

Again, this method has all of the drawbacks of the moving bars technique and one additional drawback: The trial is actually stopped during the presentation of KR. Although this prevents the KR display from distracting the subject, the subject must re-establish any cognitive or response strategies during the second part of the trial.

None of the techniques for presenting concurrent KR is completely satisfactory. Research on these techniques appears to have been abandoned, at least temporarily, because few investigators believe that such techniques are absolutely necessary to control the priorities that the subjects assign to the tasks.

## Precision of KR

If KR is used, the investigator must decide how precise the information given to the subject should be. A clear distinction should be made between inaccurate KR and imprecise KR. Inaccurate KR refers to KR that is misleading, that is, deceitful. In most cases, investigators cannot use inaccurate KR unless its use has been approved by the responsible human subjects committee. Inaccurate KR is used very rarely in human information processing research; its effects are often motivational and of little immediate interest. Imprecise KR is simply KR that is not as accurate as the data. For example, an investigator may record reaction times to millisecond accuracy but present reaction time KR to tenths of a second accuracy. In this example, the KR is imprecise but not misleading. No guidelines are available concerning the precision of the KR presented to the subject. The author's impression is that, for simplicity, most investigators present KR that has the same degree of precision as the data.

# 4. IMPLEMENTATION PROBLEMS

This chapter deals with some of the more common problems that investigators encounter when they implement existing tests on their own equipment. The reader may wish to consult Behavioral Research Methods, Instrumentation, and Comp'ters for current information on other pertinent hardware and software problems. This journal publishes articles on instrumentation, testing, computer technology, and algorithms. Most of the articles are concerned with microcomputers and many deal with problems that could occur in the development of a performance battery. The reader may also wish to consult Moerland et al. (25) for a discussion of the effects of computerizing standard tests on test-retest reliability, validity, and administration time. All of the problems discussed in this chapter should be eliminated when a standardized performance assessment battery implemented on standardized equipment becomes available. These pro'iems are grouped into three major categories: hardware, software, and subject instructions.

## HARDWARE PROBLEMS

Most human performance tests are developed in university or specialized government laboratories. Typically, these laboratories use equipment that was developed specifically to assess human performance. Investigators working in more applied settings often do not have access to comparable pieces of equipment. In most cases, substituting general apparatus for specialized apparatus has no effect. However, equipment substitution can have serious consequences for human performance testing in at least two instances.

The first of these concerns the keypads. Most keypads used in university laboratories are specially manufactured for research purposes or are selected according to very strict criteria from commercially available products. In many applied situations, the investigator may be forced to use the standard QWERTY keyboard or a keypad that comes with the computer. These devices typically have several problems. The keys are often relatively slow to respond and may stick at the contact point, causing very distorted reaction times (this sticking cannot always be detected simply by pressing the keys a few times). The keys may also "bounce." Bouncing occurs when the computer reads several responses rather than one because of the tendency of the contact points to deform repeatedly after a normal key press. This problem can be corrected in software, or the investigator can purchase keypads that register responses by changes in a magnetic field. Bouncing may also occur if the subject accidentally depresses a key for a few hundred milliseconds; the program may read several responses and store incorrect data. Finally, the QWERTY keyboards and keypads sold with most microcomputers are not well designed from a biomechanical standpoint. This may result in spuriously long reaction times from certain keys.

A second problem concerns the graphic systems. Most microcomputer displays have relatively poor resolution. Using tests that require smooth motion in either two- or three-dimensional space is almost out of the question with most of these systems. Indeed, sometimes the resolution of these displays is so poor that even relatively simple two-dimensional figures cannot be drawn accurately. This becomes a serious problem for many of the best spatial tests, such as the rotated letters (figures) test. A recent version of the NAMRL aircrew selection battery had to use letters rather than figures for exactly this reason. Color is another display problem.

A few human performance tests require color discrimination. If only monochromatic display systems are available, then the test either has to be discarded or modified to allow a different type of discrimination. Any modifications to the test must be thoroughly tested and validated before the modified version can be used.

**SOFTWARE PROBLEMS**

Three persistent problems may occur in developing software for human performance tests. The most dangerous problem stems from the small changes the programmer makes to accommodate hardware limitations; very small and apparently innocuous changes in stimulus presentation or timing can radically alter the nature of the test. The most common change of this type concerns stimulus presentation; programmers often change from simultaneous to sequential stimulus presentation to accommodate limitations in the graphics system. The introduction of almost any delay between the presentation of two stimuli will require the use of one or more memory systems. If these systems were not required in the original version of the test, the new and the original versions may have very different characteristics. Frick (26) presents a good example of the processing changes that occur when stimuli are presented sequentially rather than simultaneously.

Another "small" change that occurs frequently is the size of the stimuli. Programmers may inadvertently change the size of the stimuli when modifying the software for a new display. In some cases, such changes will have little, if any, detectable effect on the subject's performance. In other cases, however, such changes may have a noticeable effect, particularly if the stimuli are accidentally reduced in size. For example, many experimental variables--such as fatigue, drugs, and ambient illumination--may reduce the subject's visual acuity. The subject then might have difficulty perceiving a stimulus that was accidentally reduced in size but not one that was the correct size. Such perceptual difficulties might result in a variety of unanticipated (and unwanted) statistically significant performance effects.

The second problem is related to speed. Most human performance tests are written in compiled languages and many are written predominantly in assembly language. Programmers writing in a noncompiled language should ensure that no response-stimulus delays have been introduced in the program. Stimulus presentation also must be checked to ensure that the stimuli are presented in the same fashion as in the original version. The most common problem with microcomputers is that, because of the limitations of their graphics systems, stimuli are sometimes drawn rather than flashed on the screen. Drawing the stimuli allows some information to be analyzed immediately and can change the cognitive processes required by the test.

The third problem concerns the manner in which the subject's response is detected. Either interrupt-driven or software timing loops can be used to detect a response. When a program is interrupt-driven, the program stops at some point until the subject makes a response. A signal is then sent from the response device to the computer, indicating that a response has occurred. The program then processes the response and performs other functions until it again stops to wait for another response from the subject.

Many investigators believe that interrupt-driven software provides the most accurate measurement of reaction times. This, however, is not true; most of the variability in reaction time measurement occurs after a response has been detected and processed. Generally, the majority of the error of measurement is caused by variability in the time required to present the next stimulus to the subject.

Interrupt-driven software has two problems. First, because the program waits for the subject to respond to continue processing, fixed-length tiials are impossible to obtain. Typically, after the program detects a response, it checks the clock to determine if the trial duration has been exceeded. If it has, the trial is stopped at this point. If not, the program finishes the remaining functions and again waits for the subject to make a response. Thus, the trial can be stopped only after the subject makes a response. If the subject does not respond for some reason, the trial will go on indefinitely. Second, interrupt-driven software typically requires special response devices that signal the computer when a response has occurred. Many common laboratory keypads and keyboards cannot be used as interrupt-driven devices.

The second type of reaction time measurement uses software timing loops. Generally, a program using this technique performs a number of initial functions, presents a stimulus to the subject, and then enters a software timing loop. This loop may contain any number of statements, but one must be a command to check the response device(s). If no response is detected, the program continues executing this loop. As soon as a response is detected, the program typically exits the loop and reads the system clock to record the time when the response occurred.

One major advantage of using software loops to measure reaction times is that this technique can be used to create trials of specified durations. This is done by inserting a statement in the timing loop to read the system clock and compare it to the specified trial length. If the elapsed time exceeds the trial length, the trial is terminated without a response from the subject. A second advantage of this technique is that no special response devices are required. The only minor drawback of software timing loops is that they result in slightly more variability in reaction time measurement than an interrupt-driven approach. This occurs because most timing loops contain a number of statements. Because a response can occur during the execution of any statement in the loop, the number of statements to be executed before a response is detected varies. The amount of variance that occurs in measuring reaction times using this technique depends on the number of statements in the loop, but the time required to execute each statement is normally so small that this source of variance is trivial compared to the variance associated with the presentation of the stimulus.

The interrupt-driven technique may be combined with the timing loop technique by placing interrupt-driven statements in the timing loop. This hybrid technique allows fixed-length trials but requires the same special hardware needed by the normal interrupt-driven software. On the whole, the best approach for most human performance research is to measure reaction times using timing loops or a combination of timing loops and interrupt-driven software rather than using only interrupt-driven software.

## SUBJECT INSTRUCTIONS

Developing instructions for a computerized test is frequently a time-consuming process. Typically, the instructions for most human performance tests are designed for one-on-one interactions. That is, the experimenter reads or plays a tape of the instructions to the subject and allows the subject to ask questions. This procedure is not always practical in applied settings in which many subjects are tested concurrently and the experimenter cannot move from subject to subject to answer questions. The goal then in applied settings is to deliver clear instructions automatically to the subject. Written instructions generally are used rather than taped instructions in applied settings because it is easier for the subjects to reread passages they do not understand than to replay a tape.

Written instructions are not easy to develop. Using simple language and including examples of stimuli and responses either on the display screen or on loose sheets of paper placed near the computer does help. If a test seems particularly difficult for subjects to understand, a short pretest can be administered. If the subject does not score above a predetermined criterion, then the experimenter can be notified to provide additional help.

The use of a computerized test does not diminish the need for standardized procedures for interacting with subjects. This is particularly true when more than one individual will have contact with the subjects in a given context, that is, there is more than one experimenter. Standardized procedures for obtaining informed consent, introducing subjects to the testing area, and answering questions should be developed before any data are collected. These procedures should be strictly followed to minimize any experimenter-induced biases.

## 5. CLASSIFICATION TECHNIQUES FOR INFORMATION PROCESSING TESTS

After deciding on a broad- or a narrow-spectrum battery and addressing some of the methodological issues, the investigator needs to select specific tests for the battery. Currently, tests are classified using a variety of different schemes. The oldest scheme classifies tests according to what the subject is required to do. Thus, there are tracking tasks, vigilance tasks, choice reaction time tasks, psychomotor tasks, et cetera. This scheme is used today only for very well known tasks, such as tracking tasks or vigilance tasks, because it does not describe the tasks in the detail required by modern cognitive psychology.

A second relatively new scheme, which is based on Wickens' Multiple Resources Model (24), uses a number of different dimensions to describe a test: code of processing (verbal versus spatial), stage of processing (perceptual and central versus response), stimulus mode (visual versus auditory), and response mode (manual versus vocal). This scheme is used most often to describe test combinations. Because it is a relatively new scheme, it is not yet commonly used in applied research. Additionally, it has not been widely accepted in cognitive psychology.

A third scheme identifies tests according to the primary cognitive structure (i.e., short-term memory) or process (memory retrieval) they purport to measure. This scheme is based loosely on cognitive psychology and appears to be the most widely accepted classification scheme for applied research at present.

In the following chapters, tests are classified using both the second and third schemes described above to the extent possible.

## 6. RATE-OF-INFORMATION-PROCESSING TESTS

### OVERVIEW

The rate-of-information-processing tasks are among the most theoretically important and widely used tests available today. This category includes the Sternberg memory search task, the Neisser visual scan task, the mental rotation task, and the choice reaction time task. These four rate-of-information-processing tasks defy easy classification using the two schemes described earlier. Using Wickens' Multiple Resources Model, all four of these tasks require predominantly early rather than late processing. Responses to any of the four tasks may be made either verbally or manually and, except for the mental rotation task, stimuli may be presented either visually or auditorily. The tasks require verbal processing code resources except for the mental rotation task, which requires spatial processing code resources. Using the third scheme, the Sternberg and Neisser tasks require some memory functions, the mental rotation task may or may not depending on its implementation, and the choice reaction time task requires very minimal memory functions. The mental rotation task is usually assumed to require spatial processing; the other three are assumed to require verbal processing.

### TASK DEVELOPMENT

Because all four of these tasks are described in detail in the literature, no specific development information will be given. Instead, some background is provided for each task.

#### Sternberg Memory Search

The Sternberg task (3,4) probably is the most thoroughly documented cognitive test in existence today. Extensive baseline data exist, and standard values have been established for its parameters. Additionally, the task is sensitive to the effects caused by some toxic substances, such as lead (27).

#### Neisser Visual Search

This test (28) was developed using the same approach and concepts as the Sternberg task. It has been used much less extensively in both basic and applied research than the Sternberg task. No standardized version of this test exists. Consequently, no baseline data are available.

#### Mental Rotation

The mental rotation task is a relatively new cognitive test that was developed and popularized by Shepard and Cooper (a good overview of this work is given in reference 29; see also 30 and 31). Like the Sternberg task, this is a theoretically well developed test that is supported by a comprehensive and thorough body of literature. Unlike the Sternberg task, however, no standard values of its parameters are available because the rates of rotation obtained in the experiments are strongly affected by the familiarity of the object (i.e., letters versus geometrical shapes) and the type of rotation required (two- or three-dimensional). Very little is known about the general robustness of this task. Cooper and Shepard (31) maintain that there are large and consistent individual differences in rotation rates although this assertion has not been tested with large populations. Preliminary testing conducted at NAMRL indicates that the use of low resolution displays may seriously degrade performance. Additionally, instructions for this test seem to be particularly difficult to develop.

### Choice Reaction Time

The choice reaction time task is one of the oldest tests in psychology and has been studied for well over 100 years (32). This task appears to be a great deal more robust than the other three tasks described in this section. That is, the basic linearity of the function relating correct reaction time to the amount of information transmitted is affected less by methodological variations than comparable functions for the other three tasks. The major drawback to the task is that no real baseline data are available; correct reaction times from this task are affected by the stimulus and response modalities, the stimulus domain, the configuration of the response device for manual responses (described below), and practice. Thus, even though the general effect of many experimental variables on the reaction time function is known, observed correct reaction times cannot be predicted very accurately.

## DEPENDENT MEASURES

One major characteristic of the four tasks described above is that they all measure rate of information processing. Thus, for all four of these tasks, a linear regression is calculated using the raw correct reaction time scores from various conditions (degrees of stimulus rotation for the mental rotation task or set size for the other three tasks). The major dependent variable for all of these tasks is the slope of the regression equation although the intercept is also of both practical and theoretical importance.

## PRACTICAL PROBLEMS

A number of practical problems occur with these tasks because slope is the dependent variable of interest. The primary problem involves practice. Subjects must receive relatively extensive practice on these tasks to produce data that are fit well by linear regression. Enough time may not be available to allow sufficient practice in an applied situation. Even if the data are described adequately, the slopes may not be stable, and even more practice may be required. (See reference 33 for a discussion of reliability problems associated with the use of slopes.)

A second practical problem concerns the set sizes that can be investigated for the Sternberg, Neisser, and choice reaction time tasks. Because human short-term memory is limited, the number of items to be held in memory in the Sternberg and Neisser tasks is usually limited to six. However, some normal adults cannot retain six items, and occasionally the standard deviation of correct reaction times at the six-item level is much larger than at the other levels. This can cause problems with subsequent statistical analyses. Therefore, the investigator may want to limit the maximum number of items to be held in memory to five.

The comparable problem for the choice reaction time task is slightly different. Because all of the items may not be held in short-term memory at the same time, this task does not appear to be as affected by memory limitations as the Sternberg and Neisser tasks. The major limitation for this task concerns the method of making a response. If the investigator wants the subject to respond manually to the stimuli, the number of distinct responses is limited to 10. Normally, this implies that the maximum number of stimuli will also be limited to 10 unless the investigator wants to examine many-to-one mappings, which involve other considerations. On the other hand, if the investigator can allow the subject to respond vocally, the number of responses is theoretically unlimited.

Manual responses for choice reaction time tasks also involve two other related problems. If the task requires more than four different responses, the response device must be configured either to allow movement of the finger of the dominant hand or use of the nondominant hand. The nondominant hand normally produces reaction times somewhat slower than the dominant hand, introducing a bias into the data. Similarly, the ring and little fingers produce responses that are longer than those produced by the index and second fingers, particularly for the nondominant hand. Currently, only one accepted technique can eliminate this bias: use both hands and analyze only the data from the index fingers of each hand. The problem with this technique is that only a fraction of the responses emitted by the subject are used. To obtain good estimates of various parameters, the subject must make many more responses than if the data from all the fingers were analyzed.

A number of investigators have chosen to circumvent the problems related to speed of response by allowing the subject to respond using only the index finger of the dominant hand. Typically, the subject keeps the index finger on a 'home' key and moves it to the response keys. This introduces a travel time (distance) that is added to the true reaction time. If the response keys are all the same distance from the home key, the travel time should be a constant, and no bias is introduced. However, for some common response devices, such as a 4 by 4 matrix keypad, the travel time may not be constant. In this situation, either a large percentage of the responses must be excluded from the analyses, or extensive baseline data must be collected to obtain estimates of the travel time.

All four of these tests also have problems with the zero-choice or the zero-rotation situation. Usually, the linear regression equation fitting the correct reaction time scores to the set size, number of alternatives, or degrees of rotation accounts for a large percentage of the variance, typically more than 70%. However, if the correct reaction time scores from the set size 1, 1 alternative, or 0 degrees of rotation condition is added to the equation, the percentage of variance accounted for by the equation frequently drops. Visual inspection of the data usually reveals that the additional data point has a larger mean than would be predicted from the previous data points. To date, no explanation for this finding has been generally accepted, which indicates a lack of knowledge about choice or rotation situations versus no choice or no rotation situations. From a practical point of view, the investigator should calculate two equations for the experimental test. One equation should use all the available data;

the other should exclude the set size 1, 1 alternative, or zero degree rotation condition. The investigator should use the equation that explains the most variance.

Finally, one major methodological problem exists for an investigator who wishes to use the Sternberg task. This problem concerns the type of mapping, varied or constant, to be used. For applied research, the constant mapping procedure is normally used because more data can be collected in a given period of time. However, extensive practice with the constant mapping procedure can lead to 'automatic processing' (34,35). If automatic processing occurs, the slope of the function relating correct reaction time to set size becomes zero, indicating an infinitely fast rate of processing.

# 7. HIGHER PROCESSES

This chapter describes tests that are assumed to assess more complex cognitive functions than those assessed by the rate-of-information-processing tests. Tests of higher cognitive processes generally resemble real-world activities and, therefore, are of interest to investigators examining applied problems. Many tests of higher processes are currently available. Three of the most common are described below. A fourth section on task combinations is included because of the recent interest in assessing time-sharing skills and abilities.

## MENTAL ARITHMETIC

### Overview

Mental arithmetic tasks are probably the most frequently used higher processes tests. Using the second classification scheme, which is based on Wickens' Model, these tasks require verbal processing code resources and early rather than late resources. They may require either visual or auditory resources and either manual or vocal resources depending on the implementation of the task. Using the third classification scheme, these tasks all require the use of both short- and long-term memory.

Mental arithmetic tasks are used frequently in performance batteries for several reasons. One reason is that they have high face validity. That is, mental arithmetic is required in many real-world activities. By including a mental arithmetic task in a battery, the investigator appears to be examining relevant skills and abilities. A second reason for including mental arithmetic tasks is that almost all adult subjects have the necessary skills to perform at least simple versions of this task and to understand the relevant instructions quickly. A third reason for the popularity of these tasks is their diversity. Stimuli can be presented either auditorily or visually for many versions of these tasks, and subjects can respond manually or vocally. Additionally, mental arithmetic tasks vary greatly in difficulty and complexity. Thus, the investigator has a wide range of potential tasks available.[9]

### Task Development

Typically, the most difficult problem facing an investigator who wants to include a mental arithmetic task in a performance battery is selecting a task that is appropriate for the subject population. To identify such a task, the investigator should examine the difficulty of the mental arithmetic tasks under consideration carefully. One factor that affects the difficulty is the amount of information the subject must remember to perform the task. For example, some tasks require the subject to perform multiple

[9] Another reason for using mental arithmetic tasks is that at least some of them appear to follow the standard stage model of human information processing developed by Sternberg (3). See Ashcraft and Battaglia (36) for a discussion of a mental arithmetic task that follows such a model. If this is the case, then the additive factors logic can be applied to at least some mental arithmetic tasks.

operations on a pair of numbers. To do this, the subject must remember the sequence of operations to be performed as well as the results of the immediately preceding operation. The more numbers and operations, the more difficult the task, and the less appropriate the task becomes for some subject populations.

Task difficulty directly affects the probability of detecting experimental effects by influencing both the number of responses emitted in a given period and their accuracy. Thus, an easy mental arithmetic task, such as adding a constant to the stimulus, may result in a large number of responses in a given period of time while a more difficult task will result in relatively few responses. If the investigator needs numerous responses in a short period of time to detect an effect, then an easy mental arithmetic task may be preferred. Task difficulty also affects the error rate. If the task is too difficult for the subjects, they may respond at the chance accuracy level. Consequently, any performance decrements caused by experimental factors may be impossible to detect.

A list of some of the common mental arithmetic tasks is given below with at least one reference per task. The purpose of this list is to provide the reader with some idea of the types of tasks available.

**Addition of a Constant.** This is the simplest type of mental arithmetic task. The subject is required to add a constant to a number or set of numbers. Because this task is so simple, it is frequently paced or performed concurrently with another task (37).

**Running Difference.** The subject subtracts the most recent digit from the preceding digit and enters the difference. As soon as the subject responds, a new digit is presented, which the subject subtracts from the immediately preceding digit (13,38,39). This task may also be performed as a running sum task, in which the subject adds the most recent two digits and reports the sum (40).

**Two-digit Addition.** The subject is presented with two two-digit numbers to be added. The subject reports the sum and immediately is presented with two more two-digit numbers (41).

**Complex Operations.** These tasks require the addition or subtraction of multiple digit numbers that may be displayed either vertically or horizontally (see references 42 and 43). Tasks requiring multiple-digit division or multiplication also fall into this group although examples of these types of tasks are less frequent (44).

**Criteria Verification.** Three one-digit numbers are presented as a sequence to the subject. The subject must decide if the sequence meets one or more criteria. For example, Griffiths and Boyce (45) had subjects determine if a sequence met one of two criteria: (1) the first digit was the largest, and the second digit was the smallest; or (2) the third digit was the largest, and the first digit was the smallest. The subject made one response if the digits met the criteria and another response if they did not.

**Multiple Operations.** Many varieties of tasks use multiple arithmetic operations. For example, Morgan et al. (46) required subjects to add two three-digit numbers and then subtract a third three-digit number from the sum. Chiles and Jennings (47) required the same sequence of operations using two-digit rather than three-digit numbers. As another example, Williges (48) had subjects perform a mental arithmetic task with four steps. The subjects were presented with two digits and added five to the smaller of the two. After the addition, the subject compared the two digits and doubled the smaller digit. Next, the subject subtracted the smaller from the larger and compared the result with a criterion value. If the result exceeded the criterion, the subject made one response. If the result did not exceed the criterion, the subject made a second response.

### Dependent Measures

Accuracy scores have traditionally been the primary performance measures for mental arithmetic tasks. These include percentage correct, number of correct responses in a specified time, and so forth. More recently, reaction times have become common measures of performance although they are almost always used in conjunction with some type of accuracy score.

### Practical Problems

The investigator must be concerned with few practical problems beyond those associated with task selection. Mental arithmetic tasks are easy to program and debug and can be presented using relatively poor quality graphic systems. Thus, the only practical problem concerns measuring reaction time. Many mental arithmetic tasks require multiple digit responses that must be entered using some type of keypad. Subjects usually require some type of familiarization with the keypad, and the investigator must allow adequate time for this process.

The investigator may require all subjects to enter responses in a standardized fashion. That is, all subjects may be required to place their fingers on a specific row of keys or to use only one finger to respond. In contrast, some investigators feel that standardizing any aspect of the data entry process reduces the face value of the test and, consequently, do not specify how the subject should enter the responses.

## VIGILANCE TASKS

### Overview

Vigilance tasks vary widely in their stimulus mode, response mode, and the number of stimulus sources the operator must monitor. The salient characteristic of all of these tasks is that they require sustained attention for a relatively long time period, typically at least 50 min. These tasks are difficult to describe using either of the classification schemes. Using the second classification scheme based on Wickens' Model, these tasks require early rather than late processing resources. They cannot be classified in terms of the stimulus or response resource requirements because stimuli may be presented either auditorily or visually, and responses may be made either manually or vocally. Similarly, they cannot be described in terms of their code of processing resources because they may require either spatial or verbal processes. These tasks never have been described in the cognitive literature, but they seem to require short- and long-term memory and pattern recognition processes.

There are many reasons to include a vigilance task in a performance battery. One of the most important is that these tasks simulate many important real-world activities. Thus, the investigator can increase both the applicability of the results and the face validity of the battery by including vigilance tasks in the battery. Another important reason is that this type of task has been thoroughly investigated; at this time, more than 1500 studies have appeared in the open literature. Additionally, several excellent literature reviews have been published, such as Craig (49) and Parasuraman (50). The investigator may be able to decrease substantially the amount of pretesting required by using the available information to narrow the range of several task parameters, such as the intensity of the stimulus and the number of events per hour.

There are also a number of practical reasons for including vigilance tasks in a battery. These tasks are easy to program and require little central processing capacity. The speed of the central processing unit is of little concern, and almost any microcomputer can be used successfully. Additionally, very little hardware is needed for the subject's responses; in most cases, the subject responds manually by pressing a key to indicate a signal. The stimuli also can usually be presented using very simple hardware. For visual stimuli, only primitive graphics are normally necessary; for auditory stimuli, a pure tone generator and a white noise generator are often sufficient unless the investigator wants to simulate a specific real-world task, such as sonar operation.

Vigilance tasks also have two drawbacks. The first, and the most serious, is that subjects often find these tasks boring. As a result, they may fall asleep during the testing session, decide to stop monitoring the task for awhile, or adopt some new way of responding, such as pressing keys with their elbows or feet. They may also decide that the task is too tedious to be tolerated and quit the experiment. If any of these situations occurs, the experimenter may have to discard a large amount of data. The second major drawback concerns training the subject. Most normal adult subjects understand vigilance instructions, but many have difficulty learning to detect signals reliably. Subjects may repeat the training session several times before reaching the performance criteria necessary to begin the testing session. A few subjects never reach the criteria. Thus, the investigator must allow for lengthy training sessions and for replacing subjects who cannot reach criteria.

## Task Development

To develop a vigilance task, the investigator must determine the stimulus mode, the response mode, the number of stimulus sources, and type of discrimination required (successive versus simultaneous) by the task. The choice between visual and auditory stimuli appears to be completely arbitrary unless the investigator intends to apply the results to a specific real-world activity. Vigilance tasks generally require manual responses although vocal responses are theoretically as acceptable as manual responses. Manual responses probably have been used almost exclusively to date simply for convenience. Subjects may be required to monitor one display or several displays for a signal. The vast majority of the literature has examined single-source monitoring, but again the choice may depend on the desired applicability of the data.

Finally, the investigator must decide between successive versus simultaneous discrimination of signals and nonsignals. Successive discrimination requires that a stimulus change repetitively in two ways. The most common change is usually defined to be a nonsignal; the less common type of change is defined as a signal. For example, Williges (51) had an abstract geometric figure change brightness periodically from 5 to 4 fL (17.13 to 13.70 cd/m$^2$). A 1.3-s period of dimness was a signal; a 1.7-s period of dimness was a nonsignal. Simultaneous discrimination requires the presence of both the signal and the nonsignal either in the same stimulus or at the same time. A good example of simultaneous discrimination is detecting a weak pure tone (the signal) against a background of white noise (the nonsignal). The major difference between simultaneous and successive discrimination from an information processing standpoint is that successive discrimination imposes a short-term memory load on the subject that is not required by simultaneous discrimination; the subject must remember the characteristics of the signal to compare it with a nonsignal in the successive discrimination situation.

The choice between these two types of discrimination again appears to be dictated by the applicability of the results. If the investigator wants the data to be immediately applicable to a specific task, then the experimental task must use the same type of discrimination. If simulating a real-world task is not necessary, the choice of discrimination is arbitrary. However, data obtained using the simultaneous discrimination paradigm require more time-consuming analyses than those obtained using successive discrimination (see reference 50 for a succinct discussion of these problems).

## Dependent Measures

Traditionally, vigilance task performance is measured by the probability of detecting [P(D)] a signal. To obtain this measure, the experimental session is divided into a number of equal time periods, and the probability of detecting signals presented in each period is calculated. In most cases P(D) decreases across the time periods. This decrease is called the "vigilance decrement." False alarms (FA) are also often calculated for each time period, and occasionally the average reaction time for correct signal detections is obtained. Some investigators (52) maintain that calculating P(D) and FA over a period of time, such as 10 or 15 min, provides performance measures that are so crude as to be misleading. These investigators advocate a more fine-grained approach in which the probability of detecting each signal is calculated, and statements about performance are based on trends in detection evident across signals. This approach was useful in several experiments but was never widely accepted.

Both the traditional approach based on P(D) and the fined-grained approach have been replaced for the most part by Signal Detection Theory (SDT). McNicol (53) gives a good intuitive explanation of this theory with many practical examples. Green and Swets (54) provide a more rigorous explanation. The major reason for adopting SDT is that this theory separates change in the subject's ability to discriminate a signal from a nonsignal from the subject's willingness to respond "signal" or "nonsignal." Thus, SDT is an extremely powerful theory that has provided many insights into vigilance behavior.

Estimates of the subject's ability to discriminate a signal from a nonsignal are reflected in a dependent measure referred to as "d'." The subject's willingness to respond is reflected in a measure referred to as "beta." To calculate d', the P(D) and the number of FAs must be calculated for each time interval of interest. To calculate beta, the subject must be told the a priori probability of signals and nonsignals. Additionally, the subject should be given a payoff matrix with specified rewards for each correct detection of a signal and each correct rejection of a nonsignal and penalties for each missed signal and FA. Thus, to use SDT, the investigator must provide the subject with more information than is typically given when the traditional data analysis approach is followed. Correct reaction times may be obtained in addition to d' and beta, but these are secondary measures.

## Practical Problems

A few practical problems should be considered before including a vigilance task in a performance battery, but these are neither as numerous nor as serious as those associated with the rate-of-information-processing tasks. As noted earlier, some subjects have a great deal of difficulty reaching training criteria, and a few subjects never reach the criteria. Such difficulties imply that the investigator must allow a large amount of training time and must have more than the minimum number of subjects available.

A more serious problem pertaining to training concerns the ratio of the signals to nonsignals presented during the training session. Typically, investigators have used signal-to-nonsignal ratios in training that are much higher than those encountered in the testing sessions. Colquhoun and Baddeley (55,56) have demonstrated that subjects trained under a signal-to-nonsignal ratio that is higher than the ratio used in the testing session show larger vigilance decrements than subjects trained with the same ratio used in the testing session. Craig and Colquhoun (57) suggest that much of the observed vigilance decrement is caused by training with inappropriate signal-to-nonsignal ratios. Craig's (58) analysis of data in the open literature supports this assertion.

The primary reason for using training ratios that are higher than those of the testing session is to provide the subject with sufficient practice in distinguishing signals from nonsignals in as short of time as possible. If the investigator uses the same ratio in the testing and training sessions, the length of the training session must be increased to determine if the subject can detect signals reliably. Thus, investigators have had to choose between inappropriate training ratios and long training sessions. Recently, Williams (59) proposed a new training technique based on probability matching, which uses the appropriate signal-to-nonsignal ratio in a relatively short training session. Williams demonstrated that this technique could eliminate some, but not all, of the vigilance decrement.

Another problem concerns the payoff matrix used to determine beta, one of the SDT measures. Presumably, subjects must be given a payoff matrix to support the SDT assumptions underlying the calculation of beta. It may not be possible, however, to pay some subject populations, such as active duty military personnel. Only one study (60) compared the performance of subjects receiving a cash payoff matrix with those receiving no payoff matrix.

No difference in performance was found for these two groups; however, Wiener was not using SDT to analyze his data.

The detectability of the signal as compared to the nonsignals must also be given serious consideration. If the signal is obvious, the subjects will detect all of the signals and will make no FAs. Although such data can still be analyzed using the traditional approach, beta will be mathematically indeterminate. The possible loss of one of two SDT measures should be carefully considered when the signals and nonsignals are selected.

Finally, the investigator should be aware of some criticism of laboratory vigilance tasks. The four major criticisms are that the length of the testing session is too short, too few sessions are administered, the signal rate is too high to simulate any real-world activity, and naive subjects are used. All of these criticisms are justified to some extent. Only one study (61) used a signal rate typical of many real-world systems (one signal per week) and examined subject behavior over a 6-month period. The reasons for using a high signal rate, collecting data during a few short testing periods, and employing naive subjects are to generate sufficient data for analysis and to keep costs down. Thus, economic constraints and data analysis considerations may reduce the applicability of any laboratory task to real-world behavior despite the best intentions of the investigator.

## TRACKING

### Overview

Tracking tasks can be described using Wickens' Multiple Resources Model. They require response rather than perceptual or central resources, spatial rather than verbal processing code resources, manual rather than vocal response resources (with a few exceptions), and visual rather than auditory stimulus resources (also with some exceptions). Using the third classification system, the general consensus is that tracking tasks require spatial processes and may require spatial short-term memory.

There are two primary reasons for including a tracking task in a performance battery. The first is to increase the applicability of the data from the battery. This is a legitimate reason if the battery is designed to examine skills and abilities of activities that require tracking. The investigator, however, should consider the relation between the real-world activity and any potential laboratory tasks carefully before deciding to add a tracking task to the battery. Bartram et al. (62) demonstrate that tracking tasks that differ in the display type, number of dimensions of movement of the cursor, or the allocation of controls to the limbs correlate poorly. Thus, data obtained from a laboratory tracking task may not be applicable to a real-world activity if the two differ on any of the dimensions noted by Bartram et al. There is also reason to suspect that differences in other parameters, such as control order, will lower between-task correlations.

The second reason to include tracking tasks in a battery is to obtain measures of skills and abilities that are not assessed by any other type of task. This implies that performance on tracking tasks should correlate poorly with performance on other tasks. Interestingly, few data show that the computer-generated tracking tasks used by applied psychologists do correlate poorly with the skills and abilities required by other tasks.

Only one study (63) used a computer-generated tracking task to examine the structure of human abilities. (This experiment was concerned primarily with the existence of a general timesharing ability.) This tracking task performed singly and in combination with the other tasks of the battery had significant loadings on only one factor of the solution; no other tasks had significant loadings on this factor. These results seem to indicate that tracking requires unique skills and abilities, but more research is needed before any firm conclusions can be drawn.

## Task Development

Five major issues concerning the development of a tracking task are given below. Good tracking tasks are difficult to develop and almost impossible to debug. Many good tracking tasks cannot be programmed on some micro- and minicomputers because of their relatively slow processing speed and limited memory. Investigators with no experience constructing a tracking task should obtain the software from a reliable source, if possible, and consult with knowledgeable individuals about the necessary equations.

An investigator must decide to use either a pursuit or a compensatory display early in the development of the task. Pursuit displays present the input (command) information on one display element and the system output on a second display element. Compensatory displays simply present the difference between the input information and the system output on one display element. Pursuit displays tend to result in better performance than compensatory displays for several reasons (9). Chief among these is that the operator can view the input directly and learn any regularities. Additionally, the operator can distinguish between the changes caused by the input and changes caused by control responses.

Most traditional laboratory tasks use compensatory displays because it is easier to model an operator tracking a compensatory than a pursuit display. This reason is usually irrelevant to investigators constructing performance batteries. A second reason for using compensatory displays is that at one time a compensatory display may have been easier to construct than a pursuit display. This consideration is also irrelevant if the tracking tasks will be implemented on micro- or minicomputers. Thus, the decision between the two display types may be completely arbitrary.

If a compensatory display is selected, the investigator must be concerned with the point at which the forcing function[10] is injected into the tracking task. There are two primary points at which the forcing function can be introduced. In Figure 1a, the forcing function is introduced after the system dynamics have transformed the operator's response. In Figure 1b, the forcing function is first added to the operator's response, and then the sum is acted upon by the system dynamics. The configuration shown in Figure 1b is not preferred because if higher-order system dynamics are used, they may effectively filter the forcing function at higher frequencies. As a result, the operator may experience a forcing function that is considerably different from the one generated by the computer.

The investigator must also decide the control system order. The order of a system refers to the number of time integrations performed on the control responses. For example, no integrations are performed in a zero-order (position) system. Thus, moving the control stick to a given position always results in the cursor moving to a given position on the display. One time integration is performed in a first-order (rate) system. Thus, moving the control stick to a given position results in a specific velocity of the cursor. Two time integrations are performed in a second-order (acceleration) system. Moving the control stick to a specific location results in a specific acceleration of the cursor. Although higher-order systems can be constructed, they are of little interest to an investigator developing a performance battery.

Generally, zero-order control systems result in tasks that are easy enough to be boring. Consequently, they are of little interest to anyone developing a tracking task for a normal adult population. Second-order control systems are too difficult for a normal adult population if the bandwidth of the forcing function exceeds 0.4 Hz; data obtained using these types of systems tend to show nonlinearities (see reference 64). Therefore, an investigator constructing a performance battery can choose between a first-order system, a second-order system with a limited forcing function, or a system consisting of a combination of a first- and second-order system. Such systems may be constructed either by adding the weighted outputs of a first- and a second-order system or by placing a first-order system in series with a first-order lag with an adjustable time constant. Although there is little to aid in selecting between these three control orders, second-order systems are used more infrequently than first-order or weighted combination systems in most types of human performance research.

Another major decision concerns the forcing function used as the input to the system. The investigator must first decide between using band-limited noise and a function consisting of the sum of sine waves. The major disadvantage of band-limited noise is that control theory analyses of the operator's behavior are more difficult to perform. The major advantage of band-limited noise is that it requires less complicated algorithms to generate than sum-of-sine-waves function. Thus, if the investigator must use computers with limited (less than approximately 256K) random access memory (RAM), it may be impossible to generate a sum-of-sine-waves forcing function, leaving band-limited noise as the only alternative.

Constructing a sum-of-sine-waves forcing function requires the investigator to decide the number and frequency of the sine waves composing the function. At least three nonharmonically related sine waves must be used to achieve a random-appearing function. The total number of sine waves used to construct the function is usually limited by the size of RAM and the speed of the central processing unit. I use a forcing function consisting of nine sine waves, a rather common number.

After deciding on the number of sine waves to be used, the investigator must select the range of the sine waves. The range of usable frequencies is extremely limited; McRuer and Jex (64) demonstrated that performance is linearly related to the frequency of the forcing function up to approximately 1 Hz for zero- and first-order systems and 0.4 Hz for second-order

[10] The term "forcing function" usually refers to a function that is applied directly to the system dynamics. The terms "input" and "command" usually refer to a function applied directly to the subject's display.

systems. Beyond these values, performance becomes increasingly nonlinear. After the investigator has decided on the number of sine waves and the bandwidth, the specific sine waves are either selected at random or so that their frequencies are approximately equally spaced when plotted on a logarithmic scale.

Occasionally, investigators have included higher-frequency (above 0.4 Hz) sine waves in the forcing function. Such frequencies are included either when the investigator wants to study human tracking behavior, in general, or when performance on a specific system with high-frequency inputs is being examined.

Even inexperienced operators can learn a short portion of a forcing function if it is repeated consistently. The easiest way to avoid any learning effect is to choose a starting point randomly for each sine wave at the beginning of each trial. The forcing function for a given trial is constructed by adding the sine waves, beginning with the randomly selected starting point of each sine wave.

Finally, the investigator must consider one problem that occurs with inexperienced operators using a compensatory display: The operator sometimes moves the control stick in the wrong direction, causing the cursor to be displaced as far as possible on the display. Frequently, the operator does not realize the mistake and allows the cursor to remain maximally displaced for several seconds. The question confronting the investigator concerns the system response after the operator realizes the mistake and moves the control in the correct direction. In some systems, the cursor responds as soon as the control stick is moved in the correct direction. In other systems, the cursor does not move for some pre-established time after the control stick is moved correctly, penalizing the operator for not recognizing the mistake immediately. The problem with the second system is that an inexperienced operator may become even more confused when the cursor does not respond immediately to a control movement. The investigator must decide between penalizing a serious mistake and increasing the possibility that the operator will become confused and frustrated. Most investigators feel that the probability of confusing an inexperienced operator is high and have constructed systems that respond immediately to the correct movement of the control stick after the cursor has been maximally displaced.

## Dependent Measures

Tracking tasks have two major classes of dependent measures: classical error measures, such as RMS and average absolute error, and performance measures derived from control theory, such as gain and phase lag. The classical error measures are discussed in detail by Poulton (9). Currently, the most commonly used error measures are RMS and average absolute error. Wickens and Gopher (22) describe control theory measures, which give a more fine-grained analysis of performance.

## Practical Problems

One of the most serious practical problems, inadequate RAM, has been noted earlier. Another of the serious problems, the inability to debug the program, has also been mentioned but warrants further comment. In contrast to all of the other tasks discussed in this document, a tracking task is almost impossible to test by performing it. Some parts of the control dynamics can be tested by using known forcing functions, such as brief pulses, as input and recording the response of the system dynamics. This approach is time-consuming and often requires special equipment. Additionally, it only tests the system dynamics.

Tracking tasks require good graphics systems. Systems with poor resolution cause the cursor to hop around the display instead of moving smoothly. The phosphor of the display screen also may present problems; long-persistence phosphors may blur the cursor as it moves. This is very distracting to the operator and may induce visual fatigue.

Two other issues need to be discussed briefly. One concerns the system gain. For most tracking systems, the rule of thumb has been to set the gain so that the operator can overcome the maximum amplitude of the forcing function. Usually, the programmer can determine the maximum amplitude, but the gain should be subsequently testeo to ensure that the system dynamics are not too responsive.

The second issue concerns the control sticks. The investigator must first decide between isometric (no movement) or displacement (movement) control sticks. The type of control stick interacts in a complex manner with several parameters of the tracking system to affect the subject's performance. A description of these interactions is given in Poulton (9) and Repperger and Levision (65). The investigator may want to consider these interactions if the absolute level of the subject's performance is of concern. The general advantages and disadvantages of each type of control stick are discussed by Frost (66).

The investigator will discover quickly that the price of control sticks varies from less than $100 to more than $2000. Generally, the most expensive sticks are displacement sticks with a guaranteed linear relation (usually with less than 1% error) between the angle of displacement and the voltage output of the stick. Most investigators will not need this degree of accuracy between the stick displacement and the voltage output unless they plan to perform a control theory analysis of the data.

Finally, the investigator should remember that displacement sticks are subject to wearing. As a result, the null position may become wider over time (that is, the angular displacement necessary to signal a response may increase with time), and the voltages resulting from the maximum angular displacement of the stick may change. To account for this wearing, the investigator may have to recalibrate the system periodically or develop a software routine that recalibrates the system automatically when activated.

## TASK COMBINATIONS

### Overview

This section differs from the preceding ones in that it is concerned with task combinations in general rather than with a specific combination. This section was included because of an increased interest in multiple-task performance in exotic environments during the last 10 years. Many investigators, however, have not recognized the problems associated with constructing and measuring performance on task combinations. As a result, many studies have collected data that were either uninterpretable or unanalyzable.

Task combinations are usually described today using Wickens' Multiple Resources Model. That is, the combinations are described primarily in terms of the number and type of resources that are shared. For example, a combination might be described by stating that it required shared visual resources, separate code of processing resources, separate stage of processing resources, and shared manual response resources. Combinations may also be described using the third scheme. That is, the primary cognitive structure or processes required by each task is mentioned.

An investigator has only two reasons for including a task combination in a performance battery: (1) to measure timesharing skills and abilities, or (2) to make the data more applicable to a real-world activity that requires timesharing. Interestingly, almost 100 years of experiments have failed to isolate conclusively a general timesharing ability. Additionally, little evidence exists for more specific timesharing abilities (see reference 67 for a good literature review). Thus, at this time, no scientifically supportable reason exists for including a task combination in a performance battery to assess the effect of some variable on a timesharing ability.

In contrast, several timesharing skills have been identified (see references 23 and 68 for examples). Measuring these skills requires controlled laboratory conditions, practiced subjects, and a significant amount of statistical analysis. Even under the best conditions, the scores obtained for these skills are only crude estimates. Therefore, including a task combination in a performance battery to measure timesharing skills seems questionable. The only justifiable reason for including a combination in a battery is to increase the applicability of the data to a specific real-world activity. In this situation, the investigator should attempt to simulate the real-world activity as closely as possible to ensure that the same timesharing skills are required by the laboratory combination.

An investigator should keep in mind that data obtained from timeshared tasks are usually difficult to analyze and may require consultation with a statistician. Experimental designs requiring repeated measurement of performance on timeshared tasks usually produce the most difficult type of data to analyze; typically, this type of data violates most of the assumptions of analysis of variance. Some newer statistical techniques, such as correcting the repeated measures factors by adjusting the degrees of freedom (69), do offer solutions to some of the problems encountered with data from task combinations. These techniques do not, however, offer solutions to all of the problems likely to be encountered.

The investigator should also be aware of several other statistical problems that may be encountered. One of these concerns the type of analysis to be performed. If two or more dependent measures are obtained from one of the tasks of a combination, these measures probably will be significantly correlated. Thus, the data should be analyzed using multivariate, rather than univariate, statistics. My impression is that the dependent measures of a given task will be correlated more often when the task is timeshared than when it is performed singly. The investigator should be prepared, therefore, to use multivariate statistics.

Another problem concerns the use of derived scores, such as proportion scores. For example, one way to take single-task performance into account when analyzing multiple-task performance is to express performance on a task when it is timeshared as a percentage or proportion of performance on the same task performed alone. The problem with this approach is that derived scores often have unusual statistical properties. When derived scores are analyzed, the results are often misleading. I have conducted several analyses of timeshared data using the raw data for one analysis and derived scores for the other and found that the two sets of analyses occasionally gave contradictory results. The best advice, then, is to analyze raw scores and to test all the assumptions of the analyses carefully.

One final problem the investigator should keep in mind is that a given task combination may reach differential stability very slowly. If differential stability is necessary, the investigator should pretest carefully to establish the amount of practice required.

## Task Development

If the investigator includes a combination in the performance battery that simulates a real-world activity, then the development of the combination is relatively straightforward. Otherwise, the investigator must develop the combination. The easiest way to develop a combination is to use Wickens' Model to determine the number of resources to be shared. For example, the investigator might want to examine the effect of some variable on performance when all of the timeshared tasks require the same processing resources. Once the investigator has decided on the number of resources to be shared, the type of shared resources then must be determined. That is, if the processing code resources are to be shared, the investigator must decide to use either spatial or verbal tasks.

After determining the number and type of shared resources, the investigator must decide how to construct the tasks so that the combination has the desired characteristics. For example, suppose the investigator wanted to construct a combination with no shared resources. Should the task requiring spatial code resources use visual resources or auditory resources? Should this task use manual or vocal resources?

To construct the tasks, the investigator should consider Wickens' principle of S-C-R compatibility (see reference 70 for a good, brief summary) and the desired level of timeshared performance. Basically, the S-C-R compatibility principle states that the level of single-task performance on a given task depends on the stimulus/response configuration. More specifically, optimal performance on a task requiring spatial processing code resources occurs when the stimuli are presented visually and the subject responds manually. In contrast, optimal performance on a task requiring verbal processing code resources occurs when the stimuli are presented auditorily and the subject responds vocally.

Generally, the S-C-R compatibility principle has a less powerful influence on multiple-task performance than the number of shared resources. Nevertheless, this principle can be used to improve good multiple-task performance and worsen bad multiple-task performance. For instance, assume that an investigator wanted to construct a combination with no shared resources. According to Wickens' Model, such a combination should result in the best timeshared performance. This performance could, however, be improved by using visual stimuli and manual responses for the spatial task and auditory stimuli and vocal responses for the verbal task. Decreasing performance on a combination consisting of tasks that require the same resources is more problematic; constructing a combination that allows the subject to respond vocally to both tasks and that presents stimuli auditorily for both tasks without introducing unique constraints on the subject's behavior is difficult. Thus, the investigator must be satisfied with shared visual and manual resources regardless of the processing code used by the tasks. Nevertheless, the principle is same.

The investigator should remember that predicting the absolute level of timeshared performance of a task even when its single-task performance levels are known, is not possible. Wickens' Model only predicts that the relative performance of a task combination deteriorates as the number of shared resources increases; it says nothing about absolute levels. Thus, the investigator must obtain pretest data to determine the level of timeshared performance.

## Dependent Measures

Performance under timeshared conditions is measured using the same variables as under single-task conditions.

## Practical Problems

Only one software problem is commonly encountered. Most programs store the subject's responses in a buffer until they can be read and recorded. A problem occurs under timeshared conditions when the subject responds simultaneously to both tasks. In this situation, the program may read only the first response and clear the buffer, deleting the second response. This problem can be circumvented easily by checking for several responses in the buffer.

The only common hardware problem an investigator may encounter can also be easily circumvented. Subjects often become very frustrated under timeshared conditions. As a result, they may treat the response apparatus roughly. The investigator should ensure that all the equipment can withstand rough handling without breaking.

Two training problems are among the most serious practical problems the investigator will encounter. The first is that the relation between the amount of practice a subject receives on each task of the combination and subsequent performance on the combination is not known. Thus, establishing an efficient training schedule that results in an acceptable level of performance under timeshared conditions requires extensive pretesting. Second, several investigators have noted that about 5% of normal adults do not learn under multiple-task conditions. That is, with practice these individuals' performance never improves. The major problem with including this type of subject in an experiment is that their performance has little effect on the mean group performance calculated on a trial-by-trial basis but increases the standard deviation. Usually, the increase is large enough to cause a violation of the assumption of homogeneity of variance in subsequent analyses. Presently, these individuals cannot be identified from either their single-task performance or their early multiple-task performance, and the investigator must simply decide post hoc to include or exclude the r data from subsequent analyses.

Probably the worst problems an investigator will encounter involve controlling the subjects' priorities. Under any timeshared conditions, a subject will tend to favor one task over the other(s). This bias can be so extreme that a subject may actually ignore one task completely. The investigator's problem, therefore, is to control the subject's priorities in some way that reduces individual differences. At a minimum, the investigator can give the subject explicit priorities. This technique, of course, has it limitations; most normal adults understand equal priority instructions but may not really understand instructions such as "Give 45% of your attention to Task A and 65% to Task B." The traditional primary-secondary task designation also rarely works. This method requires the subject to maintain single-task performance levels on the primary task. The vast majority of experiments using this technique show that primary task performance detoriorates in the presence of the secondary task.

The best way to control a subject's task priorities is to use terminal KR as described in the Knowledge of Results Section (Chapter 3), and, if necessary, on-line KR. Both of these techniques have shortcomings, as noted earlier, but they are currently the best methods available.

Finally, the investigator should consider the problem of fatigue. Performing under timeshared conditions for any period of time is usually fatiguing. Unless the purpose of the research is to examine the effect of fatigue on timeshared performance, the investigator must schedule periodic breaks. These breaks, however, may not be sufficient to ensure acceptable performance, particularly if the subject is also frustrated with his/her performance. The investigator may want to consider some type of incentive based on an individual's performance to compensate for the effects of fatigue and frustration.

## 8. CONCLUDING REMARKS

As stated in the preface, this monograph is directed towards the construction of a battery of human information processing tests for use in repeated-measures testing situations. It attempts to provide some guidelines for individuals who have little knowledge of human information processing and the problems associated with computerized testing. It is not an exhaustive discussion of all the issues and tests an investigator must consider in constructing a battery. Rather, it attempts to convey the type of questions an investigator must bear in mind while constructing the battery. It also attempts to alert the reader to some of the common pitfalls associated with repeated-measures testing.

Some readers may find this monograph too abstract for dealing with the complicated and interrelated problems associated with the development of a battery. Such readers may find the Army Air Force Aviation Psychology Program Research reports valuable. This series, documenting the World War II aircrew selection efforts, provides a detailed, step-by-step commentary on the development of the battery. The fourth report on apparatus testing (71) may be particularly useful.

Finally, the reader should remember that the methodology for repeated-measures testing and the associated statistical tests are still being developed. New techniques for analyzing repeated measures data probably will be available in a few years that will be more powerful than the techniques currently available. The development of increasingly powerful microcomputers will increase both the number and the types of available tests. Thus, investigators should be alert for new developments in methodology, statistical techniques, and hardware that could be used in repeated-measures batteries.

# REFERENCES

1. Perez, W.A., Masline, P.J., Ramsey, E.G., and Urban, K.E., Unified Tri-service Cognitive Performance Assessment Battery: Review and Methodology, AAMRL-TR-87-007, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, pp. 1-338, March 1987.

2. Oborne, D.J. and Rogers, Y., "Interactions of Alcohol and Caffeine on Human Reaction Time." Aviation, Space and Environmental Medicine, Vol. 54, pp. 528-534, 1983.

3. Sternberg, S., "The Discovery of Processing Stages: Extensions of Donders' Method." Acta Psychologica, Vol. 30, pp. 276-315, 1969.

4. Sternberg, S., "Memory Scanning: New Findings and Current Controversies." Quarterly Journal of Experimental Psychology, Vol. 27, pp. 1-32, 1975.

5. Bradley, J., "Practice to an Asymptote?" Journal of Motor Behavior, Vol. 1, pp. 285-295, 1969.

6. Bittner, A.C., Jr., Carter, R.C., Kennedy, R.S., Harbeson, M.M., and Krause, M., "Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 Measures." Perceptual and Motor Skills, Vol. 63, pp. 683-708, 1986.

7. Kirk, R.E., Experimental Design: Procedures for the Behavioral Sciences, Brooks/Cole, Belmont, CA, 1968, pp. 1-577.

8. Simon, C., Methods of Handling Sequence Effects in Human Factors Engineering Experiments, Report No. P74-541A, Hughes Aircraft Company, Culver City, CA, December 1975.

9. Poulton, E.C., Tracking Skill and Manual Control, Academic Press, New York, NY, 1974, pp. 1-427.

10. Poulton, E.C. and Freeman, P.R., "Unwanted Asymmetrical Transfer Effects with Balanced Experimental Designs." Psychological Bulletin, Vol. 66, pp. 1-8, 1966.

11. Poulton, E.C., "Influential Companions: Effects of One Strategy on Another in the Within-subjects Designs of Cognitive Psychology." Psychological Bulletin, Vol. 91, pp. 673-690, 1982.

12. Damos, D.L., "The Effect of Asymmetric Transfer and Speech Technology on Dual-task Performance." Human Factors, Vol. 27, pp. 409-421, 1985.

13. Damos, D.L. and Lyall, E.A., "The Effect of Varying Stimulus and Response Modes and Asymmetric Transfer on the Dual-task Performance of Discrete Tasks." Ergonomics, Vol. 29, pp. 519-533, 1986.

14. Johansson, G., "Psychoneuroendocrine Correlates of Unpaced and Paced Performance." In G. Salvendy and M.J. Smith (Eds.), Machine Pacing and Occupational Stress, Taylor & Francis, London, England, 1981, pp. 277-286.

15. Drury, C. and Coury, B., "Stress, Pacing, and Inspection." In G. Salvendy and M.J. Smith (Eds.), Machine Pacing and Occupational Stress, Taylor & Francis, London, England, 1981, pp. 223-229.

16. Salvendy, G. and Humphrey, A., "Effects of Personality, Perceptual Difficulty, and Pacing of a Task on Productivity, Job Satisfaction, and Physiological Stress." Perceptual and Motor Skills, Vol. 49, pp. 219-222, 1979.

17. Holding, D., Principles of Training, Pergamon Press, Oxford, England, 1965, pp. 1-156.

18. Salmoni, A., Schmidt, R., and Walter, C., "Knowledge of Results and Motor Learning: A Review and Critical Reappraisal." Psychological Bulletin, Vol. 95, pp. 355-386, 1984.

19. Gopher, D. and North, R., "The Measurement of Attention Capacity Through Concurrent Task Performance with Individual Difficulty Levels and Shifting Priorities." In E.L. Saenger and M. Kirkpatrick, III (Eds.), Proceedings of the Human Factors Society 18th Annual Meeting, 1974, pp. 480-485.

20. Gopher, D. and North, R., "Manipulating the Conditions of Training in Time-sharing Performance." Human Factors, Vol. 19, pp. 583-593, 1977.

21. North, R. and Gopher, D., "Measures of Attention as Predictors of Flight Performance." Human Factors, Vol. 18, pp. 1-13, 1976.

22. Wickens, C. and Gopher, D., "Control Theory Measures of Tracking as Indices of Attention Allocation Strategies." Human Factors, Vol. 19, pp. 349-365, 1977.

23. Damos, D. and Wickens, C., "The Identification and Transfer of Timesharing Skills." Acta Psychologica, Vol. 46, pp. 15-39, 1980.

24. Wickens, C., "The Structure of Attentional Resources." In R.S. Nickerson (Ed.), Attention and Performance VIII, Lawrence Erlbaum, Hillsdale, NJ, 1980, pp. 239-257.

25. Moerland, M., Aldenkamp, A., and Alpherts, W., "A Neuropsychological Test Battery for the Apple II-E." International Journal of Man-machine Studies, Vol. 24, pp. 453-467, 1986.

26. Frick, R., "Testing Visual Short-term Memory: Simultaneous Versus Sequential Presentations." Memory and Cognition, Vol. 13, pp. 346-356, 1985.

27. Smith, P. and Langolf, G., "The Use of Sternberg's Memory-scanning Paradigm in Assessing Effects of Chemical Exposure." Human Factors, Vol. 23, pp. 701-708, 1981.

28. Neisser, U., "Decision Time Without Reaction Time: Experiments in Visual Scanning." American Journal of Psychology, Vol. 76, pp. 376-385, 1963.

29. Shepard, R. and Cooper, L., Mental Images and Their Transformations, MIT Press, Cambridge, MA, 1986, pp. 1-364.

30. Cooper, L., "Mental Rotation of Random Two-dimensional Shapes." Cognitive Psychology, Vol. 7, pp. 20-43, 1975.

31. Cooper, L. and Shepard, R., "Transformations on Representations of Objects in Space." In E. Cartarette and M. Fiedman (Eds.), Handbook of Perception, Vol. VIII, Academic Press, New York, NY, 1978, pp. 105-146.

32. Donders, F.C., "Over de Snelheid van Psychische Processen." Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868-1869, Tweede Reeks, Vol. II, pp. 92-120, 1868. Translated by W.G. Koster, "On the Speed of Mental Processes." Acta Psychologica, Vol. 30, pp. 412-431, 1969.

33. Carter, R., Krause, M., and Harbeson, M., "Beware the Reliability of Slope Scores for Individuals." Human Factors, Vol. 28, pp. 673-683, 1986.

34. Schneider, W. and Shiffrin, R., "Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention." Psychological Review, Vol. 84, pp. 1-66, 1977.

35. Shiffrin, R. and Schneider, W., "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and a General Theory." Psychological Review, Vol. 24, pp. 127-190, 1977.

36. Ashcraft, M. and Battaglia, J., "Cognitive Arithmetic: Evidence for Retrieval and Decision Processes in Mental Addition." Journal of Experimental Psychology: Human Learning and Memory, Vol. 4, pp. 527-538, 1978.

37. Mohs, R., Tinklenberg, J., Roth, W., and Kopell, B., "Methamphetamine and Diphenhydramine Effects on the Rate of Cognitive Processing." Psychopharmacology, Vol. 59, pp. 13-19, 1978.

38. Bahrick, H., Noble, M., and Fitts, P., "Extra-task Performance as a Measure of Learning a Primary Task." Journal of Experimental Psychology, Vol. 48, pp. 298-302, 1954.

39. Harris, S., North, R., and Owens, J., A System for the Assessment of Human Performance in Concurrent Verbal and Manual Control Tasks, NAMRL-1254, Naval Aerospace Medical Research Laboratory, Pensacola, FL, November 1978.

40. Laabs, G. and Stager, P., "Monitoring the Information-processing Demands of Attention Switching." Canadian Journal of Psychology, Vol. 30, pp. 47-54, 1976.

41. Midwest Research Institute, Effects of Pyridostigmine on Psychomotor and Visual Performance, Midwest Research Institute, Kansas City, MO, September 1984, pp. 1-85.

42. Bittner, A., Jr., Carter, R., Krause, M., Kennedy, R., and Harbeson, M., "Performance Evaluation Tests for Environmental Research (PETER): Moran and Computer Batteries." Aviation, Space, and Environmental Medicine, Vol. 54, pp. 923-928, 1983.

43. Moran, L. and Mefferd, R., "Repetitive Psychometric Measures." Psychological Reports, Vol. 5, pp. 269-275, 1959.

44. Seales, D., Kennedy, R., and Bittner, A., Jr., "Development of Performance Evaluation Tests for Environmental Research (PETER): Arithmetic Computation." Perceptual and Motor Skills, Vol. 51, pp. 1023-1031, 1980.

45. Griffiths, I. and Boyce, P., "Performance and Thermal Comfort." Ergonomics, Vol. 14, pp. 457-468, 1971.

46. Morgan, B., Winnie, P., and Dugan, J., "The Range and Consistency of Individual Differences in Continuous Work." Human Factors, Vol. 22, pp. 331-340, 1980.

47. Chiles, W. and Jennings, A., "Effects of Alcohol on Complex Performance." Human Factors, Vol. 12, pp. 605-612, 1970.

48. Williges, R., "Within-session Criterion Changes Compared to an Ideal Observer Criterion in a Visual Monitoring Task." Journal of Experimental Psychology, Vol. 81, pp. 61-66, 1969.

49. Craig, A., "Human Engineering: The Control of Vigilance." In J. Warm (Ed.), Sustained Attention in Human `erformance, John Wiley & Sons Ltd., New York, NY, 1984, pp. 247-291.

50. Parasuraman, R., "Vigilance, Monitoring, and Search." In K. Boff, L. Kaufman, and J. Thomas (Eds.), Handbook of Perception and Human Performance: Vol. II, John Wiley & Sons, New York, NY, 1986, pp. 43-1--43-39.

51. Williges, R., "The Role of Payoffs and Signal Ratios in Criterion Changes During a Monitoring Task." Human Factors, Vol. 13, pp. 261-267, 1971.

52. Jerison, H., Experiments on Vigilance: V. The Empirical Model For Human Vigilance, Technical Report 58-526, U.S. Air Force, Wright-Patterson Air Development Center, Dayton, OH, 1959.

53. McNicol, D., A Primer of Signal Detection Theory, George Allen & Unwin Ltd., London, England, 1972, pp. 1-242.

54. Green, D. and Swets, J., Signal Detection Theory and Psychophysics, Robert E. Krieger, Huntington, NY, 1974, pp. 1-479.

55. Colquhoun, W. and Baddeley, A., "Role of Pretest Expectancy in Vigilance Decrement." Journal of Experimental Psychology, Vol. 68, pp. 156-160, 1964.

56. Colquhoun, W. and Baddeley, A., "Influence of Signal Probability During Pretraining on Vigilance Decrement." Journal of Experimental Psychology, Vol. 73, pp. 153-155, 1967.

57. Craig, A. and Colquhoun, W., "Vigilance: A Review." In C. Drury and J. Fox (Eds.), Human Reliability in Quality Control, Taylor & Francis, London, England, 1975, pp. 71-88.

58. Craig, A., "Is the Vigilance Decrement Simply a Response Adjustment Towards Probability Matching?" Human Factors, Vol. 20, pp. 441-446, 1978.

59. Williams, P., "Processing Demands, Training, and the Vigilance Decrement." Human Factors, Vol. 28, pp. 567-579, 1986.

60. Wiener, E., "Money and the Monitor." Perceptual and Motor Skills, Vol. 29, pp. 627-634, 1969.

61. Warrick, M., Kibler, A., and Topmiller, D., "Response Time to Unexpected Stimuli." Human Factors, Vol. 7, pp. 81-86, 1965.

62. Bartram, D., Banerji, D., Rothwell, D., and Smith, P., "Task Parameters Affecting Individual Differences in Pursuit and Compensatory Tracking Performance." Ergonomics, Vol. 28, pp. 1633-1652, 1985.

63. Wickens, C., Mountford, S., and Schreiner, W., "Multiple Resources, Task-hemispheric Integrity, and Individual Differences in Timesharing." Human Factors, Vol. 23, pp. 211-229, 1981.

64. McRuer, D. and Jex, H., "A Review of Quasi-linear Pilot Models." IEEE Transactions on Human Factors in Electronics, Vol. HFE-8, pp. 231-249, 1967.

65. Repperger, D. and Levison, W., "Effects of Control Stick Parameters on Human Controller Response." In S. Hart and E. Hartzell (Eds.), 20th Annual Conference on Manual Control, National Aeronautics and Space Administration Conference Publication 2341, NASA, Ames Research Center, Moffett Field, CA, 1984, pp. 157-172.

66. Frost, G., "Man-machine Dynamics." In H. Van Cott and R. Kinkade (Eds.), Human Engineering Guide to Equipment Design, Government Printing Office, Washington, DC, 1972, pp. 227-309.

67. Ackerman, P., Schneider, W., and Wickens, C., "Deciding the Existence of a Time-sharing Ability: A Combined Methodological and Theoretical Approach." Human Factors, Vol. 26, pp. 71-82, 1984.

68. Damos, D., Smist, T., and Bittner, A., Jr., "Individual Differences in Multiple-task Performance and a Function of Response Strategy." Human Factors, Vol. 25, pp. 215-226, 1983.

69. Huynh, H. and Feldt, L., "Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-plot Designs. Journal of Educational Statistics, Vol. 1, pp. 69-82, 1976.

70. Vidulich, M. and Wickens, C., "Stimulus-Central Processing-Response Compatibility: Guidelines for the Optimal Use of Speech Technology." Behavioral Research Methods, Instruments, and Computers, Vol. 17, pp. 243-249, 1985.

71. Melton, A.W. (Ed.), Apparatus Tests, Report #4, Army Air Force Aviation Psychology Program Research Report, Washington, DC, Government Printing Office, 1947.

72. Cohen, J., Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York, NY, 1977, p. 1.

73. Steiger, J., "Testing Pattern Hypotheses on Correlation Matrices: Alternative Statistics and Some Empirical Results." Multivariate Behavioral Research, Vol. 15, pp. 335-352, 1980a.

74. Steiger, J., "Tests for Comparing Elements of a Correlation Matrix." Psychological Bulletin, Vol. 87, pp. 245-251, 1980b.

75. Bittner, A., Jr., "Statistical Tests for Differential Stability." In C. Bensel (Ed.), Proceedings of the Human Factors Society 23rd Annual Meeting, 1979, pp. 541-545.

76. Damos, D., Kennedy, R., and Bittner, A., "The Effects of Extended Practice on Dual-task Tracking." Human Factors, Vol. 23, pp. 627-632, 1981.

## GLOSSARY

**Algorithm.** A rote or mechanical procedure for solving a problem.

**Assembly language.** A low-level computer language one step above the binary machine language.

**Average correct reaction time.** The average of the reaction times when the subject responded correctly.

**Bandwidth.** The difference between the frequency limits of a band containing the useful frequency components of a signal.

**Baseline.** A measure of behavior under control conditions or before the experiment begins. Later experimental treatments are expected to modify the baseline.

**Between-subjects design.** A design in which no repeated measures are obtained on the subjects.

**Bimodal distribution.** A distribution that has two distinct modes.

**Carry-over effect.** An effect that occurs in repeated measures experiments when the administration of one treatment level affects a subject's performance on subsequent levels.

**Compile.** To prepare a machine language program automatically from a program written in a higher programming language, usually generating more than one machine instruction for each symbolic statement.

**Computerized test.** Any test that is presented using a computer, that is, the stimuli are presented on the CRT of the computer and the computer does all the response recording.

**Constant mapping.** A procedure for presenting stimuli in the Sternberg task such that memory set items are never distracters and distracters are never memory set items.

**Counterbalancing.** The process of arranging a series of experimental treatments in such a way as to minimize practice effects, fatigue, or other order effects. A simple form of counterbalancing would be to administer two experimental conditions in the order ABBA.

**Cursor.** A moving display element that represents the system output on a pursuit display. On a compensatory display the cursor represents the difference between the system input and the output.

**Debug.** To test for, locate, and remove mistakes from a program or malfunctions from a system.

**Dependent variable.** The variable that is observed and measured in an experiment. The dependent variable is the response predicted to change as a result of and in relation to the manipulation of the independent variable by the investigator.

**Discrete task.** A task requiring discrete responses, such as key presses.

**Experimental design.** A specific plan for assigning subjects to experimental conditions and the statistical analysis associated with that plan.

**Feedback.** Information provided by the various sense organs, particularly information received when a response is made.

**Fit.** To adjust a smooth curve of a specified type to a given set of points in such a way as to minimize the sum of the squares of the distances measured parallel to the axis of the ordinates from the given points to the curve.

**Gain.** The increase in signal power that is produced by an amplifier.

**Hz.** Cycles per second.

**Independent variable.** A variable under control of the investigator.

**Knowledge of results.** Augmented feedback.

**Latin square.** An experimental design in which treatments are administered in orders that are systematically varied.

**Linear regression.** A regression analysis that assumes that the predictor variable is related to the predicted variable along a straight line.

**Mixed design.** A design that contains both between- and within-subject factors.

**Normal distribution.** A bell-shaped probability curve showing the expected value of sampling a random variable. Also called Gaussian distribution, normal curve, normal probability curve.

**Normative score.** A person's score compared with the scores of other individuals, such as a percentile ranking in a particular group.

**Paced.** Each stimulus of a test follows the preceding stimulus at a certain interval.

**Paradigm.** A model, pattern, or design of the functions and interrelations of a process. In psychological research, a paradigm is an experimental design or plan of the various steps of the experiment, or a model of the process or behavior under study.

**Phase lag (lag angle).** The negative of the phase difference between a sinusoidally varying quantity and a reference quantity, which varies sinusoidally at the same frequency, when the phase difference is negative.

**Power of a statistical test.** "The probability that the test will yield statistically significant results," (72).

**Root mean square (RMS) error.** A common measure of tracking performance obtained by squaring uncorrected error values, dividing by the number of errors, and then taking the square root of the result.

**Sensitivity.** The extent to which performance on a test changes in response to changes in some variable.

**Sequence effects.** The portion of the carry-over effects that depends on the order of specific treatments.

**Slope.** The change in the ordinate of a function divided by the change in the abscissa.

**Validation.** The process of determining the accuracy of a test in measuring what it purports to measure.

**Varied mapping.** A procedure for presenting stimuli in the Sternberg task such that the memory set items and the distracters are randomly intermixed over trials.

**Task analysis.** The detailed breakdown of a job into its component skills, required knowledge, and specific operations.

**Within-subject design.** A design in which repeated measures are obtained on all of the independent (experimental) variables.

**Z-score (standardized score).** A score showing the relative status of a score in a distribution. The mean of a distribution of standardized scores is always 0.0, and the standard deviation is always 1.0.

# APPENDIX. DIFFERENTIAL STABILITY

Differential stability is concerned with performance in a repeated measures situation. Its calculation allows the investigator to determine when group performance has "stabilized" according to mathematically defined criteria. Thus, the investigator does not have to rely on "eyeballing" the data to determine when a test has been learned sufficiently to avoid confounding learning effects with experimental effects (see reference 5 for an excellent example of problems associated with estimating asymptotic performance by "eyeball").

To be differentially stable, performance on a given test must have three characteristics, which are assessed statistically with a specified level of experimental error. First, the means of the dependent measure must either be constant or increase in a slow, linear fashion. Second, the variances on each trial must be constant. Third, the rank order of subjects must be constant from trial to trial.

To determine differential stability, the investigator must first calculate the intertrial correlation matrix. This is a square matrix with 1.00s in the diagonals. All cell entries represent the correlation between performance on the ith trial with performance on the jth trial. For most tests of interest, the intertrial correlation matrix initially has what is known as superdiagonal form. This form is characterized by decreasing correlations across any given row and up any given column. After some amount of practice, the superdiagonal form disappears, and the correlations become constant (within some specified error level). When the superdiagonal form disappears, performance on the test has become differentially stable.

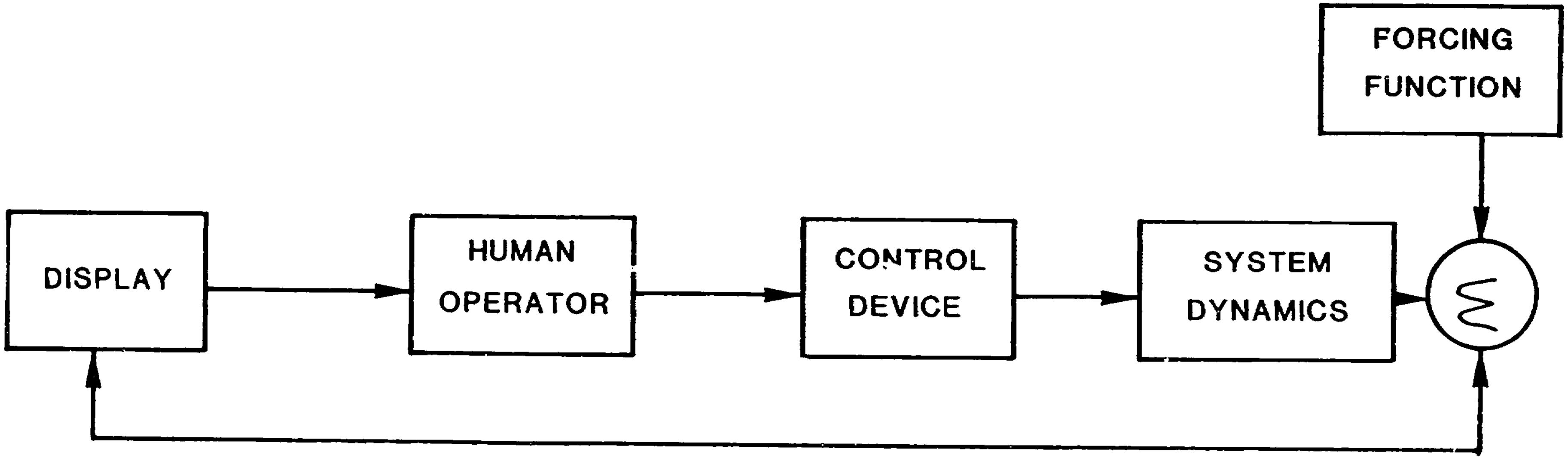
The trial on which the test becomes differentially stable can be identified by performing a number of different statistical analyses on the intertrial correlation matrix. The most common analyses are the early-versus-late analysis of variance, the Lawley Chi Squared Test, and the Steiger Test (73,74). Bittner (75) describes the first two of these in some detail and discusses the advantages and disadvantages of each. The Steiger Test has not been widely used to date because it is difficult to program. Consequently, little information is available on its characteristics.

The reader should note that some tests never achieve differential stability (see reference 76 for an example). Other tests are stable immediately; their intercorrelation matrices never show superdiagonal form. Such tests usually either assess some skill that has been practiced extensively, such as simple mental arithmetic, or a skill that is learned extremely easily, such as turning a screw a given number of rotations.

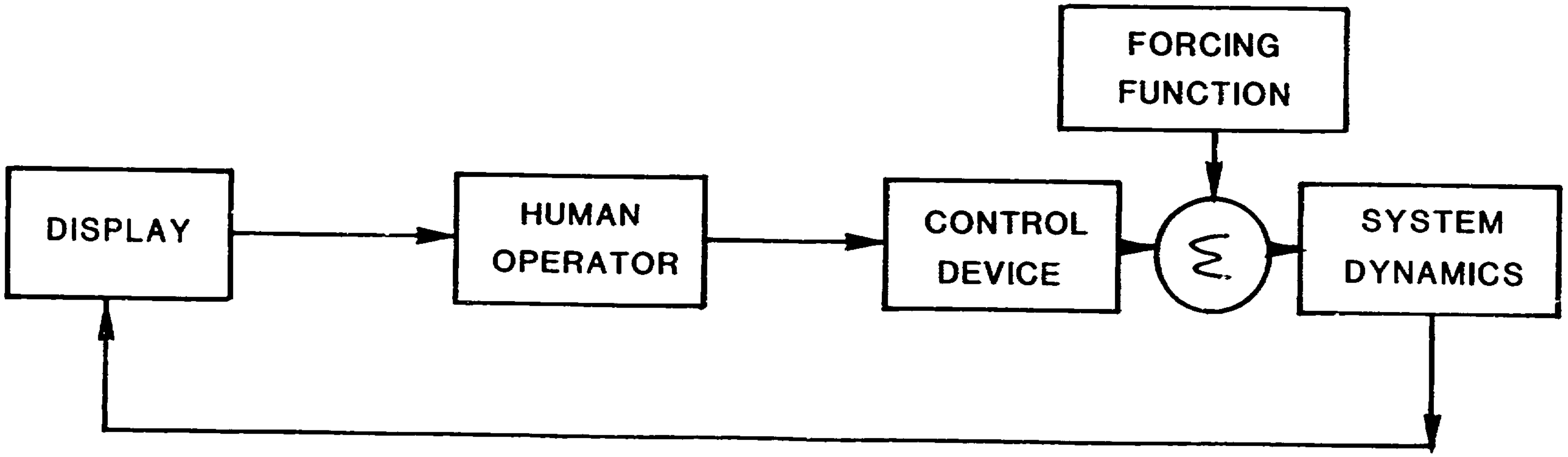
Figure 1. Schematic representations of tracking tasks.

Figure 1a shows the preferred implementation with the forcing function added to the task after the system dynamics have acted on the operator's response. Figure 1b shows a less desirable implementation with the forcing function added to the operator's c :tput before the system dynamics have acted on the operator's output.

1a.


FORCING FUNCTION
DISPLAY
HUMAN OPERATOR
CONTROL DEVICE
SYSTEM DYNAMICS
Σ

1b.


FORCING FUNCTION
DISPLAY
HUMAN OPERATOR
CONTROL DEVICE
Σ
SYSTEM DYNAMICS