REPORT

Fitness-For-Duty Testing in the Transit Workplace (TCRP F-1)

To

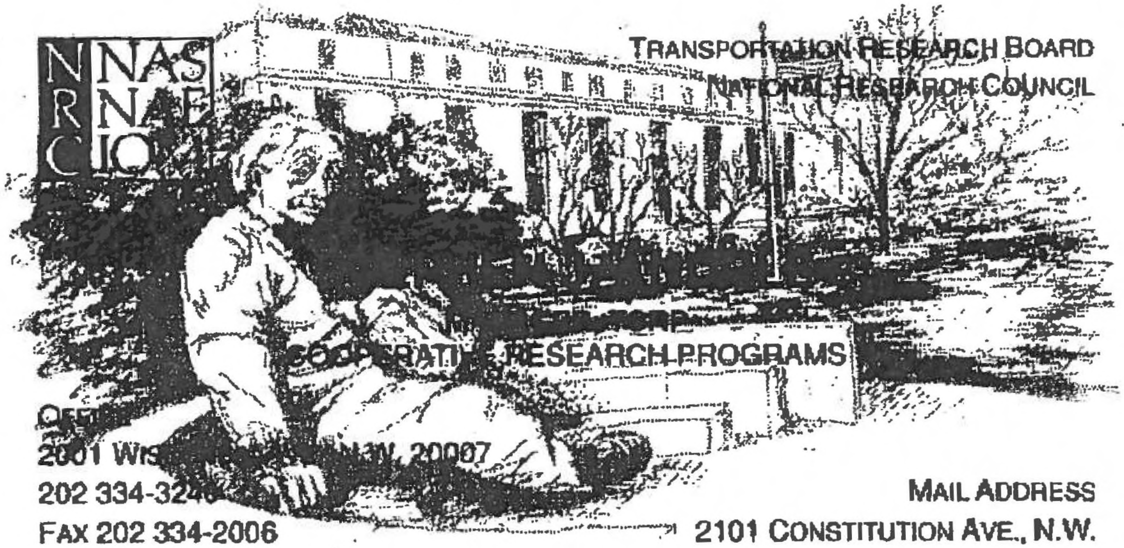
Transportation Research Board

October 31, 1996

Battelle
. . . Putting Technology To Work

3103121554 P.01/01

TOTAL P.01



NRC NAS NAE IOM

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

COOPERATIVE RESEARCH PROGRAMS

OFFI...
2001 WIS... N.W. 20007
202 334-324...
FAX 202 334-2006
INTERNET SANDRLE@NAS. EDU

MAIL ADDRESS
2101 CONSTITUTION AVE., N.W.
WASHINGTON, D.C. 20418

FITNESS-FOR-DUTY TESTING IN THE TRANSIT WORKPLACE (TCRP F-1)

FINAL REPORT

Prepared for

TRANSPORTATION RESEARCH BOARD
Transportation Cooperation Research Program
Mr. Stephen J. Andrle

Battelle
. . . Putting Technology To Work

Southern California Research Institute
Transportation Resource Associates

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES

LIST OF TABLES (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

Page

EXECUTIVE SUMMARY

To explore the feasibility of using fitness–for–duty testing for the transit industry, a study was conducted for the Transit Cooperative Research Program (TCRP) to:

- Define transit industry needs
- Determine feasibility of fitness-for-duty testing by
  - Conducting a literature search
  - Developing evaluation criteria
  - Identifying promising tests
  - Selecting independent variables for testing
  - Evaluating selected tests.

Transit Industry Needs

Selected safety–sensitive jobs were analyzed to determine the processes critical to their safe performance. The seven selected job categories were:

- Bus operator
- Rail operator
- Dispatcher
- Maintenance worker
- Bus supervisor
- Rail supervisor
- Maintenance supervisor.

Critical incident workshops were held, and employees in safety–sensitive jobs were interviewed for the purpose of developing a list of processes, abilities, and skills common to the seven job categories. The list follows:

- Memory
- Attention to detail
- Reaction time
- Reading comprehension
- Analytical ability

- Ability to understand written communication
- Ability to communicate orally.

Determination of Feasibility

A determination of feasibility of fitness-for-duty testing is difficult in part due to imprecise and varying expectations on the part of the transit industry. A consensus as to the essential characteristics of a test appears not to exist at the present time. It is evident, however, that vendors will attempt to provide tests which will meet the industry's needs. Fitness-for-duty tests characteristics that were found to be preferred by transit agencies are summarized below:

- Two minutes or less per employee per test
- Minimal training of employees and administrators
- A baseline against which to assess daily test performance
- A test requiring involuntary responses
- A test with high "face" validity
- Minimal expense.

Literature Search

A literature search was conducted to identify candidate Fitness-For-Duty tests. To the extent that reported data permitted, tests developed in research laboratories, as well as tests developed by vendors specifically for workplace applications, were evaluated for validity, reliability, and sensitivity. A report lists the tests and summarizes the available research. To further explore questions about available tests, a Vendor/Transit Symposium was convened. Nine vendors responded to the invitation to demonstrate their tests at the symposium.

Evaluation Criteria

Fitness-for-duty tests must meet the *scientific criteria* which apply to testing in general: validity, reliability, sensitivity, specificity. In addition, operational criteria apply to the tests and to the environment in which the tests will be used. *Operational criteria* related to the tests are:

- Time required to perform a single test
- Time required per employee to establish a baseline
- Time required to train staff to administer or supervise a test
- Test apparatus cost
- Ease of administering/performing the test
- Clarity of test results and interpretation

- Demonstrated correlation of test results and fitness–for–duty
- Employee acceptance of test as a valid measure of fitness–for–duty.

Operational criteria related to the transit agency environment are:

- Testing costs
- Testing frequency
- Human resource policies
- Impact on mandated substance abuse testing
- Impact on collective bargaining agreements
- Employer response to test failure
- Impact on personnel requirements/work schedules
- Test validation data
- Relevant statutes and regulations
- Legal issues
- Cost/benefit issues
- Durability and integrity of testing apparatus
- Logistics of daily or frequent testing.

These criteria were the basis for transit personnel evaluations of the tests, which were presented at the Vendor/ Transit Symposium. The evaluations were an important source of information during the process of selecting apparatus for an initial pilot experiment and subsequent full-scale experiment.

Test selection relied on data available in the literature and information provided by vendors, together with the transit industry representatives' evaluations of systems presented at the Vendor/Transit Symposium. Six systems were selected for testing:

- Enhanced Performance System (EPS–100)
- Factor 1000
- NovaScan
- Personal Safety Analyzer (PSA)
- Delta WP
- Fitness Impairment Tester (FIT).

Stressor Variable

The following criteria guided the selection of the stressor to be used in laboratory study of the selected tests:

- The stressor should not endanger the safety or health of the subjects
- The stressor should have known effects on human performance
- It should be possible in a laboratory setting to precisely administer, control and measure the stressor
- The stressor should be relevant to the transit industry.

Alcohol was selected as the stressor best meeting the criteria, and was administered to subjects in the laboratory studies.

Fitness-For-Duty Test Evaluation

A pilot study was conducted to determine sensitivity of fitness-for-duty technology to the presence of the alcohol stressor. Five vendors provided test apparatus. Individual subject baseline performance levels for each of the 25 subjects on each of the five tests were established during two four-hour sessions held the week preceding the alcohol session. During the alcohol session, subjects were were given vodka and orange juice in amounts calculated to produce 0.08% Blood Alcohol Concentration (BAC) as measured in breath speciments. They were then tested five times on each apparatus:

- Predose
- Expected Peak (EP) when an approximately 0.08% BAC is achieved
- 1 hour after EP
- 2 hours after EP
- 3 hours after EP

Because of problems with the testing apparatus, complete data were not obtained for two of the tests. Of the remaining three tests, two tests (one physiological test, one performance test) failed 60% of subjects at a mean BAC of 0.077%. These data suggest that fitness-for-duty tests may be feasible for use in the Transit environment.

To further assess fitness-for-duty testing feasibility, a full experiment was performed. Four vendors provided fitness-for-duty tests. Alcohol was again used as the stressor and each of the 24 subjects was tested five times as in the pilot study. Unlike in the pilot study, however, each subject was given three treatments:

- Treatment A: a placebo with no alcohol
- Treatment B: target peak BAC of 0.08%
- Treatment C: target peak BAC of 0.08%

The repetition of the target peak BAC in Treatments B and C offered a preliminary measure of test-retest reliability.

The most sensitive test detected impairment in 79% of subjects at 0.08% BAC, 62.5% of subjects at 0.06% BAC, 38.46% of subjects at 0.04% BAC, and 19.35% of subjects at 0.02% BAC. These results suggest that one of the tests is especially promising as a fitness-for-duty assessment tool based upon the use of alcohol as the stressor.

Conclusions

Although these results are promising, additional testing needs to be accomplished to determine whether similar sensitivity to other stressors can be achieved. In additional, important logistical issues must be addressed to ensure that the appropriate policies are in place to support the use of fitness-for-duty testing in the workplace, including procedures for dealing with individuals who fail the tests. Finally, resolution must be achieved on the concept of fitness-for-duty and how it is to be defined operationally so as to provide a better "fit" between pass/fail criteria embodied in the tests and the Transit industry's own requirements.

CHAPTER ONE

INTRODUCTION AND RESEARCH APPROACH

Problem Statement

This report describes a study conducted for the Transit Cooperative Research Program (TCRP) which explored the feasibility of fitness-for-duty testing in the transit workplace. The study was performed under TCRP Project F-1, *Fitness-For-Duty Testing in the Transit Workplace*.

The TCRP was authorized as part of the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA) and represents three cooperating organizations: the Federal Transit Administration (FTA) of the DOT; the National Academy of Sciences, acting through the Transportation Research Board (TRB); and the Transit Development Corporation, Inc. (TDC), a nonprofit educational and research organization established by the American Public Transit Association (APTA).

In 1993, the TCRP chose the team of Battelle (prime contractor), the Southern California Research Institute (SCRI), and Transportation Resource Associates (TRA) to perform this study. Work was directed by the TCRP Project Manager with oversight by a TCRP Project Panel.

Fitness-For-Duty

Safety has always been a primary concern in transit operations throughout the United States, and how well individual transit employees conduct their duties plays an integral part in the overall safety of the transit system. Employees whose performance is impaired by alcohol, drugs, fatigue, or emotional distress compromise the safety of the American public. Transit agencies have historically relied on observation and evaluation by direct supervisors to ensure that employees are mentally and physically prepared to perform their jobs safely each day. Some transit agencies have instituted biochemical drug and/or alcohol testing to further uncover and deter alcohol and drug use.

In recent years, direct observation to detect impairment has become more formalized with the training of "drug recognition experts" in law enforcement. Highly trained officers use a standardized evaluation protocol that integrates multiple signs and symptoms to form an opinion about an individual's impairment. Central to the evaluation are vital signs and eye signs, which are not observable to an untrained observer in an unstructured situation. The issue is particularly difficult since stimulant influence (such as by cocaine) is difficult to detect. The lengthy drug recognition procedures, as developed for law enforcement, are probably not suitable for screening large numbers of employees on a routine basis and are not known to be occurring in transit today.

The use of biochemical tests to detect alcohol and drug consumption has become more common; and, by U.S. Department of Transportation (DOT) regulation, mandatory testing for all safety-sensitive transit employees was phased in, beginning January 1, 1995. Safety-sensitive employees include those whose performance directly affects the safety of the public or other employees. They include bus drivers, train operators, dispatchers, and certain maintenance and security personnel, among others.

While biochemical tests for alcohol and drug consumption are generally considered accurate and successful in deterring alcohol and drug use, many transit agencies are concerned about the safety implications of other stressors that are not now being formally evaluated (such as fatigue and emotional stress) and the impracticality of a lengthy direct observation of each employee, or the administration of biochemical alcohol and drug tests, on a daily basis, to detect drug or alcohol consumption and impairment.

Fitness-for-duty tests have been proposed as a possible response to these challenges and to complement traditional testing methods. Many vendors have a variety of fitness-for-duty tests in development and several are in use in transportation and non-transportation workplaces. Transit agencies have expressed a growing interest in using fitness-for-duty tests.

Fitness-For-Duty Testing

It is difficult to give a single definition for fitness-for-duty testing because it is an emerging field with generally complementary, but sometimes conflicting, expectations on the parts of employers, employees, test vendors, and scientists. A generic working definition encompasses a broad array of tests proposed by various vendors to assist employers in determining whether an employee is physically and mentally fit to begin work each day. Generally speaking, fitness-for-duty tests are intended for frequent (probably daily) administration to all safety-sensitive employees with immediate results. They are usually envisioned as being computer-based, sensitive to multiple stressors, and of short duration. Fitness-for-duty tests generally are not suggested for post-accident testing (because of the stress induced by the accident itself) and presently do not satisfy regulatory requirements for alcohol and drug testing of safety-sensitive employees in transportation industries. Finally, in contrast to most of the traditional testing methods, fitness-for-duty testing has not yet been subject to legal challenges—and validation—as to its application and use of results. In some fields, particularly the nuclear industry and aviation, fitness-for-duty tests are also known as "readiness-to-perform" tests.

The relationship of broad categories of fitness-for-duty tests within the larger scope of employee "impairment" tests is shown in Figure 1-1. Traditional tests include *biochemical* testing (urine, blood, and breath testing for alcohol and drug consumption), *psychophysiological* testing (finger to nose and walking the line for alcohol), and *sensory evaluation* (personal appearance, odor of alcohol). Fitness-for-duty tests fall into two broad categories. *Performance-based* fitness-for-duty tests directly measure an employee's impairment on a simple task that requires some of the same cognitive processes and skills as the employee's regular job. The presumption is that if an employee cannot perform the simple analogous task properly, he or she will not be able to perform the actual job properly and safely. *Physiological* fitness-for-duty tests examine an employee's involuntary reactions to the presence of a stressor such as alcohol or fatigue. These include pupil response to light, changes in voice patterns, and changes in brain activity. The presumption is that abnormal responses reflect impairment.

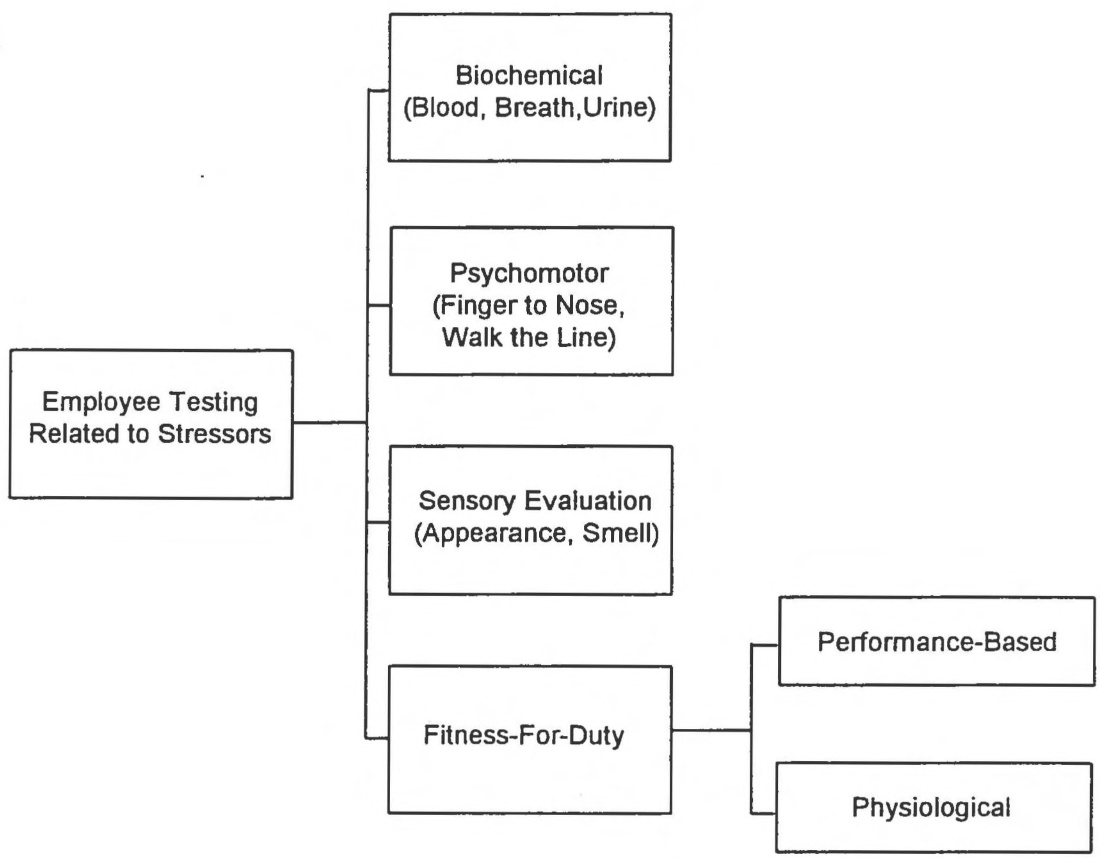


Figure 1-1. Relationship of Fitness-For-Duty Tests to Other Types of "Impairment" Tests

Potential Advantages of Fitness-For-Duty Tests

The most commonly claimed advantages of fitness-for-duty testing over traditional testing are the frequency with which the tests can be given and the low cost per test. Proponents have also suggested that fitness-for-duty tests can detect impairment caused by a wider range of stressors than can traditional testing methods. These stressors could include alcohol and drugs, which are generally the focus of traditional tests, but also stressors that are more difficult to quantify, such as fatigue and emotional stress. Proponents also claim that fitness-for-duty tests can detect impairment caused by a combination of stressors, each of which might not exceed some regulatory threshold. Fitness-for-duty tests are also advanced as being non-presumptive, in that most fitness-for-duty tests do not attempt to identify the actual stressor but rather assess only whether an employee can do the job, thereby avoiding legal implications and moral judgments. Finally, proponents suggest better acceptance by transit officials and employees because the tests often resemble the actual job that is to be performed and do not require the employee to provide a sample of blood, urine, or breath.

Expectations For Fitness-For-Duty Testing

The scientific and transit communities have already begun to formulate a set of expectations for fitness-for-duty tests. These expectations fall into two broad categories, scientific and operational. Scientific expectations center around the principal relationship that must be established between the tests and the job—that employee performance on a fitness-for-duty test accurately predicts on-the-job performance. This means that:

- The tests *measure* variations in human functioning (sensitivity) and provide consistent scores for the same level of functioning across different days (reliability)
- The tests are *sensitive* to the variety of stressors (alcohol, fatigue, etc.) likely to be found in the transit environment
- The tests possess "criteria validity," that is, they must accurately measure one or more aspects of human functioning that is *relevant to the job being performed.*

Operational expectations center around the integration of fitness-for-duty testing within a transit agency's daily functioning. These needs can be defined in terms of attributes that must be satisfied, including the cost to purchase and maintain the tests, the time required to train employees to take the tests, and additional administrative requirements. Parallel to the "criteria validity" expected by the scientific community, there is a strong expectation in the transit community that fitness-for-duty tests will possess "face validity," that is, they will resemble jobs typically performed in transit. For example, a fitness-for-duty test for bus drivers would mimic driving tasks, such as hand-eye coordination and response to colored signals.

Objective

This study was conducted in order to determine the feasibility of fitness-for-duty testing in the transit environment. Feasibility, as used in this study, refers to the satisfaction of both scientific and operational expectations for fitness-for-duty tests. Given the early stage of development of fitness-for-duty tests, and the inevitably conflicting expectations, feasibility cannot be defined too rigorously at this time. Instead, the goal is to determine whether there is sufficient evidence to suggest that fitness-for-duty testing may, with further development, succeed in meeting those expectations key to its successful implementation in the transit environment.

In order to assess feasibility, this study determined the testing needs of transit agencies, the state of development of fitness-for-duty tests, and the scientific credibility of those tests. By comparing the needs of transit agencies to fitness-for-duty tests offered by vendors, the study made a broad assessment of the feasibility of using fitness-for-duty testing in the transit industry. Although proprietary tests were examined, it was not the purpose of this study to pick "winners and losers" among the participating vendors. Instead, the research team focused on identifying the advantages and disadvantages that characterize fitness-for-duty tests *as a group*. Assessments within this study were conducted for these specific purposes and should not be used to compare specific vendors or their products.

Research Plan

Because two constituencies, the scientific community and the transit community, must be satisfied for fitness-for-duty testing to be considered feasible, this study had two complementary, and partially overlapping, research perspectives.

Scientific Perspective

The scientific perspective was addressed within the framework of information processing theory, which proposes that any complex task is performed with a set of cognitive processes, such as perception, attention, memory, decision making, and response selection and execution. An impairment of any one of these processes would be expected to reduce the ability of an employee to perform the task. Since stressors such as alcohol and drugs are known to influence cognitive processes in characteristic ways that affect an individual's performance, any test that could measure changes in an employee's cognitive processes could be presumed to be able to predict changes in the employee's ability to perform a task. This relationship is shown in Figure 1-2.

To achieve face validity and criteria validity, the cognitive processes that are measured by the fitness-for-duty tests would need to be a subset of those required to perform the task; and they

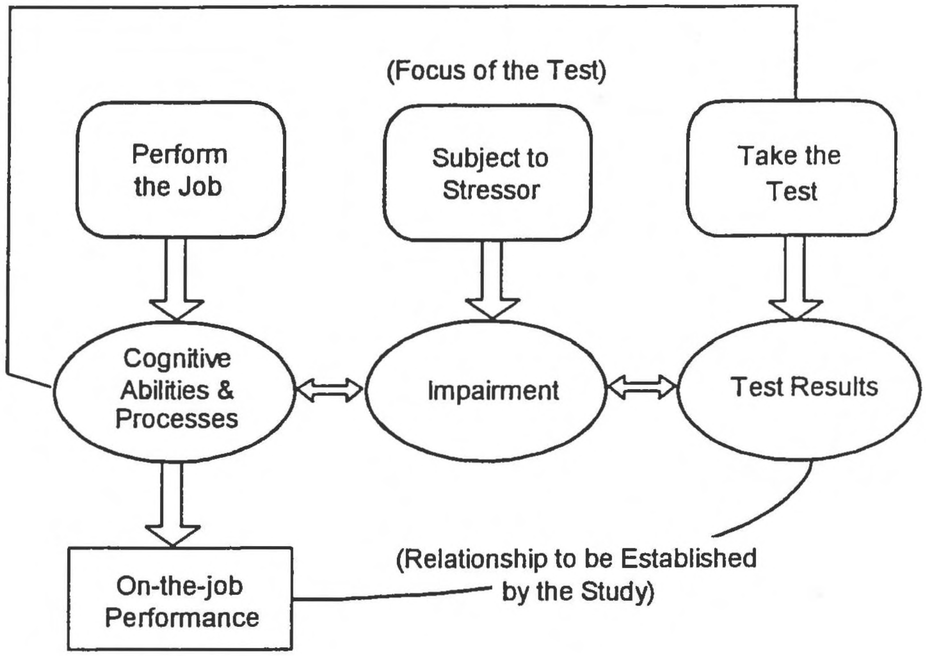


Figure 1-2. Relationship of Stressors to Impairment in the Information Processing Framework

must be sensitive to stressors likely to be found in the transit environment. Fitness-for-duty tests satisfying these requirements could be considered to be measuring *performance*. At the beginning of this study, it was expected that all fitness-for-duty tests could be evaluated within this information processing framework.

During the study, however, the research team identified several proposed fitness-for-duty tests that, instead of evaluating performance, assess an employee's *physiological* condition. In contrast to performance-based fitness-for-duty tests, which attempt to measure the effects of a particular stressor, and biochemical tests, which measure the presence of a stressor (alcohol or drugs), physiological fitness-for-duty tests attempt to detect symptoms of an employee subjected to a stressor and then equate those symptoms to performance. These symptoms could include changes in pupil size, eye tracking, speech patterns, and neural activity. Although physiological testing is similar to traditional biochemical testing in that performance is not directly measured, physiological testing can, for purposes of this study, be considered a form of fitness-for-duty testing because the tests are intended to be administered on a frequent basis before an employee goes on duty and because they do not require collection of breath, blood, or urine.

Physiological tests cannot be evaluated within an information processing framework, however, because they do not directly evaluate an employee's cognitive processing, nor do the tests make any effort to mimic an employee's tasks. Instead, these tests assess fundamental aspects of physiological functioning that are independent of the tasks an employee must perform.

Operational Perspective

Operational questions raised by the transit community (such as time to test, cost to test, etc.) were addressed by comparing the logistical and policy needs of the transit industry to current or projected attributes of fitness-for-duty tests. Of course, many of the questions of interest to the scientific community are of interest to the transit community as well. Many of the tasks undertaken during this study address both perspectives.

Research Tasks

This study was conducted in a series of seven tasks, as shown in Figure 1-3.

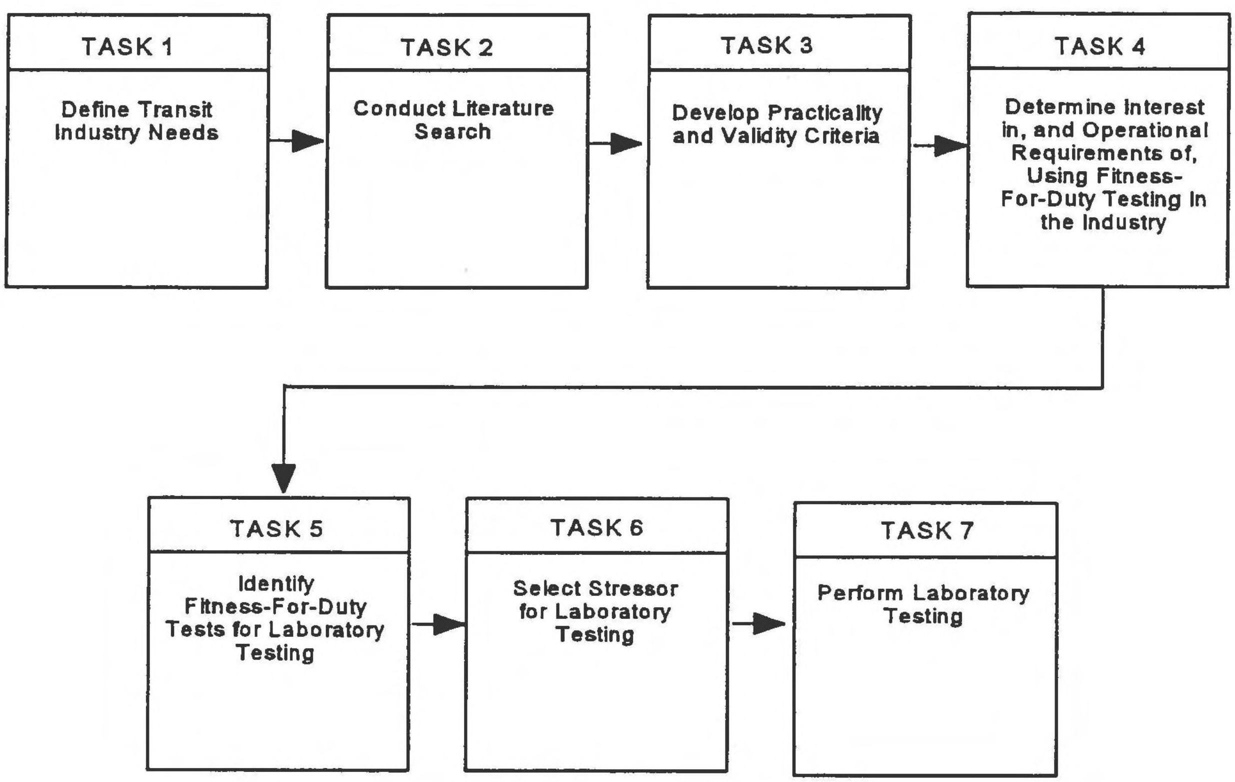


Figure 1-3. Tasks to Complete Fitness-For-Duty Study

Task 1: Define the Processes and Abilities Required to Perform Safety-Sensitive Jobs

Criteria validity and face validity suggest that performance-based fitness-for-duty tests can be assumed to reflect an employee's fitness to perform the job to the extent that the test measures the performance of cognitive processes and abilities that are also required by the job. The first task in this study had the goal of identifying the cognitive processes and abilities required by transit safety-sensitive jobs.

Identifying the particular cognitive processes and abilities required a three-step process: (1) identify which jobs are safety-sensitive by examining job descriptions and talking to transit officials; (2) identify attributes of jobs that affect safety by examining job descriptions and other literature, interviewing transit officials, and talking to transit employees at job analysis workshops; and (3) describe the processes and abilities that fitness-for-duty testing must assess. Chapter Two of this report describes the specific steps that were performed and the outcome of this work.

Task 2: Conduct Literature Search

The second task involved identifying performance-based tests that could be candidates for use as fitness-for-duty tests and are worthy of closer examination in a laboratory study. An extensive literature search was conducted to provide a catalogue of information about existing performance-based tests. Information collected included the conditions under which each test has been evaluated and any conclusions that were drawn concerning the validity, reliability, and sensitivity of the test. Candidate tests were described and evaluated using several criteria, including test administration time, reliability, and sensitivity. In addition, candidate vendor products that might also be evaluated in a laboratory study were identified. In most cases, these products were based on one or more of the individual tests identified in the literature review. [Additional vendors and tests were identified during Task 4.] The results of this task are described in Chapter Three.

Task 3: Develop Practicality and Validity Criteria

Using the results of Task 1 interviews and the Task 2 literature review, a preliminary list of criteria for assessing fitness-for-duty tests was developed. These criteria were intended to be used, in particular, for assessing which tests should be included in the laboratory experiment. To ensure that candidate tests were feasible for use in the transit environment, criteria were developed that reflect both scientific and operational aspects of fitness-for-duty testing. The criteria identified are described in Chapter Four.

Task 4: Determine Interest in, and Operational Requirements of, Using Fitness-For-Duty Testing in the Transit Industry

Inherent in the concept of fitness-for-duty test feasibility used in this study are the specific operational requirements of the transit agencies. A test must not only meet the scientific requirements that define feasibility but must also be feasible for actual use in the transit environment. To assess the operational requirements of the transit agencies, two activities were performed. First, transit agency managers were interviewed to access information concerning the conditions of test administration, including the number of employees to be tested, time constraints, space availability, and labor agreements.

A second activity involved holding, in October, 1993, a Vendor/Transit Symposium at the Battelle facilities in Columbus. Fitness-for-duty test vendors were invited to attend the two–day symposium, to demonstrate their tests, and to answer questions from representatives of transit agencies. To ensure that all fitness-for-duty vendors had the opportunity to attend, all known vendors of fitness-for-duty tests were contacted. In addition, a notice was placed in the *Commerce Business Daily* which described the purpose of the Symposium and invited all interested vendors to participate. The only requirement was that each vendor had to bring a working unit that could be demonstrated to symposium participants.

Selected transit agencies were asked to send representatives who could describe their requirements and preferences for fitness-for-duty tests and answer questions from potential vendors. The project research team, representatives from the TRB/TCRP, and several DOT modal agencies, including the FTA, the Federal Aviation Administration, and the Maritime Administration also attended.

The symposium provided agency representatives and the research team hands–on experience with fitness-for-duty tests, while vendors had the opportunity to gain additional insights into transit industry needs. As a final activity of the symposium, transit agency representatives and vendors were interviewed in separate groups to obtain their insights on using fitness-for-duty testing in the transit environment. Transit agency representatives also completed questionnaires that assessed their opinions on the tests which vendors had presented. The information gathered during the Symposium is summarized in Chapter Four.

Task 5: Identify Fitness-For-Duty Tests for a Pilot Experiment

Based upon the scientific and operational criteria identified for fitness-for-duty tests during Tasks 1, 3, and 4, tests that best met these criteria were identified. These tests became candidates for participation in the pilot laboratory study conducted in Task 7. The tests that were selected and the basis for their selection are described in Chapter Four.

Task 6: Select Stressor for Pilot Experiment

Task 6 involved choosing a stressor to serve as the impairing agent in the laboratory test conducted in Task 7. Factors that entered into the selection of the stressor are described in Chapter Four.

Task 7: Conduct Pilot Experiment

Of the vendors whose tests were identified as being feasible in the near term, several were then requested to provide equipment for clinical testing. The number of vendors who were invited to participate in the testing was determined by the resources available for testing, not by the number of vendors who met any minimum criteria.

Laboratory testing was conducted from May through August 1994 at the Southern California Research Institute (SCRI). Human subjects were trained to perform the various tests while unimpaired. After consuming a measured amount of alcohol, the subjects performed the same tests at various blood alcohol concentrations (BAC). The ability of each test to detect the resulting impairment was measured. Subjects received an alcohol dosage that produced a mean peak BAC of 0.08 percent, as measured in breath specimens. Five test times were used: before receiving alcohol, at the estimated time of peak BAC, and at one-hour intervals for three additional tests on the descending alcohol curve. The results of the pilot study are described in Chapter Five.

A full-scale laboratory experiment was then performed. Each subject participated in three treatment sessions, two alcohol (expected peak BAC of 0.08%) and a placebo (peak BAC of 0.00%). As in the pilot experiment, subjects performed the fitness-for-duty tests just prior to receiving alcohol, at expected peak BAC, and at one-hour intervals for three additional tests on the descending alcohol curve. The results of the full-scale experiment are described in Chapter Six.

On the basis of the results obtained from these seven tasks, preliminary conclusions about the feasibility of fitness-for-duty testing in the transit workplace were reached. These conclusions are described in Chapter Seven and are followed, in Chapter Eight, by recommendations on steps this investigation might take in future efforts.

CHAPTER FIVE

PILOT EXPERIMENT EXPERIMENTAL DESIGN

The pilot experiment design and objectives were limited to the issue of feasibility of the technology as a whole, not the evaluation of specific fitness-for-duty tests. Subjects established their baseline performance levels during two four-hour sessions on the week preceding the alcohol session. During the alcohol session, subjects were tested five times, as shown in Table 5-1.

Table 5-1. Testing Schedule

Battery	Name	Time
1	Predose	Before drinking and at 0.00% BAC
2	Expected Peak (EP)	30 minutes after end of last drink and at approximately 0.08% BAC
3	EP + 1 Hour	1 hour after EP
4	EP + 2 Hours	2 hours after EP
5	EP + 3 Hours	3 hours after EP

Each battery was comprised of five FFD tests. Each test yielded a Pass/Fail result. The pilot experiment was conducted at the SCRI facility in Los Angeles.

Subject Characteristics

As seen in Table 5-2, more than 100 individuals telephoned SCRI in response to recruitment ads and word–of–mouth information about the experiment. Telephone interviews led to in–person interviews for 63 applicants.

One male, Subject No. 12, was unable to complete the alcohol session, and data were obtained from replacement Subject No. 30. Because the reasons for atypical BAC curves for two women were not understood, an additional female (No. 25) was enrolled. Data are reported for 12 men and 13 women. Their ages, body weights, and ethnicity are summarized in Table 5-3 and Figures 5-1 and 5-2.

Table 5-2. Applicant Interviews

APPLICANT INTERVIEWS			
Eligible telephone interviewees			63
Not interviewed in person			
	Failed to schedule in-person interview	-3	
	Scheduled but canceled in-person interview	-2	
	Failed to appear for in-person interview	-2	-7
Interviewed in person			56
Ineligible:			
	Over age 40 years	-2	
	Anemic, grossly obese	-1	
	MMPI profile not acceptable	-7	
	Q-F-V drinking category excludes	-19	-29
Eligible by in-person interview			27
Failed to appear for baseline session			-1
ENROLLED IN EXPERIMENT			26

Table 5-3. Subjects' Characteristics (Pilot)

Male Subjects	Age (Years)	Weight (Lbs.)	Height (Inches)	Ethnic Group
1	28	174	71.0	Caucasian
2	25	184	69.5	Asian
3	24	268	76.8	African-Am
4	21	181	71.8	Caucasian
5	21	156	67.5	Caucasian
6	21	172	70.2	Caucasian
7	21	180	68.5	Caucasian
8	21	193	70.0	Caucasian
9	22	188	71.2	Caucasian
10	25	154	69.5	Caucasian
11	23	161	66.2	Asian
12	21	192	70.0	Asian
30	21	151	68.5	Caucasian
Mean	22.6	181.1	70.1	
Std. Dev.	2.3	29.8	2.5	
Female Subjects	**Age (Years)**	**Weight (Lbs.)**	**Height (Inches)**	**Ethnic Group**
13	23	128	67.0	Caucasian
14	32	136	69.0	Caucasian
15	25	136	62.0	Latina
16	22	135	69.0	Caucasian
17	21	127	64.0	Asian
18	21	107	62.5	Asian
19	22	105	62.0	Caucasian
20	24	141	62.0	Asian
21	38	106	58.0	Caucasian
22	21	110	62.0	Caucasian
23	23	190	65.0	Caucasian
24	21	123	63.0	Caucasian
25	23	101	64.0	Caucasian
Mean	24.3	126.5	63.8	
Std. Dev.	5.1	23.6	3.1	
All Subjects	**Age (Years)**	**Weight (Lbs.)**	**Height (Inches)**	
Mean	23.5	155.9	67.0	

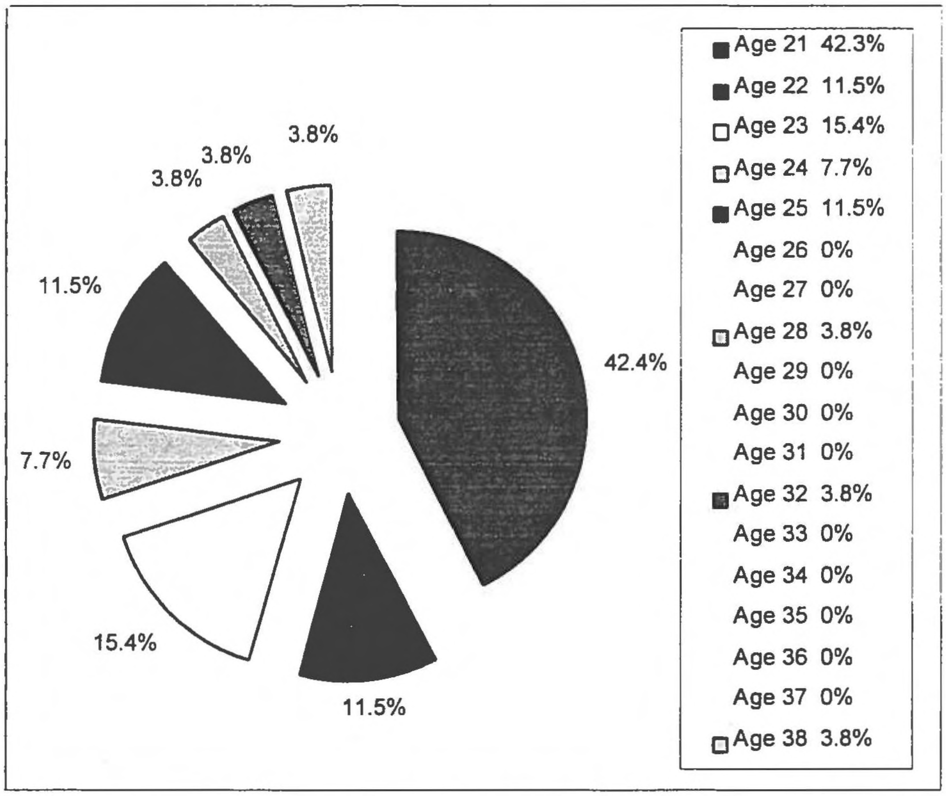


Figure 5-1. Subjects' Ages

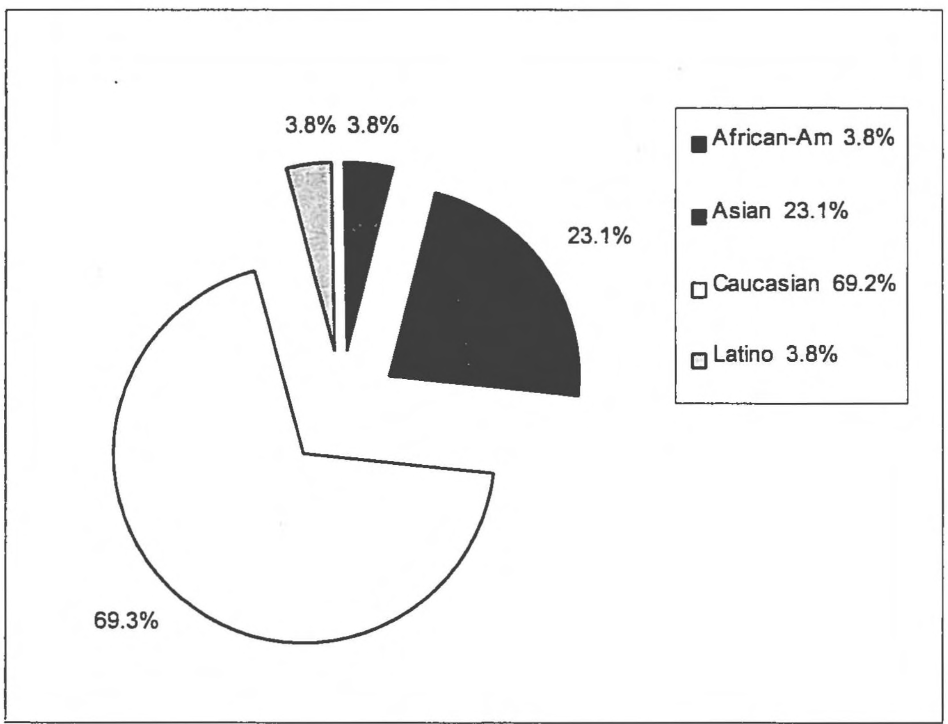


Figure 5-2. Subjects' Ethnic Group

The protocol for the experiment limited subject ages to 21 to 40 years but did not otherwise specify an age distribution. Qualified subjects in that range were enrolled as they applied without further regard to age. The result is a clear bias toward young subjects, with a mean age for the sample of 23.5 years. Over–representation of ages 21 to 25 years is attributable largely to the timing of the experiment, which was in progress during the spring and summer months when classes had just ended for college students. Many applicants were seeking short–term job opportunities, which would coincide with the summer school break.

Application of a quantity–frequency–variability (Q–F–V) scale (Cahalan et al., 1969) to the subject pool resulted in 18 subjects being categorized as moderate drinkers, five as low heavy drinkers (men), and two as light drinkers (women). The scale was administered initially during the telephone interview. When it was repeated during the in–person interview, two men were switched to a heavier drinking category and two women to a lighter drinking category. Possibly the changed responses reflect subjects' efforts to be more accurate than on first hearing the questions. It is also possible that they were trying to second guess the alcohol–use criteria to ensure acceptance into the study.

SCRI experience with self–reports of alcohol consumption has proven them sufficiently accurate to avoid overdosing. Self–reports are less likely to be an accurate index of tolerance or sensitivity to alcohol. In these data, it is possible that differences in tolerance or sensitivity obscure a significant relationship of BAC and test failures. Unfortunately, those variables can be only roughly approximated by the self–reports of alcohol exposure.

Alcohol Data

The alcohol doses (ounces of alcohol per pound of body weight) were expected to produce a mean peak BAC of 0.08% as measured in breath. The alcohol was given to fasted subjects, who consumed it over a 30–minute period. The time–to–peak was estimated as approximately 30 minutes after the end of drinking.

The mean peak BAC for 25 subjects, as measured with an Intoxilyzer 4000, was 0.077%, with a range of 0.07% to 0.09% (Table 5–4, Figure 5-3). There is, of course, no direct evidence that the actual alcohol peak coincided with the first breath test, nor can the exact time–to–peak be determined. Despite the admonition to subjects to fast, it is possible that some of them had consumed enough food to prevent full absorption within 30 minutes. Although a subject's BAC may have continued to rise after the first breath test, it would have peaked prior to the second breath test.

The mean rate of metabolism was 0.016% BAC per hour and, in general, the measured BACs followed the expected absorption–metabolism pattern. Other data characteristics merit comment.

Table 5-4. Subjects' BACs

Male Subjects	Predose	Expected Peak (EP)	EP+1 Hour	EP+2 Hours	EP+3 Hours
1	0.000	0.071	0.059	0.043	0.030
2	0.000	0.073	0.066	0.050	0.040
3	0.000	0.076	0.060	0.043	0.028
4	0.000	0.090	0.074	0.054	0.048
5	0.000	0.081	0.065	0.050	0.029
6	0.000	0.093	0.065	0.049	0.035
7	0.000	0.094	0.070	0.052	0.041
8	0.000	0.084	0.077	0.059	0.045
9	0.000	0.079	0.059	0.049	0.033
10	0.000	0.081	0.061	0.048	0.035
11	0.000	0.088	0.064	0.054	0.039
30	0.000	0.073	0.065	0.050	0.032
Mean	0.000	0.082	0.065	0.050	0.036
Std. Dev.	0.000	0.008	0.006	0.004	0.006
Female Subjects	**Predose**	**Expected Peak (EP)**	**EP+1 Hour**	**EP+2 Hours**	**EP+3 Hours**
13	0.000	0.076	0.055	0.041	0.022
14	0.000	0.082	0.054	0.046	0.026
15	0.000	0.076	0.057	0.047	0.028
16	0.000	0.081	0.051	0.038	0.017
17	0.000	0.072	0.054	0.045	0.024
18	0.000	0.077	0.059	0.043	0.020
19	0.000	0.060	0.062	0.057	0.044
20	0.000	0.073	0.060	0.048	0.030
21	0.000	0.076	0.060	0.048	0.020
22	0.000	0.073	0.067	0.045	0.025
23	0.000	0.059	0.066	0.048	0.032
24	0.000	0.067	0.065	0.034	0.017
25	0.000	0.075	0.058	0.046	0.006
Mean	0.000	0.073	0.059	0.045	0.024
Std. Dev.	0.000	0.007	0.005	0.006	0.009
All Subjects	**Predose**	**Expected Peak (EP)**	**EP+1 Hour**	**EP+2 Hours**	**EP+3 Hours**
Mean	0.000	0.077	0.062	0.048	0.031
Std. Dev.	0.000	0.009	0.006	0.006	0.009

Alcohol doses were adjusted to accommodate gender differences and were expected to produce 0.08% BACs with both male and female subjects. The mean BACs, however, were 0.073% for females and 0.082% for males. The metabolism rates were 0.016% per hour for females and 0.015% per hour for males. Although the differences are statistically non–significant, the consistency of females' lower peak BACs is striking. As can be seen in the table, peak measured BACs were below 0.075% for almost half the females.

Test Performance

Subjects took the five–test battery a total of five times during the alcohol session, which was expected to yield a total of 20 passes/fails per subject. As will be discussed, however, the data are valid for only three of the systems, yielding 12 opportunities for each subject to pass or fail. Total failures by subjects ranged from none (3 subjects) to 10 (1 subject). As expected, the largest number of failures occurred at peak BACs, and the numbers declined thereafter. Women failed more tests and showed more between–subject variability.

Analysis of the data to examine the feasibility of fitness-for-duty testing has been restricted largely to analysis of the pass/fail output of the systems. Although raw scores from pre–dosing and alcohol testing are of interest, those scores typically will not be viewed directly or immediately in the workplace to determine whether an employee is fit for duty. Thus, the feasibility question at its simplest level must be addressed in terms of a system's pass and fail results.

It is not possible in all cases to determine whether a subject was or was not impaired and whether a particular test is reliably sensitive to impairment. Note, for example, that a pass by a subject for whom a positive BAC has been measured may be the outcome of at least the following three circumstances:

1) The subject was not impaired at that BAC on the skill or skills measured by that test.
2) The subject was impaired, but the test is not a sensitive measure of alcohol impairment at that BAC.
3) The subject was impaired, and the test is sensitive (i.e., it measures changes in critical skills), but the apparatus utilizes an inappropriate scoring criterion, or cutoff, and the "pass" actually was a false negative.

Since the overall study objective was limited to an examination of the feasibility of fitness-for-duty testing in the transit industry, it is viewed as inappropriate for this report of results to rate or directly compare the systems. For that reason, they will be referred to only as Tests 1 through 5.

Objective Measures

Test 1

As shown in Table 5-5 and Figure 5-4, this system failed 0% of the subjects at the Predose battery (0.00% BAC), 60% of the subjects at the Expected Peak (EP) battery (mean BAC 0.077%), 28% of the subjects at the EP+1 hour battery (mean BAC 0.062%), 8% of subjects at EP+2 hours battery (mean BAC 0.047%), and 0% of subjects at the EP+3 hours battery (mean BAC 0.023%). Test 1's passes at the last two batteries can be viewed either as "false negatives" or "correct rejections," depending on the safety demands, BAC criterion, and objectives of a particular work environment. Also of note is a low incidence of "False Positives" (i.e., Fails at 0.00% BAC).

Test 2

The data for this system appear in Table 5-6 and Figure 5-5. The test protocol for this device allows four trials per test, and satisfactory performance of those trials is a pass, which concludes the test. If the first four trials are failed, however, the individual tries again with four more trials, and satisfactory performance of the second set of trials also produces a pass; i.e., a failure requires that two sets of trials be failed. By that scoring system, Test 2's hit rate is 40% for both the first and second test times, 24% at the third test time, and 8% at the last test time.

In alternative scoring, if failure of the first set of trials is scored as a fail (i.e., second tries not considered), the hit rate at peak alcohol levels increases to 60% (Table 5-7, Figure 5-6). Another consequence of this more stringent scoring, of course, is an increased number of test failures at low BACs, and five additional failure would be scored at BACs below 0.05%. Whether those failures are defined as hits or as false positives will depend on workplace objectives.

Test 3

Although sensitivity to alcohol was expected for Test 3, given the demands of the task, relatively few subjects failed at the alcohol concentrations of this experiment (Table 5-8, Figure 5-7). Since frequent screen warnings during testing to subjects about borderline performance had been observed, further analysis was undertaken.

Test 3 allows a second attempt when a failure occurs. The re–analysis did not allow the second attempt. In addition, the failure criterion was changed from 2 to 1.5 standard deviations based on inspection of graphs generated by the system. The rationale for the more stringent criterion was the observation of numerous pre–dosing–to–peak BAC declines in performance which *just* failed to reach the two standard deviations criterion. Table 5-9 and Figure 5-8 display the results of changing the criteria.

Scoring criteria must reflect many concerns that fall outside the scope of the pilot study. Also, the re–analysis is incomplete as a result of missing data. Nonetheless, it suggests that Test 3 is sensitive to alcohol, as expected, and that it captured the impairment, but that the pass/fail criteria were too lenient.

Table 5-5. Pass/Fail Results for FFD Test 1 (+=Pass, X=Fail)

Male Subjects	Predose (Mean BAC 0.00%)	EP (Mean BAC 0.077%)	EP+1 Hour (Mean BAC 0.062%)	EP+2 Hours (Mean BAC 0.048%)	EP+3 Hours (Mean BAC 0.031%)
1	+	X	+	X	+
2	+	+	+	+	+
3	+	X	X	+	+
4	+	X	X	+	+
5	+	X	+	+	+
6	+	X	+	+	+
7	+	+	+	+	+
8	+	X	+	+	+
9	+	X	+	+	+
10	+	+	+	+	+
11	+	X	+	+	+
30	+	+	+	+	+
Total Passes	12	4	10	11	12
Total Fails	0	8	2	1	0
Female Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
13	+	X	X	+	+
14	+	+	+	+	+
15	+	X	X	+	+
16	+	+	+	+	+
17	+	X	+	+	+
18	+	+	+	+	+
19	+	+	+	+	+
20	+	X	X	+	+
21	+	X	X	+	+
22	+	X	+	+	+
23	+	X	X	X	+
24	+	+	+	+	+
25	+	+	+	+	+
Total Passes	13	6	8	12	13
Total Fails	0	7	5	1	0
All Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
Total Passes	25	10	18	23	25
Total Fails	0	15	7	2	0

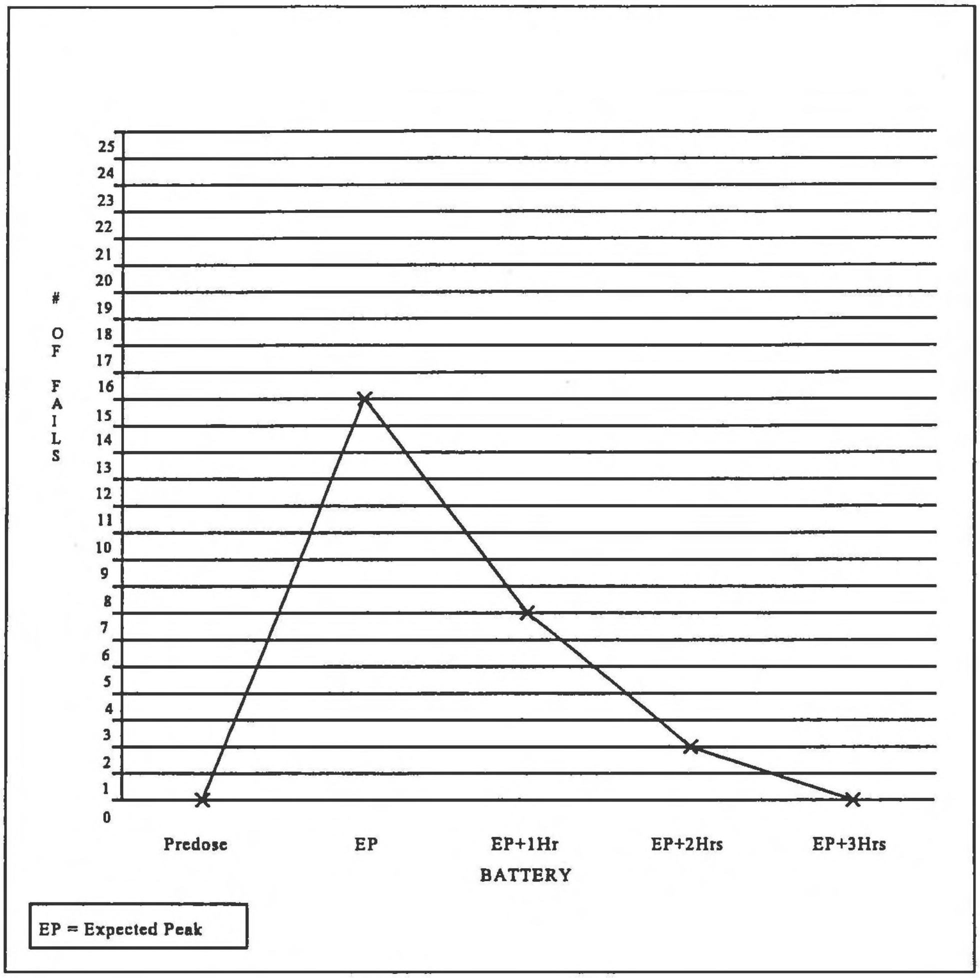


Figure 5-4. Number of Fails for FFD Test 1

Table 5-6. Pass/Fail Results for FFD Test 2 (+=Pass, X=Fail, X+= Pass)

Male Subjects	Predose (Mean BAC 0.00%)	EP (Mean BAC 0.077%)	EP+1 Hour (Mean BAC 0.062%)	EP+2 Hours (Mean BAC 0.048%)	EP+3 Hours (Mean BAC 0.031%)
1	XX	X+	+	+	+
2	+	X+	+	+	X+
3	+	+	+	+	+
4	+	XX	XX	X+	+
5	+	X+	+	+	+
6	+	+	+	+	+
7	+	+	X+	+	+
8	+	+	+	+	X+
9	+	+	+	+	+
10	+	XX	XX	XX	XX
11	+	X+	+	+	+
30	+	XX	XX	+	+
Total Passes	11	9	9	11	11
Total Fails	1	3	3	1	1
Female Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
13	+	XX	XX	+	+
14	+	+	+	+	+
15	+	XX	XX	XX	X+
16	+	X+	XX	X+	+
17	+	+	XX	+	+
18	+	+	+	+	+
19	+	XX	+	+	+
20	+	+	+	XX	XX
21	XX	XX	XX	XX	+
22	+	XX	+	XX	+
23	+	XX	XX	X+	+
24	+	+	+	+	+
25	+	XX	XX	XX	+
Total Passes	12	6	6	8	12
Total Fails	1	7	7	5	1
All Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
Total Passes	23	15	15	19	23
Total Fails	2	10	10	6	2

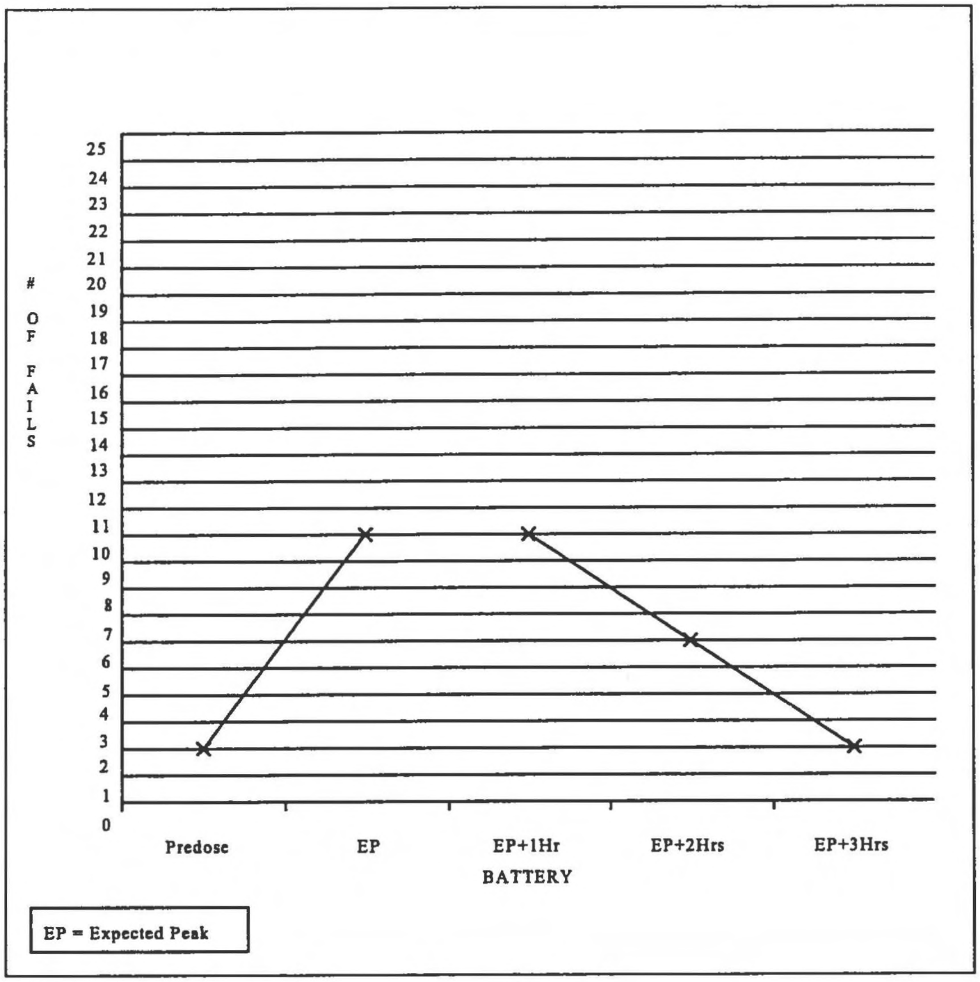


Figure 5-5. Number of Fails for FFD Test 2

Table 5-7. Pass/Fail Results for FFD Test 2, Without Retest (+=Pass, X=Fail)

Male Subjects	Predose (Mean BAC 0.00%)	EP (Mean BAC 0.077%)	EP+1 Hour (Mean BAC 0.062%)	EP+2 Hours (Mean BAC 0.048%)	EP+3 Hours (Mean BAC 0.031%)
1	X	X	+	+	+
2	+	X	+	+	X
3	+	+	+	+	+
4	+	X	X	X	+
5	+	X	+	+	+
6	+	+	+	+	+
7	+	+	X	+	+
8	+	+	+	+	X
9	+	+	+	+	+
10	+	X	X	X	X
11	+	X	+	+	+
30	+	X	X	+	+
Total Passes	11	5	8	10	9
Total Fails	0	7	4	2	3
Female Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
13	+	X	X	+	+
14	+	+	+	+	+
15	+	X	X	X	X
16	+	X	X	X	+
17	+	+	X	+	+
18	+	+	+	+	+
19	+	X	+	+	+
20	+	+	+	X	X
21	X	X	X	X	+
22	+	X	+	X	+
23	+	X	X	X	+
24	+	+	+	+	+
25	+	X	X	X	+
Total Passes	12	5	6	6	11
Total Fails	0	0	7	7	2
All Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
Total Passes	23	10	14	16	20
Total Fails	2	15	11	9	5

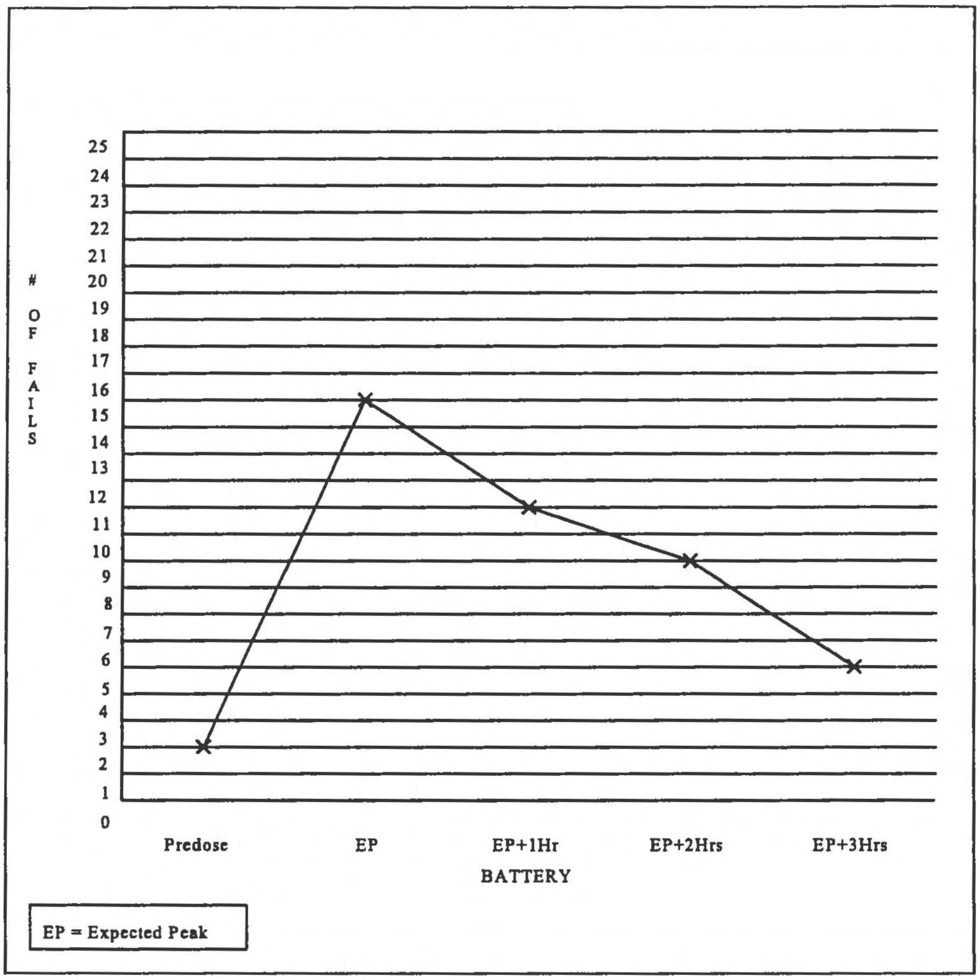


Figure 5-6. Number of Fails for FFD Test 2 Without Retest

Table 5-8. Pass/Fail Results for FFD Test 3 (+=Pass, X=Fail, X+= Pass)

Male Subjects	Predose (Mean BAC 0.00%)	EP (Mean BAC 0.077%)	EP+1 Hour (Mean BAC 0.062%)	EP+2 Hours (Mean BAC 0.048%)	EP+3 Hours (Mean BAC 0.031%)
1	+	+	+	+	+
2	+	+	+	+	+
3	+	XX	XX	+	+
4	+	+	XX	+	+
5	XX	+	XX	X+	XX
6	+	+	+	X+	+
7	+	+	XX	+	+
8	+	+	X+	+	+
9	+	+	+	+	+
10	+	+	+	+	+
11	+	+	+	+	+
30	+	+	+	X+	+
Total Passes	11	11	8	12	11
Total Fails	1	1	4	0	1
Female Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hours (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
13	+	+	+	+	+
14	+	+	+	+	+
15	+	+	+	+	+
16	+	+	XX		
17	+	+	+	+	+
18	+	+	+	+	+
19	+	+	X+	+	+
20	+	+	+	X+	+
21	+	+	+	+	+
22	+	+	+	+	+
23	+	XX	X+	+	+
24	+	+	+	+	+
25	+	+	+	+	+
Total Passes	13	12	12	12	12
Total Fails	0	1	1	0	0
All Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
Total Passes	24	23	20	24	23
Total Fails	1	2	5	0	1

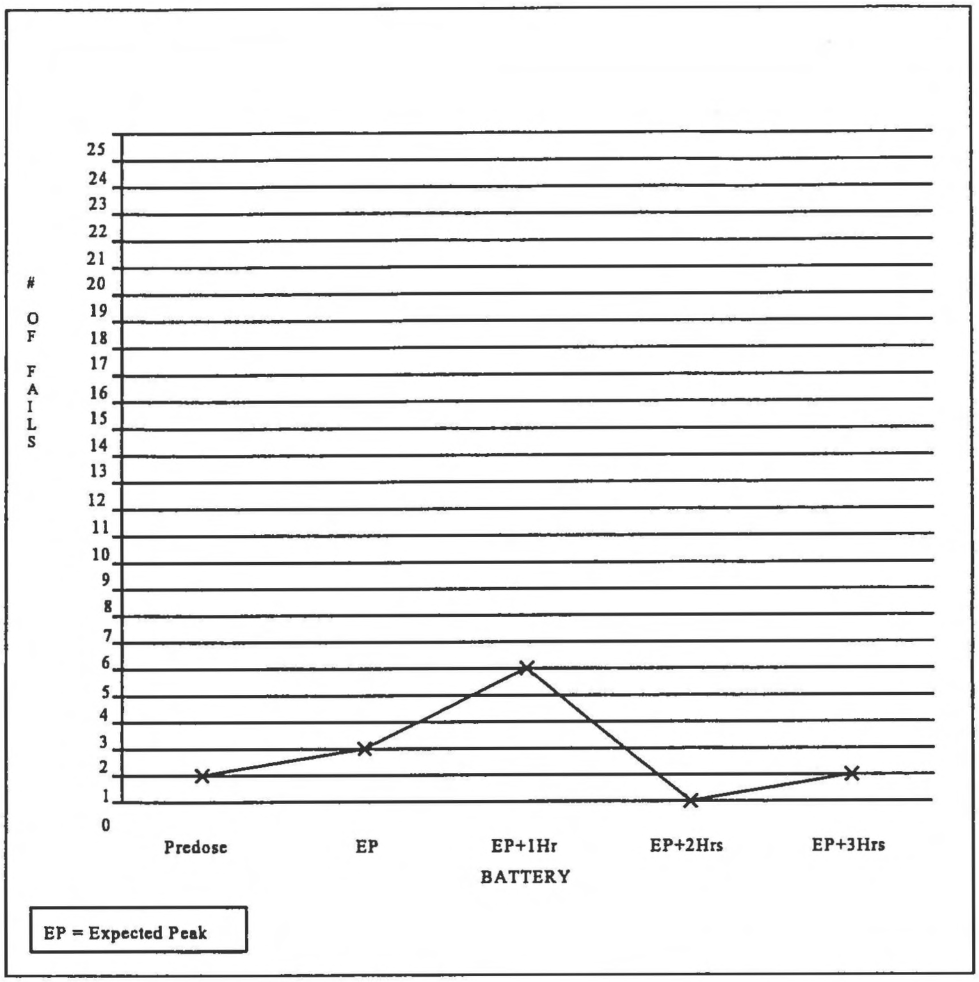


Figure 5-7. Number of Fails for FFD Test 3

Table 5-9. Pass/Fail Results for FFD Test 3: Modified Pass/Fail Criterion (+=Pass, X=Fail)

Male Subjects	Predose (Mean BAC 0.00%)	EP (Mean BAC 0.077%)	EP+1 Hour (Mean BAC 0.062%)	EP+2 Hours (Mean BAC 0.048%)	EP+3 Hours (Mean BAC 0.031%)
1	+	+	+	+	+
2	+	X	+	+	+
3	+	X	X	+	+
4	+	X	X	+	+
5	X	X	X	X	X
6	+	+	+	X	+
7	+	+	X	X	+
8	+	X	X	+	+
9	+	+	+	+	+
10	+	+	+	+	+
11	+	+	+	+	+
30	+	+	+	X	+
Total Passes	11	7	7	8	11
Total Fails	1	5	5	4	1
Female Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
13	+	+	+	+	+
14	+	+	+	+	+
15	+	+	+	+	+
16	+	+	X		
17	+	+	+	+	+
18	+	+	+	+	+
19	+	X	X	+	+
20	+	+	+	X	+
21	+	+	+	+	+
22	+	+	+	+	+
23	+	X	X	+	+
24	+	+	+	+	+
25	+	+	+	+	+
Total Passes	13	11	10	11	12
Total Fails	0	2	3	1	0
All Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
Total Passes	24	18	17	19	23
Total Fails	1	7	8	5	1

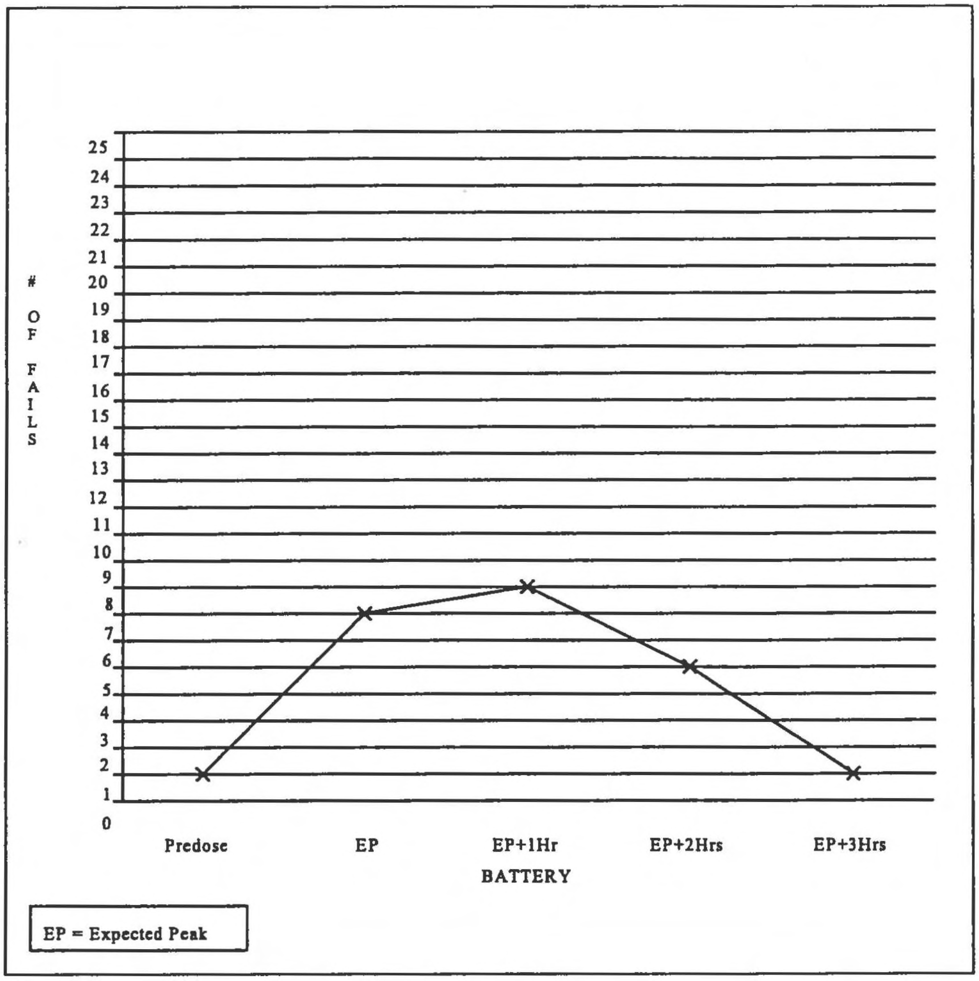


Figure 5-8. Number of Fails for FFD Test 3 With Modified Pass/Fail Criterion

Test 4

This system failed 23.37% of the subjects (Table 5-10, Figure 5-9) at the first test time (mean BAC 0.077%). At the second test time, the mean BAC was 0.062% and the rate decreased to 10.52%. At the third test time with a mean BAC of 0.047%, the system failed 21.05% of subjects, and at the last test time (mean BAC 0.023%), the system failed 36.84% of subjects. Of note, however, is the 21.05% Fail rate among subjects at 0.00% BACs (False Positive).

The multi–task approach to fitness-for-duty testing used by Test 4 makes it a highly viable system. Unfortunately, three factors may have contributed to the pilot experiment not achieving a valid evaluation of it. First, multiple trials over an extended period are needed to establish a valid baseline, and subjects did not achieve a stable performance level within the number of trials possible during two training sessions.

Second, after an employee establishes a baseline on Test 4, he or she thereafter will be expected to perform the test "as usual." Scores falling outside a specified region of the distribution of baseline scores, as a result of either unusually good or unusually poor performance, will yield a fail. Penalizing a test taker for good performance is contrary to the approach of much testing, and probably to the test–taking "set" of most individuals. It remains to be demonstrated that it is appropriate to the workplace and that it is compatible with fitness-for-duty objectives.

Given the "perform as usual" approach to testing, SCRI's training procedures, which encourage the best possible performance, may have been counterproductive. During the experiment, subjects often expressed frustration over repeated failures despite genuine efforts to excel. The effects on Test 4 scores of encouraging subjects to continue to perform at their highest capability cannot be determined.

Analysis revealed that approximately 45% of failures occurred because scores were significantly better than subjects' baseline scores. That is, subjects failed not because of poor performance but because they were continuing to improve. It is not possible to do a thorough analysis without raw scores, but it appears that steep learning curves obscure whatever changes in performance occurred with alcohol.

Table 5-10. Pass/Fail Results for FFD Test 4

Male Subjects	Predose (Mean BAC 0.00%)	EP (Mean BAC 0.077%)	EP+1 Hour (Mean BAC 0.062%)	EP+2 Hours (Mean BAC 0.048%)	EP+3 Hours (Mean BAC 0.031%)
1	X	+	+	+	+
2	+	+	+	+	+
3	+	+	+	+	+
4	+	+	+	+	+
5	+	+	+	+	X
6	DATA NOT PROCESSED				
7	+	X	+	X	X
8	+	X	+	X	+
9	DATA NOT PROCESSED				
10	+	X	+	+	X
11	+	+	+	+	X
30	+	+	+	+	X
Total Passes	9	7	10	8	5
Total Fails	1	3	0	2	5
Female Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
13	DATA NOT PROCESSED				
14	+	X	X	X	+
15	X	+	+	+	+
16	DATA NOT PROCESSED				
17	X	+	+	+	X
18	X	+	+	+	X
19	DATA NOT PROCESSED				
20	DATA NOT PROCESSED				
21	+	+	+	+	+
22	+	+	+	+	+
23	+	X	X	+	+
24	+	+	+	X	+
25	+	+	+	+	+
Total Passes	6	7	7	7	7
Total Fails	3	2	2	2	2
All Subjects	**Predose (Mean BAC 0.00%)**	**EP (Mean BAC 0.077%)**	**EP+1 Hour (Mean BAC 0.062%)**	**EP+2 Hours (Mean BAC 0.048%)**	**EP+3 Hours (Mean BAC 0.031%)**
Total Passes	15	14	17	15	12
Total Fails	4	5	2	4	7

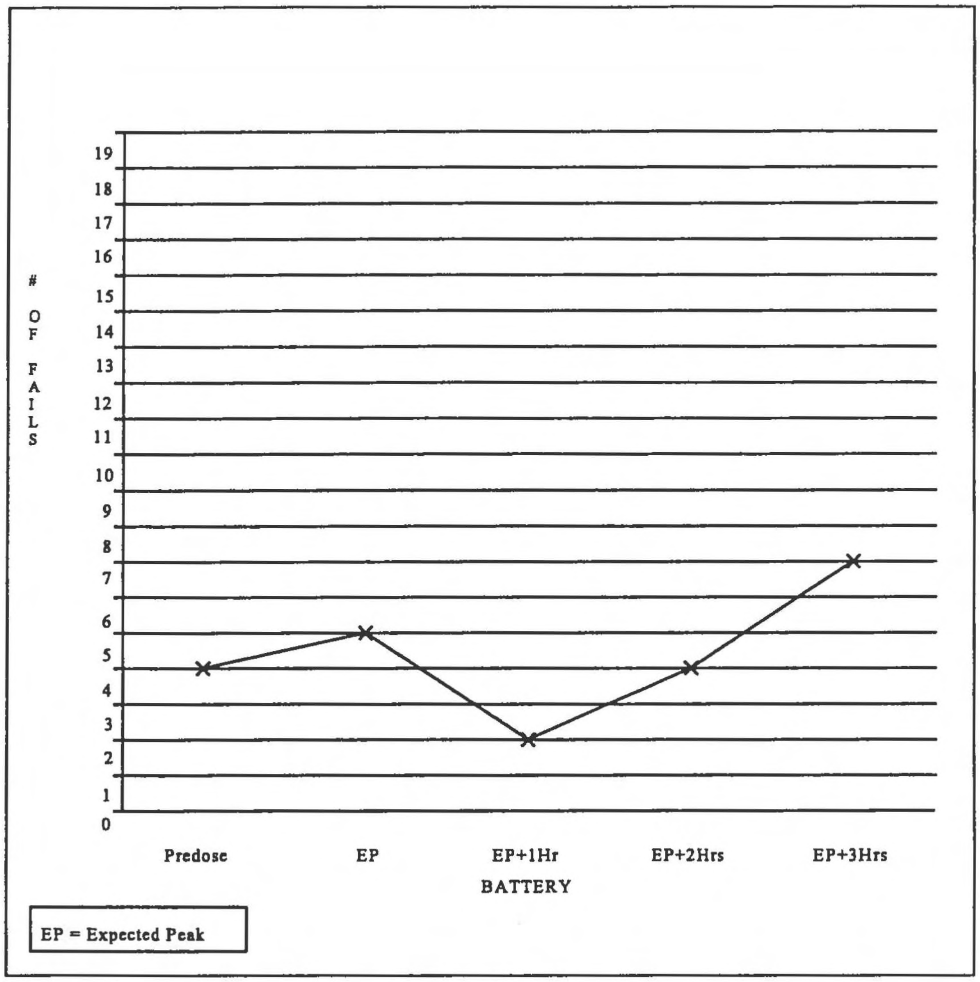


Figure 5-9. Number of Fails for FFD Test 4

Third, data were not processed for six subjects due to the equipment being turned off overnight. Normally, data processing routines are automatically executed at midnight.

This system merits further testing. A valid examination in a laboratory experiment will be difficult, however, because of the requirement for daily baseline trials over an extended period.

Test 5

This system was not used correctly in the pilot study, and obtained data are not a valid test. Procedural errors by SCRI resulted in insufficient data, and the relationship of BAC and pass/fail by the proprietary scoring of the system cannot be determined. Additional analyses indicate that the tests are alcohol sensitive, but the range of sensitivity cannot be established. Test 5 merits additional, properly executed examination.

Subjective Effects

Subjects evaluated their intoxication and impairment and rated the difficulty and interest of the tests (Table 5-11, Figure 5-10). Despite being at slightly higher BACs, the intoxication/impairment ratings by male subjects were consistently lower than the ratings by female subjects. On a 100 mm scale, half the men and three–quarters of the women rated their intoxication with a line at or above 70 mm.

Impairment ratings ranged from very low (10 mm, almost no impairment) to very high (94 mm, severe impairment). A rating, of course, reflects the individual's own criterion and may or may not be closely related to performance. Two men and one woman, who believed they were only slightly impaired, correctly assessed themselves on the performance–based tests. Interestingly, however, they failed FFD Test 1, but that test provides no feedback, and the failures could not have entered into their evaluations.

Some tests were viewed by subjects as being significantly more difficult than others (Chi Square = 14.00, 4 degrees of freedom, $p < .01$). Overall, FFD Test 2 was considered the most difficult test and FFD Test 4 was considered the most interesting test (see Figures 5-11 and 5-12), but men and women assessed the tests differently (Chi Square = 11.37, 3 degrees of freedom, $p < .01$). FFD Test 1 was rated most difficult by six women but by only one man. Seven men, but only four women, found FFD Test 2 to be the most difficult.

The results of the pilot study provided some preliminary evidence that fitness-for-duty testing may provide sufficient sensitivity to be of use in the transit environment. However, because of problems with the testing apparatus, complete data were not obtained for two of the tests. To better assess fitness-for-duty testing feasibility, additional data were needed. Consequently, a full experiment was performed, which is described in the next chapter.

Table 5-11. Subjective Measures

Male Subjects	Intoxication	Impairment	Most Difficult FFD Test	Most Interesting FFD Test
1	63	74	2	5
2	76	54	2	4
3	62	51	2	4
4	68	48	2	1, 2
5	18	18	2	1, 3
6	74	75	3	1
7	70	51	3	4
8	58	48	3	3, 4
9	35	10	2	1
10	77	84	2	5
11	80	28	1	2
30	70	94	3	1, 4
Mean	62.58	52.92		
Std. Dev.	18.42	25.81		
Female Subjects	**Intoxication**	**Impairment**	**Most Difficult FFD Test**	**Most Interesting FFD Test**
13	89	50	1	3
14	73	31	4	2
15	97	94	1	3
16	70	59	2	4
17	27	10	1	3
18	77	30	4	2
19	78	64	1	4
20	67	88	2	4
21	74	74	1	5, 3
22	32	32	2	1
23	80	82	4	3
24	70	84	1	2
25	79	77	2	2, 5
Mean	70.23	59.62		
Std. Dev.	19.81	26.81		
All Subjects	**Intoxication**	**Impairment**		
Mean	66.04	55.54		
Std. Dev.	19.39	26.20		

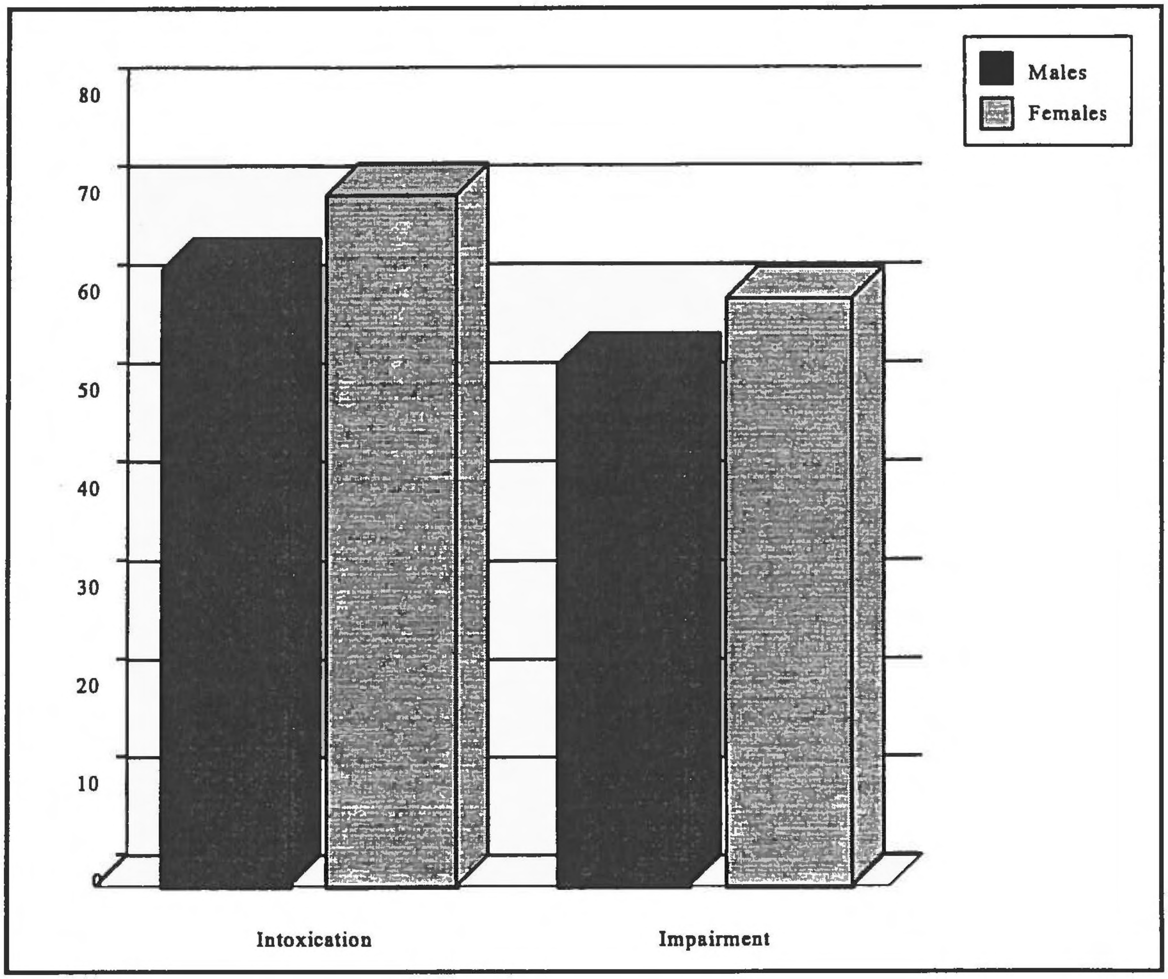
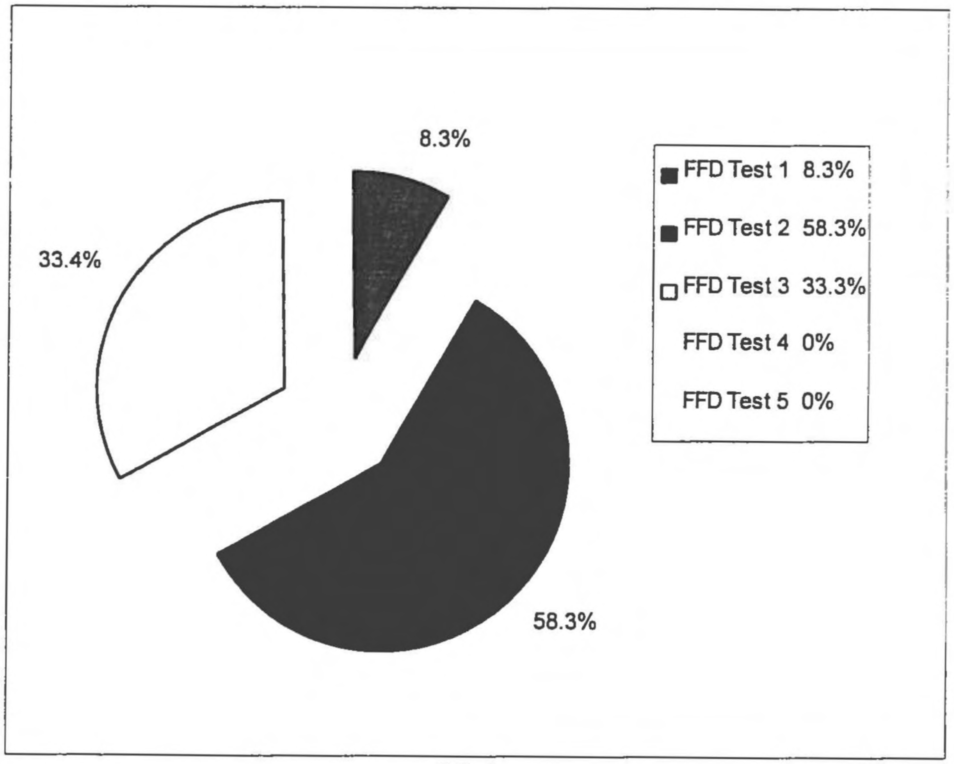
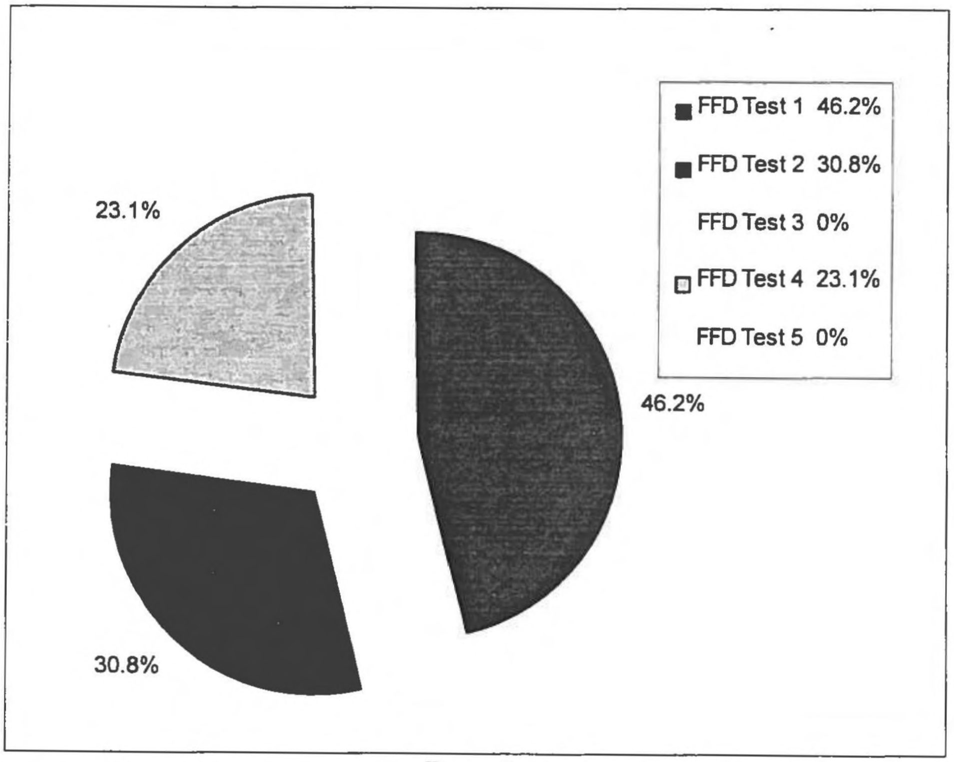


Figure 5-10. Subjects' Rating of Intoxication and Impairment on a Visual Analogue Scale (0="Not at all", 100="Very")

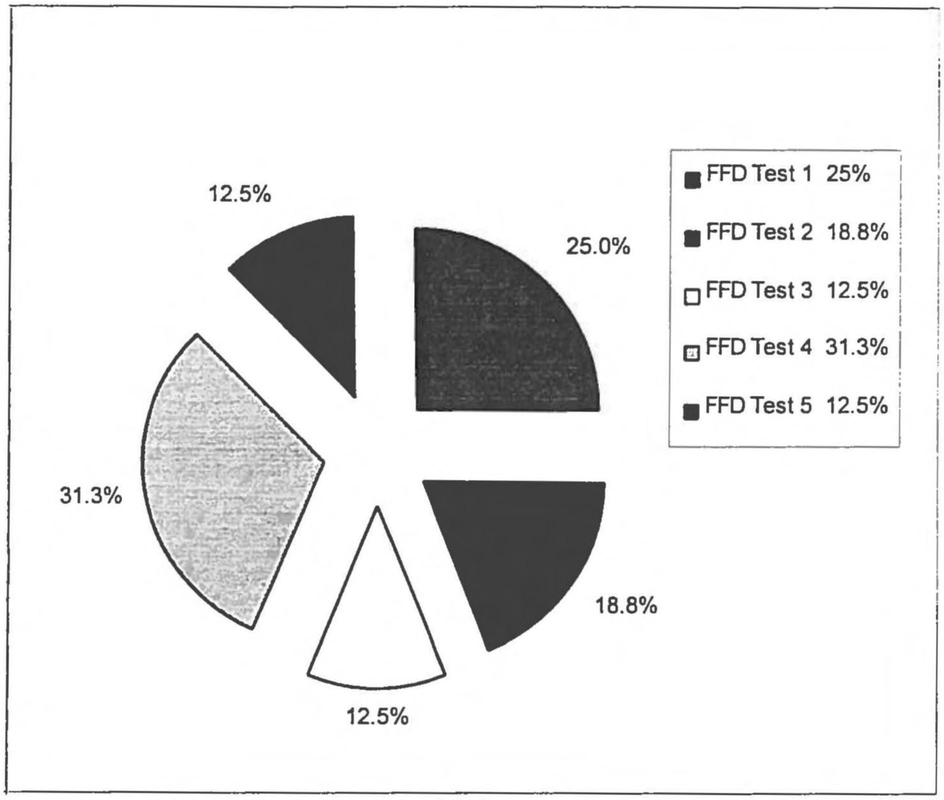


Male

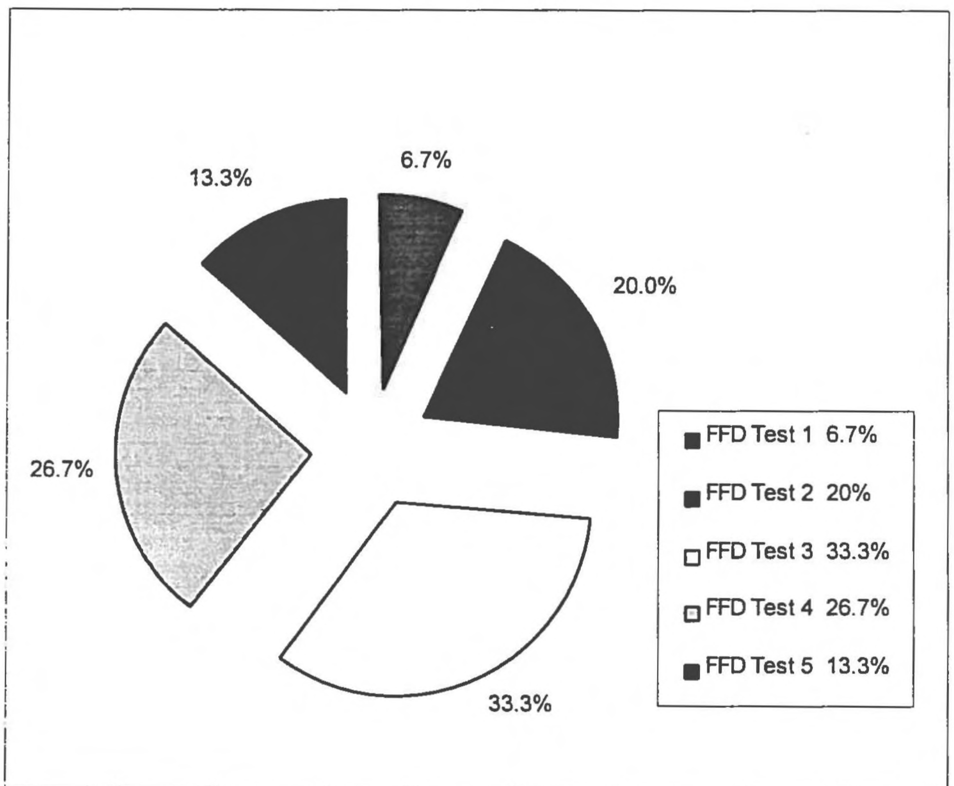


Female

Figure 5-11. "Most Difficult Test"



Male



Female

Figure 5-12. "Most Interesting Test"

CHAPTER SIX

LABORATORY STUDY

The findings of the pilot study, as previously reported, supported a continuation of the evaluation of FFD tests. Five vendors who participated in the pilot study, and one vendor who had been unable to provide apparatus at that time, were invited to participate in the second phase of the study. Four agreed to participate.

As in the pilot study, measured BAC served as the standard of impairment. Unlike the pilot study however, each subject was given three treatments: A, B, and C. In Treatment A, or placebo, no measurable alcohol was administered, and all measured BACs were 0.00%. In Treatments B and C the target peak BAC was 0.08%. The two separate treatments with identical target BACs yielded a preliminary measure of test-retest reliability. As in the pilot study, each subject was tested five times at each treatment.

Experimental Design

In a repeated measure design, 24 subjects participated in training and test/treatment sessions. Figure 6-1 illustrates the 3x5 design.

Subjects

Twenty-four healthy men and women (12 men, 12 women) were recruited to participate as paid volunteer subjects. Applicants responded to advertisements in community newspapers and were screened in terms of their health history, current health status, drug and alcohol use. Applicants who scored as moderate to low-heavy drinkers on the Cahalan, Cisin, and Crossley Quantity-Frequency-Variability scale (1969) and whose Minnesota Multiphasic Personality Inventory (MMPI) profiles reflected emotional stability were enrolled as subjects.

Apparatus

Four vendors provided fitness-for-duty tests for the experiment. The four tests, two performance and two physiological, were a balanced sample of current fitness-for-duty technology.

BATTERY	TEST	TREATMENT A	TREATMENT B	TREATMENT C
Predose	Test 1	Ss 1-24	Ss 1-24	Ss 1-24
	Test 4	Ss 1-24	Ss 1-24	Ss 1-24
	Test 5	Ss 1-24	Ss 1-24	Ss 1-24
	Test 6	Ss 1-24	Ss 1-24	Ss 1-24
Expected Peak (EP)	Test 1	Ss 1-24	Ss 1-24	Ss 1-24
	Test 4	Ss 1-24	Ss 1-24	Ss 1-24
	Test 5	Ss 1-24	Ss 1-24	Ss 1-24
	Test 6	Ss 1-24	Ss 1-24	Ss 1-24
EP+1 Hour	Test 1	Ss 1-24	Ss 1-24	Ss 1-24
	Test 4	Ss 1-24	Ss 1-24	Ss 1-24
	Test 5	Ss 1-24	Ss 1-24	Ss 1-24
	Test 6	Ss 1-24	Ss 1-24	Ss 1-24
EP+2 Hours	Test 1	Ss 1-24	Ss 1-24	Ss 1-24
	Test 4	Ss 1-24	Ss 1-24	Ss 1-24
	Test 5	Ss 1-24	Ss 1-24	Ss 1-24
	Test 6	Ss 1-24	Ss 1-24	Ss 1-24
EP+3 Hours	Test 1	Ss 1-24	Ss 1-24	Ss 1-24
	Test 4	Ss 1-24	Ss 1-24	Ss 1-24
	Test 5	Ss 1-24	Ss 1-24	Ss 1-24
	Test 6	Ss 1-24	Ss 1-24	Ss 1-24

Figure 6-1. Experimental Design

The vendors were responsible for installing and maintaining their equipment, training the SCRI staff, and providing support. Prior to beginning the data collection phase, two steps were taken to guarantee proper apparatus use.

Test Protocols

A protocol was written for each fitness-for-duty test to provide SCRI personnel with instructions for:

- Establishing subjects' baseline
- Administering tests
- Creating/maintaining subjects' data files
- Collecting data
- Backing up data files.

The protocols also insured that test systems were used in compliance with vendors' instructions. Protocols were developed at SCRI and submitted to the vendors for review. All vendors approved the protocols.

Trial. To demonstrate proper administration of tests and to detect procedural problems before beginning the experiment, a single subject was tested under placebo and alcohol conditions. Upon completion, each vendor was provided with the data obtained with its apparatus and was asked to evaluate the quality of the results. None of the vendors reported problems with the data.

Procedures

As shown in Table 6-1, a subject's schedule for the study spanned four weeks. Training days were scheduled at least a day apart. Treatment sessions were at least two but no more than three days after the last training session and were separated by a week. At all sessions, subjects provided urine and breath specimens to test for recent drug use and alcohol consumption. The subject's blood pressure and pulse rate were measured and recorded. Urine specimens obtained from female subjects were tested for pregnancy prior to treatment administration. No pregnancy test was positive.

Table 6-1. Subject's Schedule

Week	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1		Training 1 (4-6 Hours)		Training 2 (4-6 Hours)		Training 3 (4-6 Hours)	
2		Treatment 1 (8 Hours)					
3		Treatment 2 (8 Hours)					
4		Treatment 3 (8 Hours)					

Training Sessions. Upon arriving at the testing facility for training session 1, subjects were asked to read and sign a copy of the Subject's Bill of Rights, as required by California law, and an Informed Consent document, and the requirements of the study were explained in detail. There were three test periods at each training session, separated by a 15-minute rest period. Subjects were monitored at all times during training, and were given standard levels of instruction and feedback.

Treatment Sessions

On treatment days, subjects were transported by taxi between their residences and SCRI. Upon arrival, they completed a questionnaire which inquired about the following:

- Sleep over the preceding 24 hours
- Food and stimulant intake during the preceding 4 hours
- Alcohol consumption during the preceding 48 hours
- Illicit, prescription, OTC drug use during the preceding 7 days
- Present health status
- Present stress levels.

The battery of four fitness-for-duty tests was repeated five times during treatment days, as shown below:

Battery	Name	Time
1	Predose	Before drinking and at 0.00% BAC
2	Expected Peak (EP)	30 minutes after end of last drink (0.00% BAC for Treatment A and approximately 0.08% BAC for Treatments B&C)
3	EP + 1 Hour	1 hour after EP
4	EP + 2 Hours	2 hours after EP
5	EP + 3 Hours	3 hours after EP

Before and after each battery, the subjects' BAC, blood pressure, and pulse rate were measured. After the final battery, subjects completed a final questionnaire and were transported home when their BACs declined to 0.00.

Treatments. The alcohol treatments were .68 g absolute alcohol/Kg body weight for male subjects and .58 g absolute alcohol/Kg body weight for female subjects. The drinks were administered as a 1:1.5 mixture of 80 proof vodka and orange juice.

The placebo beverage was a 1:1.5 mixture of water and orange juice in the same total volume as the alcohol beverage. To provide an initial odor and taste of alcohol, 10 mL vodka was floated on the orange juice-water mixture and the rim of the glass was rubbed with vodka-soaked, sterile cotton.

The total alcohol and placebo beverage was given as three equal drinks at 10-minute intervals. Subjects were instructed to pace their drinking evenly so as to complete each drink within 10 minutes. They were under continual observation during the drinking and absorption period.

A period of 30 minutes, intended to allow for full absorption of the alcohol, followed the termination of the third drink. After the absorption period, BACs were measured with a breath sampling Intoxilyzer 5000. If the BAC was 0.08% +/- 0.01%, the first test battery was initiated. If the reading exceeded 0.09%, testing was delayed to allow the BAC to decline. If the BAC was less than 0.07%, a second breath sample was obtained after 15 minutes to determine whether additional time was required due to slow absorption or whether additional alcohol was required. The initiation of testing was not delayed for more than one hour.

Objective and Subjective Measures

The following objective and subjective measures were used.

Fail:	Detection of impairment as measured by the fitness-for-duty apparatus
Pass:	No detection of impairment as measured by the fitness-for-duty apparatus
Invalid:	A test that did not yield a result
Consistency Index:	The number of identical test results (i.e., Pass or Fail) between Treatments B and C, divided by the total number of comparisons (120)
Fail Rate:	The number of Fails divided by the sum of Passes, Fails, and Invalids, and multiplied by 100
False Positive:	A Fail when the measured BAC was below the defined impairment criterion
False Negative:	A Pass when the measured BAC was at or above the defined impairment criterion
Hit:	A Fail when the measured BAC was at or above the defined impairment criterion.
Correct Rejection:	A Pass when the measured BAC was below the defined impairment criterion
Self-report of Intoxication:	Subjects rated their intoxication on a 0 to 100 mm visual analogue scale (VAS)

Self-report of Impairment.	Subjects rated their level of impairment on a 0 to 100 mm VAS
Difficulty of Use Rating:	Subjects rated each test's difficulty on a 1 to 10 scale ("very easy" to "very difficult")
Personal Interest Rating:	Ratings of tests' interest to subjects on a 1 to 10 scale ("very interesting" to "not at all interesting")

Results

Subject Characteristics

Table 6-2 and Figures 6-2 to 6-4 summarize subjects' age, weight, height, Q-F-V category, ethnic group, and whether English was a subject's second language.

Subject 2 was dismissed after his third training session because carboxy-THC, a metabolite of tetrahydrocannabinol, was detected in his urine. Subjects 10, 12, and 15 terminated their participation due to scheduling conflicts. Subjects 11, 20, 21, and 25 did not successfully complete their training and baselining. Subject 27, who could not perform some performance tests, revealed that he was dyslexic and was dismissed. Subjects 7, 8, 16 and 32 had a slower than normal absorption rate.

Alcohol Data

As shown in Tables 6-3 and 6-4 and Figure 6-5, the average peak BACs for the alcohol Treatments B and C for all 24 subjects was 0.080% and 0.079%, respectively. In Treatment B, the mean peak BAC for both males and females was 0.08%. In Treatment C, on the other hand, the mean peak BAC for males was 0.077% and the mean peak BAC for females was 0.081%.

The mean rate of metabolism across subjects and treatments was 0.017% BAC per hour.

Analysis of Variance

In this experiment, impairment varied as a function of BAC. BAC, in turn, varied as a function of both Treatment (alcohol/no alcohol) and Battery (before drinking/after drinking). An effective fitness-for-duty test would, therefore, not fail subjects throughout Treatment A and at the Predose batteries of both Treatments B and C; would fail the most subjects at the Expected Peak (EP) battery of Treatments B and C; and would linearly decrease the number of Fails over batteries EP+1 hour, EP+2 hours, and EP+3 hours. If such a pattern of number of Fails occurred with a fitness-for-duty test in the laboratory study, meaning that Fails varied as a function of the

Table 6-2. Subject Characteristics

Male Subjects	Age (Years)	Weight (Lbs.)	Height (Inches)	Q-F-V Rating	Ethnic Group	English as 2nd Language
1	27	191	71	Low Moderate	African-American	No
3	26	149	69	Mid Moderate	Caucasian	No
4	25	173	71	Mid Moderate	Caucasian	No
5	38	183	71	Low Heavy	Caucasian	No
6	33	196	78	Low Heavy	Caucasian	No
7	30	184	71	Mid Moderate	Latino	Yes
8	25	196	73	Low Moderate	African-American	No
9	27	169	69	Low Moderate	Native American	No
26	27	170	70	Mid Moderate	Caucasian	No
28	32	171	68	Low Heavy	Latino	No
29	26	114	63	Mid Moderate	Caucasian	Yes
38	27	135	69	Low Heavy	Caucasian	No
Mean	28.6	169.2	70.2			
Std. Dev.	3.9	25.2	3.5			
Female Subjects	**Age (Years)**	**Weight (Lbs.)**	**Height (Inches)**	**Q-F-V Rating**	**Ethnic Group**	**English as 2nd Language**
13	26	119	62	Mid Moderate	Latino	Yes
14	28	182	66	Low Moderate	Caucasian	No
16	25	186	70	Low Heavy	Latino	No
17	26	141	65	Low Heavy	Asian	No
18	29	214	65	Low Moderate	Latino	No
19	25	125	64	Mid Moderate	Caucasian	Yes
22	39	118	63	Mid Moderate	Caucasian	No
23	28	155	68	Low Heavy	Caucasian	Yes
24	33	131	64	Mid Moderate	African-American	No
30	30	163	61	Low Moderate	Latino	No
31	27	152	68	Low Moderate	African-American	No
32	25	144	64	Low Heavy	Caucasian	No
Mean	28.4	152.5	65.0			
Std. Dev.	4.1	29.6	2.6			
All Subjects	**Age (Years)**	**Weight (Lbs.)**	**Height (Inches)**			
Mean	28.5	160.9	67.6			
Std. Dev.	3.9	28.2	4.0			

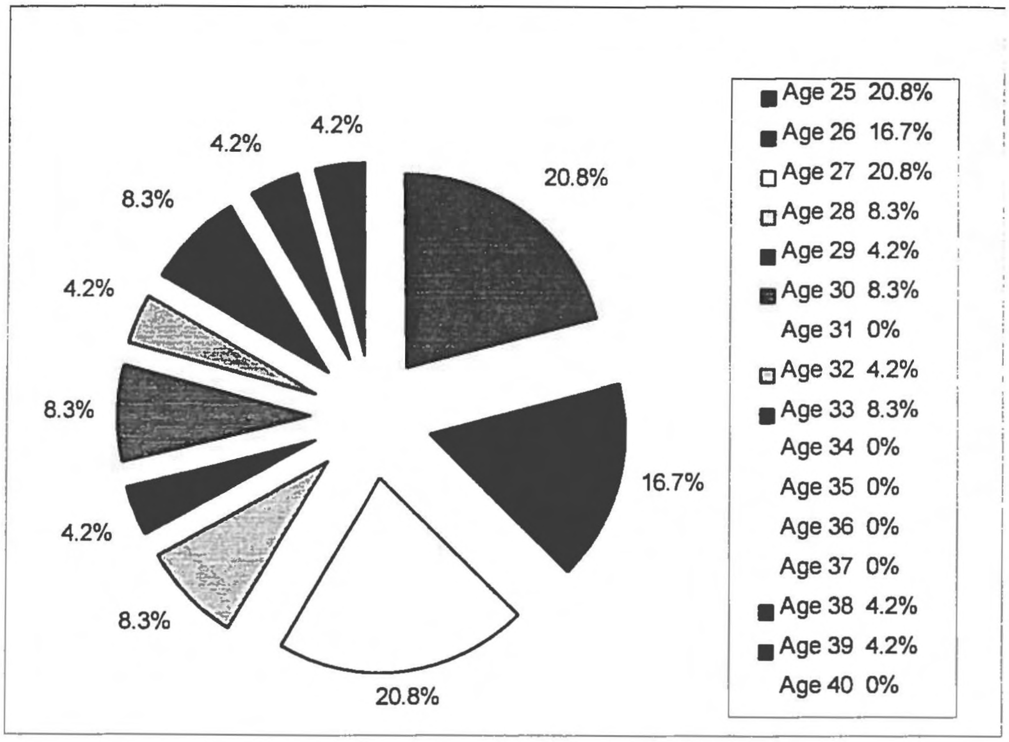


Figure 6-2. Subjects' Ages

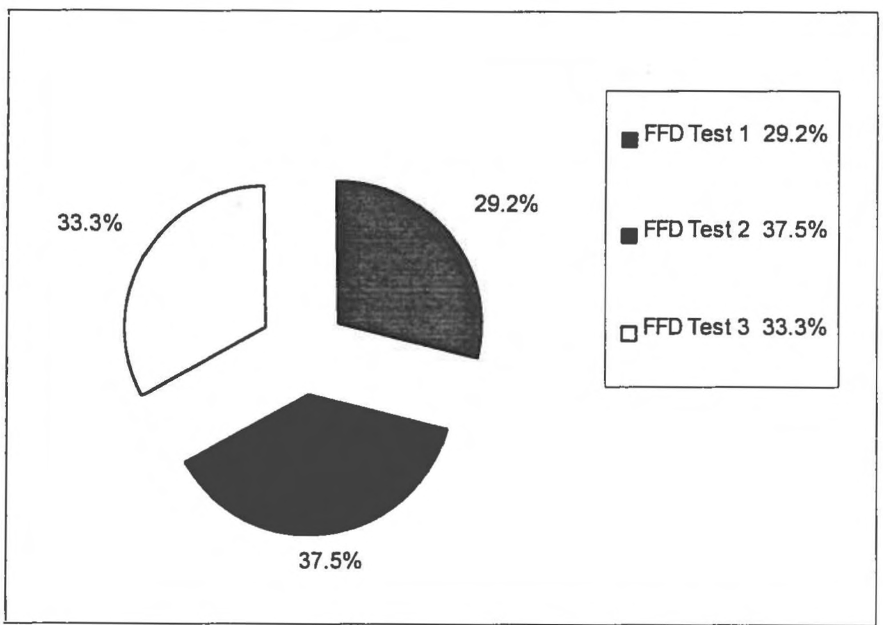


Figure 6-3. Subjects' Q-F-V Categories

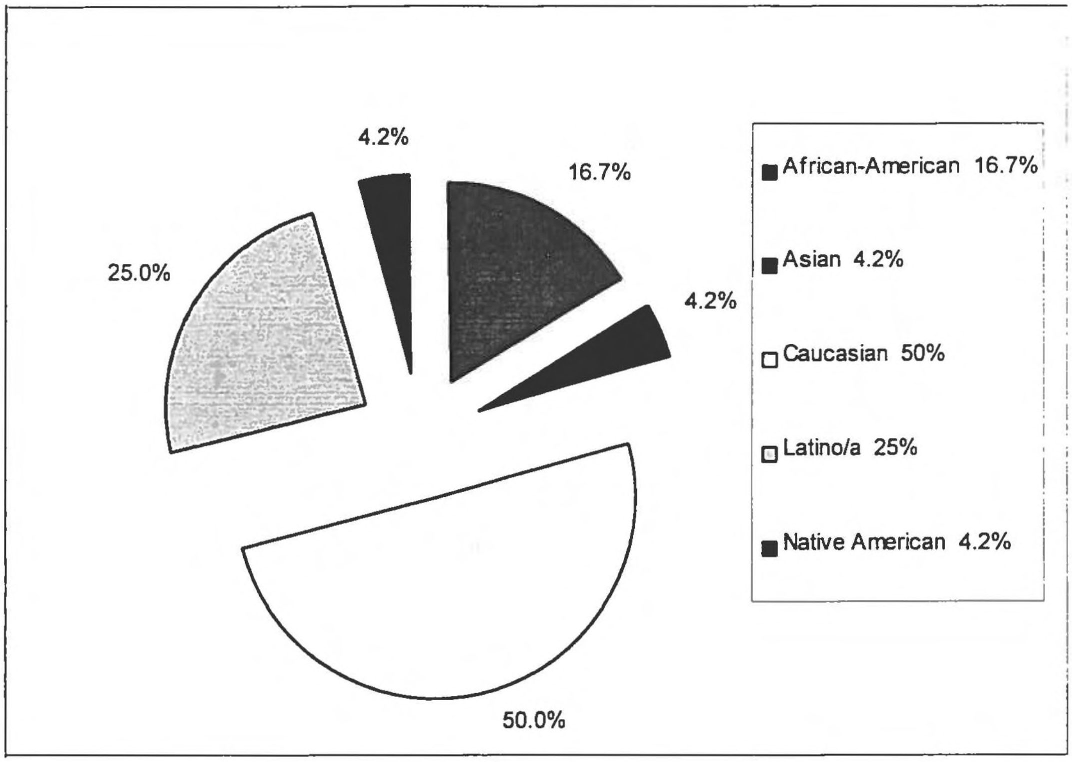


Figure 6-4. Subjects' Ethnic Groups

combined effect of Treatment and Battery, it would be detected by the analysis of variance (ANOVA) and would result in a statistically significant *Treatment by Battery (TxB) Interaction.* A 3x5 within-subject factorial ANOVA was performed to determine whether the number of Fails for each of the fitness-for-duty tests varied as a function of the TxB interaction. The analyses were performed with the statistical software packages BMDP2V and BMDP4V. Table 6-5 reports the results of the statistical analyses.

FFD Test 1. The number of Fails varied as a function of the combined effect of Treatment and Battery. In Treatment A (placebo) the number of Fails remained stable over the five batteries. In Treatments B and C, on the other hand, statistically significant differences in the number of Fails occurred across the five batteries. As expected, the number of Fails rose between Predose (no alcohol) and Expected Peak (EP) followed by a linear drop between EP and EP+1 hour, EP+2 hours, and EP+3 hours. The number of Fails for FFD Test 1 are reported in Figure 6-6.

FFD Test 4. Although the ANOVA detected the presence of a statistically significant TxB interaction, there were no differences in the number of Fails among batteries in Treatments B and C, and the number of Fails did not describe the BAC pattern. Therefore, FFD Test 4 cannot be considered a viable candidate for fitness-for-duty testing at this time. The number of Fails for FFD Test 4 are reported in Figure 6-7.

Table 6-3. Treatment B BACs

Male Subjects	Predose	Expected Peak (EP)	EP+1 Hour	EP+2 Hours	EP+3 Hours
3	0.000	0.073	0.078	0.064	0.051
4	0.000	0.084	0.073	0.051	0.037
5	0.000	0.082	0.060	0.037	0.017
6	0.000	0.084	0.073	0.067	0.046
7	0.000	0.080	0.082	0.059	0.044
8	0.000	0.075	0.080	0.069	0.061
9	0.000	0.088	0.069	0.052	0.045
26	0.000	0.076	0.060	0.046	0.027
28	0.000	00.082	0.071	0.048	0.030
29	0.000	0.081	0.051	0.033	0.013
38	0.000	0.075	0.053	0.045	0.027
Mean	0.000	0.080	0.067	0.051	0.035
Std. Dev.	0.000	0.005	0.011	0.012	0.015

Female Subjects	Predose	Expected Peak (EP)	EP+1 Hour	EP+2 Hours	EP+3 Hours
13	0.000	0.078	0.058	0.041	0.024
14	0.000	0.083	0.054	0.044	0.025
16	0.000	0.097	0.084	0.061	0.044
17	0.000	0.084	0.061	0.041	0.021
18	0.000	0.069	0.060	0.044	0.029
19	0.000	0.082	0.073	0.053	0.032
22	0.000	0.075	0.054	0.034	0.018
23	0.000	0.074	0.057	0.038	0.018
24	0.000	0.079	0.067	0.049	0.040
30	0.000	0.075	0.062	0.046	0.028
31	0.000	0.083	0.060	0.054	0.036
32	0.000	0.082	0.067	0.043	0.025
Mean	0.000	0.080	0.063	0.046	0.028
Std. Dev.	0.000	0.007	0.009	0.008	0.008

All Subjects	Predose	Expected Peak (EP)	EP+1 Hour	EP+2 Hours	EP+3 Hours
Mean	0.000	0.080	0.065	0.048	0.032
Std. Dev.	0.000	0.006	0.010	0.010	0.012

Table 6-4. Treatment C BACs

Male Subjects	Predose	Expected Peak (EP)	EP+1 Hour	EP+2 Hours	EP+3 Hours
1	0.000	0.075	0.059	0.044	0.026
3	0.000	0.069	0.047	0.038	0.025
4	0.000	0.086	0.057	0.046	0.031
5	0.000	0.073	0.047	0.033	0.013
6	0.000	0.073	0.049	0.032	0.015
7	0.000	0.080	0.063	0.046	0.033
8	0.000	0.070	0.072	0.057	0.046
9	0.000	0.089	0.071	0.056	0.045
26	0.000	0.086	0.067	0.048	0.032
28	0.000	0.083	0.059	0.044	0.026
29	0.000	0.067	0.049	0.025	0.004
38	0.000	0.078	0.060	0.049	0.036
Mean	0.000	0.077	0.058	0.043	0.028
Std. Dev.	0.000	0.007	0.009	0.010	0.012
Female Subjects	**Predose**	**Expected Peak (EP)**	**EP+1 Hour**	**EP+2 Hours**	**EP+3 Hours**
13	0.000	0.071	0.049	0.038	0.022
14	0.000	0.081	0.059	0.043	0.027
16	0.000	0.086	0.069	0.042	0.026
17	0.000	0.084	0.071	0.052	0.027
18	0.000	0.083	0.067	0.051	0.033
19	0.000	0.084	0.063	0.046	0.034
22	0.000	0.075	0.059	0.038	0.020
23	0.000	0.086	0.065	0.047	0.024
24	0.000	0.076	0.059	0.049	0.035
30	0.000	0.082	0.057	0.035	0.016
31	0.000	0.072	0.052	0.041	0.028
32	0.000	0.091	0.082	0.059	0.053
Mean	0.000	0.081	0.063	0.045	0.029
Std. Dev.	0.000	0.006	0.009	0.007	0.010
All Subjects	**Predose**	**Expected Peak (EP)**	**EP+1 Hour**	**EP+2 Hours**	**EP+3 Hours**
Mean	0.000	0.079	0.061	0.044	0.028
Std. Dev.	0.000	0.007	0.009	0.008	0.011

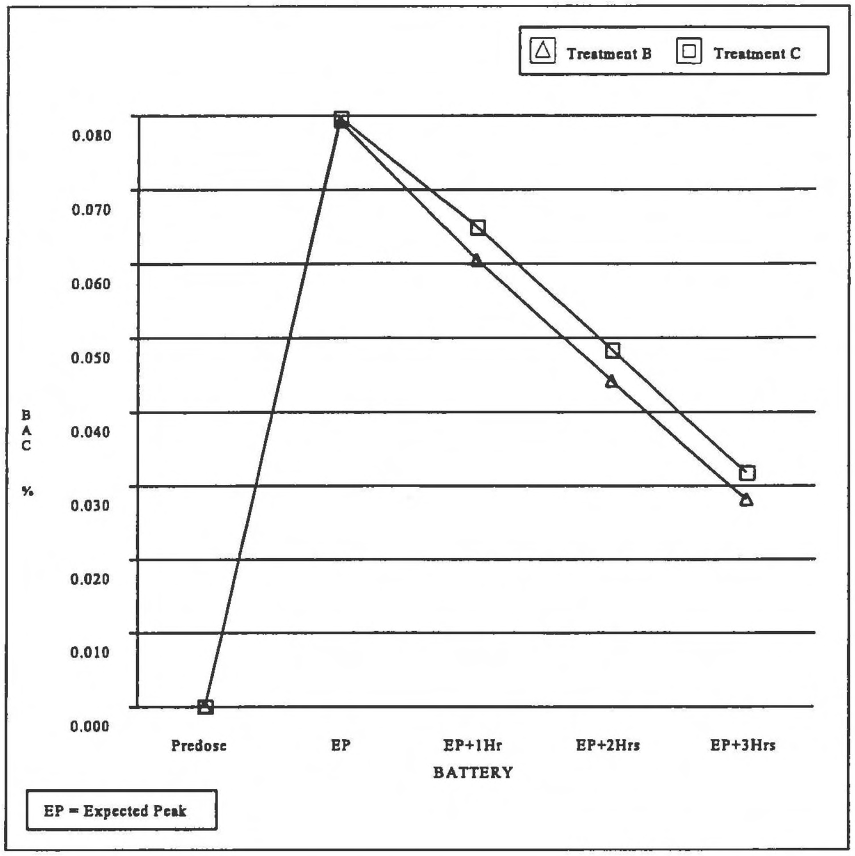


Figure 6-5. Average Peak BACs for the Alcohol Treatments

FFD Test 5. The ANOVA indicated that the number of Fails did not vary as a function of the combined effect of Treatment and Battery. In view of Figure 6-8, however, such results are puzzling. It is possible that a high Pass/Fail criterion might have skewed the data. To further examine the issue, additional analyses were undertaken with raw scores rather than the Pass/Fail results.

When the ANOVA was performed using the raw scores, the number of Fails were found to vary as a function of the TB interaction. In Treatment A (placebo) the number of Fails remained stable over the five batteries. In Treatments B and C the number of Fails rose between Predose and EP, but was not followed by a linear drop between EP and EP+1 hour, EP+2 hours, and EP+3 hours, indicating a lack of sensitivity at lower BACs. Figure 6-9 shows the interaction.

FFD Test 6. The number of Fails did not vary as a function of the combined effect of Treatment and Battery. Figure 6-10 reports the number of Fails for FFD Test 6.

Table 6-5. Statistical Analysis

ANOVAs

	Source	df	F	p
FFD Test 1	Treatment (T)	2, 44	25.47	0.0000
	Battery (B)	4, 88	27.71	0.0000
	TB	8, 176	7.95	0.0000
FFD Test 4*	Treatment (T)	2, 42	2.07	0.1390
	Battery (B)	4, 84	0.61	0.6571
	TB	8, 168	3.08	0.0029
FFD Test 5** Pass/Fail Data	Treatment (T)	2, 42	8.80	0.0006
	Battery (B)	4, 84	6.91	0.0001
	TB	8, 168	1.53	0.1503
FFD Test 5** Raw Data	Treatment (T)	2, 42	9.43	0.0004
	Battery (B)	4, 84	15.69	0.0000
	TB	8, 168	2.23	0.0278
FFD Test 6	Treatment (T)	2, 44	0.05	0.9515
	Battery (B)	4, 88	1.44	0.2277
	TB	8, 176	0.75	0.6476

* = S 38 reported test malfunction on EP+3 Hours of Treatment C. Although no anomalies were found, data was not included in the analysis.

**= S 22 reported test malfunction on Predose of Treatment C. Although no anomalies were found, data was not included in the analysis.

Simple Main Effects (with adjusted alpha level of 0.001)

	Source	df	F	p
FFD Test 1	B at T A	4, 88	0.65	0.6271
	B at T B	4, 88	20.85	0.0000
	B at T C	4, 88	15.90	0.0000
FFD Test 4	B at T A	4, 88	1.35	0.2563
	B at T B	4, 88	2.59	0.0422
	B at T C	4, 88	1.82	0.1321
FFD Test 5 Raw Data	B at T A	4, 84	0.73	0.5709
	B at T B	4, 84	11.57	0.0000
	B at T C	4, 84	7.45	0.0000

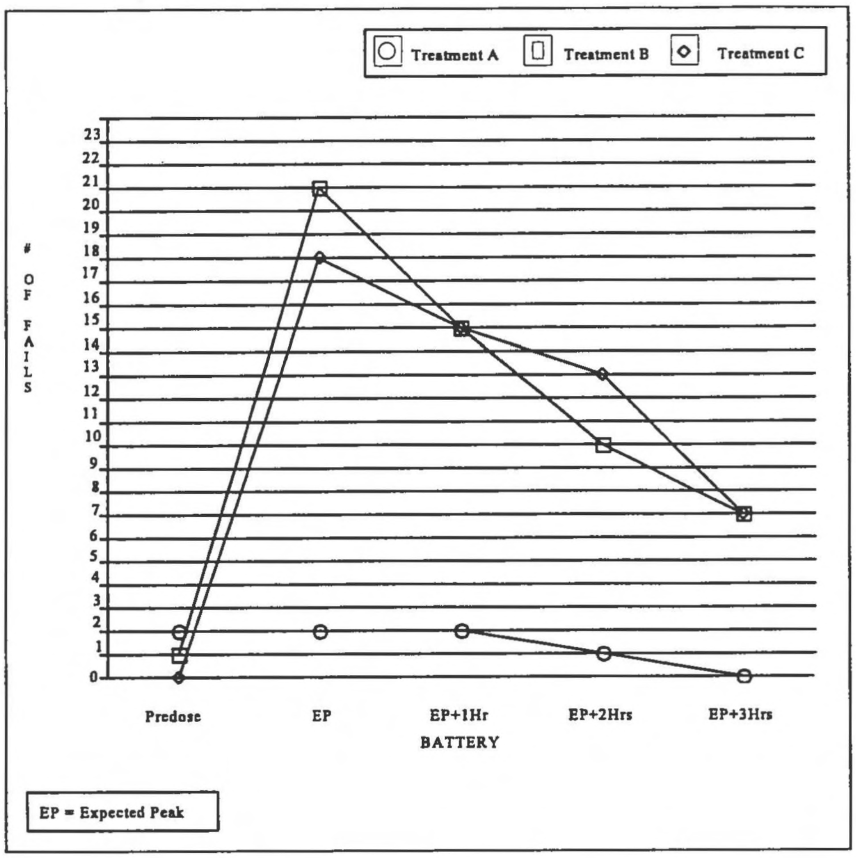


Figure 6-6. Number of Fails for FFD Test 1

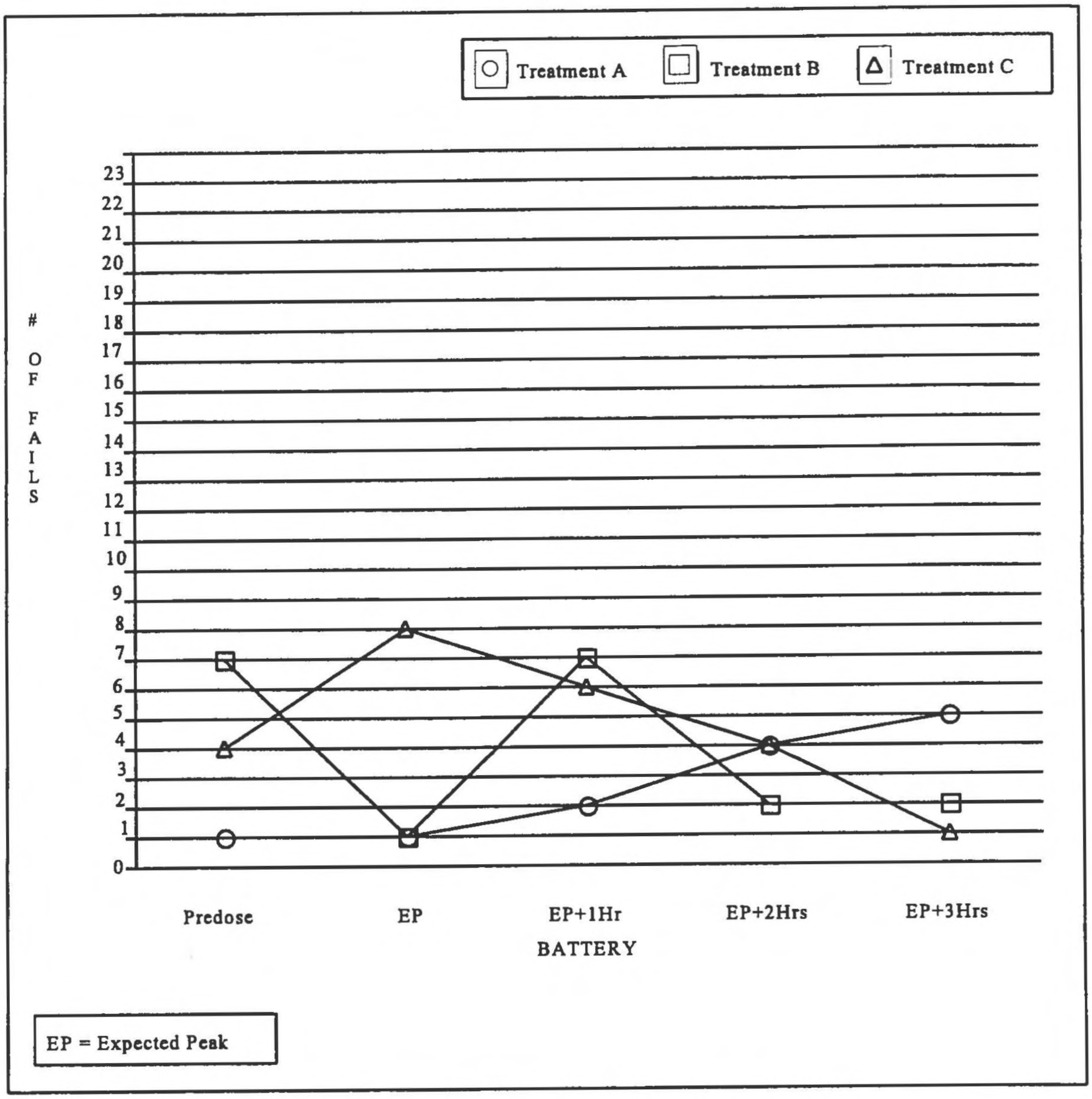


Figure 6-7. Number of Fails for FFD Test 4

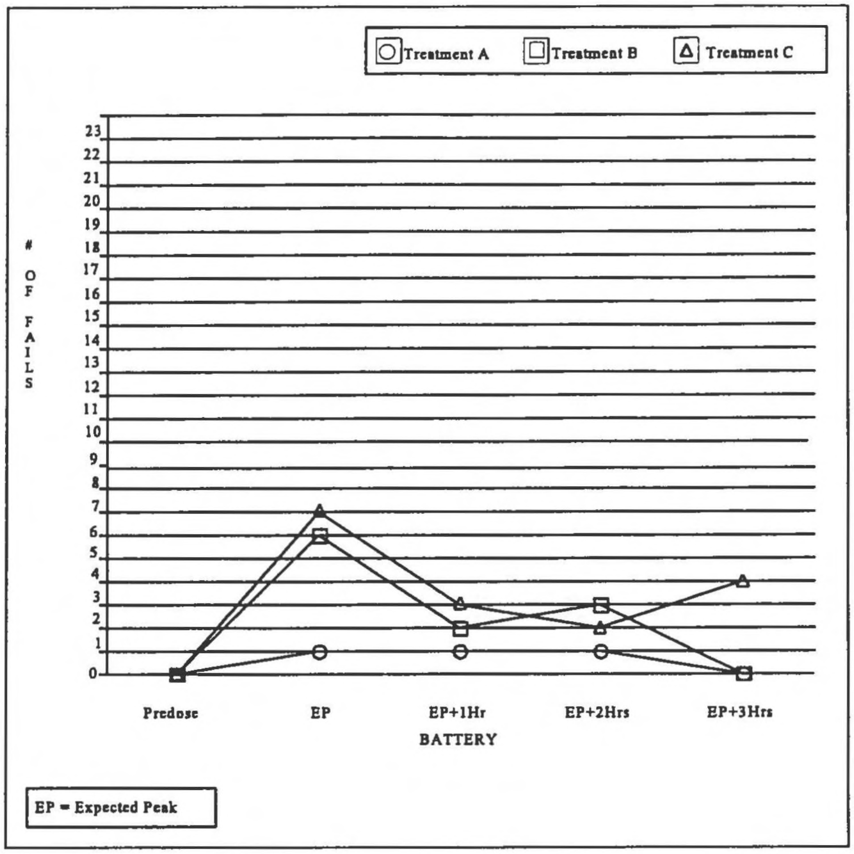


Figure 6-8. Number of Fails for FFD Test 5

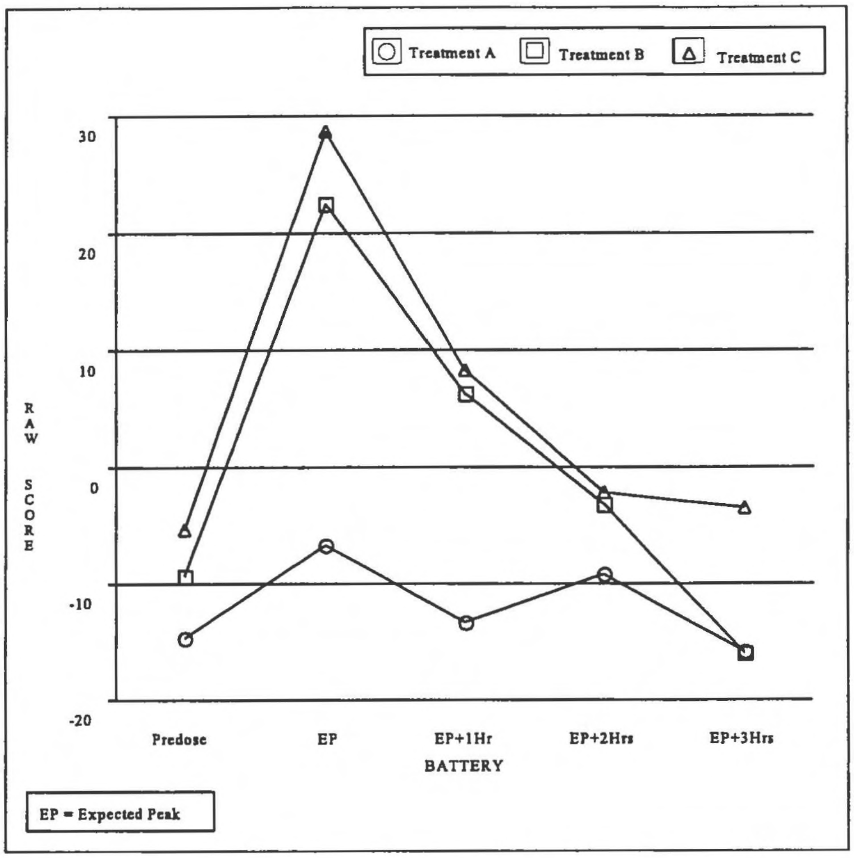


Figure 6-9. Raw Scores for FFD Test 5

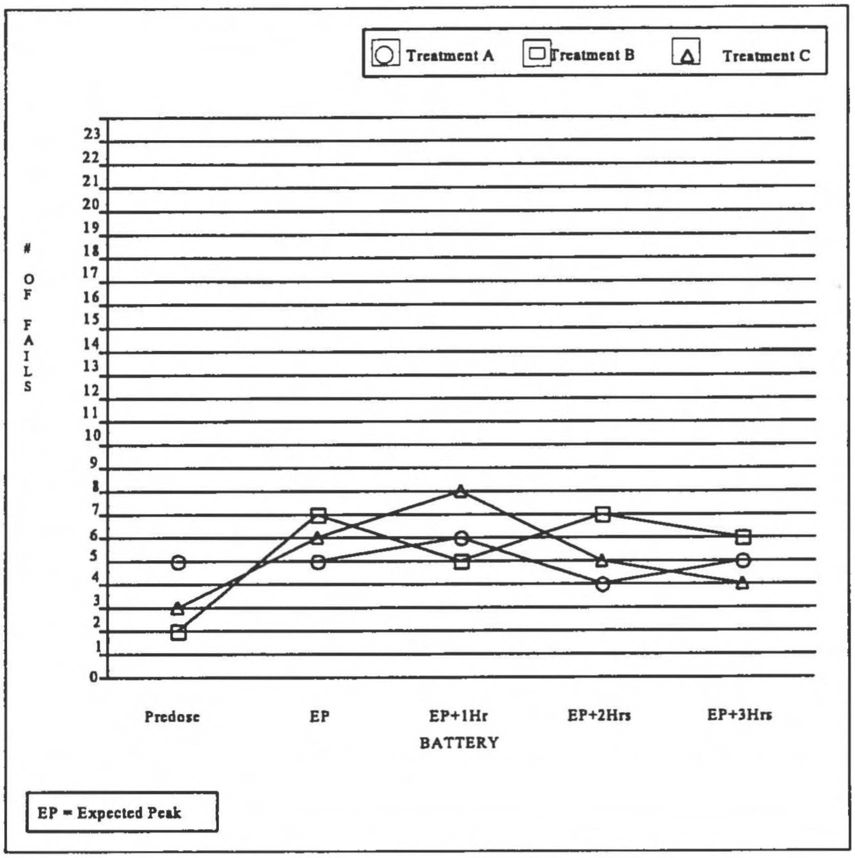


Figure 6-10. Number of Fails for FFD Test 6

Fail Percentage at Various BACs

Because subjects were tested at time intervals rather than at specific BACs, some ordering of batteries was necessary to compute the Fail Percentage of each test at 0.08%, 0.06%, 0.04%, ≤ 0.02 BACs. In arranging the batteries by BAC, two concerns were addressed. First, the breath analyzing instrument error range of ±0.01 was considered. To compensate for this possible measurement error, batteries were grouped by BAC ± 0.01 (e.g., a battery at 0.078% BAC and another at 0.089% BAC were grouped into the same 0.08% BAC group). Second, the average duration of a test battery was 13 minutes, and the average BAC declined 0.007% during that time. Because test order was randomized, no test was administered at a systematically lower BAC. However, to adjust for the subject's individual metabolism rate, the pre- and post-battery BACs were averaged into a single measure. Fail Percentages by BAC are reported in Figure 6-11.

Consistency Index

The Consistency Index, a measure of test/retest reliability across the two treatments, was calculated by dividing the number of identical test results between Treatments B and C by the number of total comparisons (24 subjects x 5 batteries = 120 comparisons). It addresses the question "if a test failed a subject at a given BAC at one alcohol session, did it also fail that subject at the other alcohol session?" Although BACs were not exactly identical across treatments, they were sufficiently similar for this measure to be useful. An index of 0.00 indicates no reliability, while an index of 1.00 indicates perfect reliability. The reliability data displayed in Figure 6-12 should be considered in conjunction with previously reported measures. To be useful, a measure must first be valid, i.e., it must measure what it claims to measure *and* it must do it reliably, i.e., consistently. Note that it is possible for a test to be a reliably poor measure. It may consistently fail individuals who are not impaired or consistently pass individuals who are impaired.

Signal Detection Analysis

The number of Hits, False Negatives, False Positives, and Correct Rejections were calculated at Low (0.001% BAC or above), Medium (0.04% BAC or above), and High (0.08% BAC or above) impairment criteria. In the interest of brevity, only calculations pertaining to FFD Test 1, the most promising test, are reported.

Sixteen tests with FFD 1 were invalid. If a BAC of 0.08% is taken as the impairment criterion (high threshold), 252 of the 344 valid tests are correct; i.e., the test failed subjects at BACs ≥ 0.08% and passed subjects with BACs with less than .08% BACs. By this criterion, the test yielded 92 incorrect results. In ten cases, the test passed subjects whose measured BAC were 0.08% or higher. In 82 cases, the test failed subjects whose BACs were below 0.08%. These results appear in the following matrix.

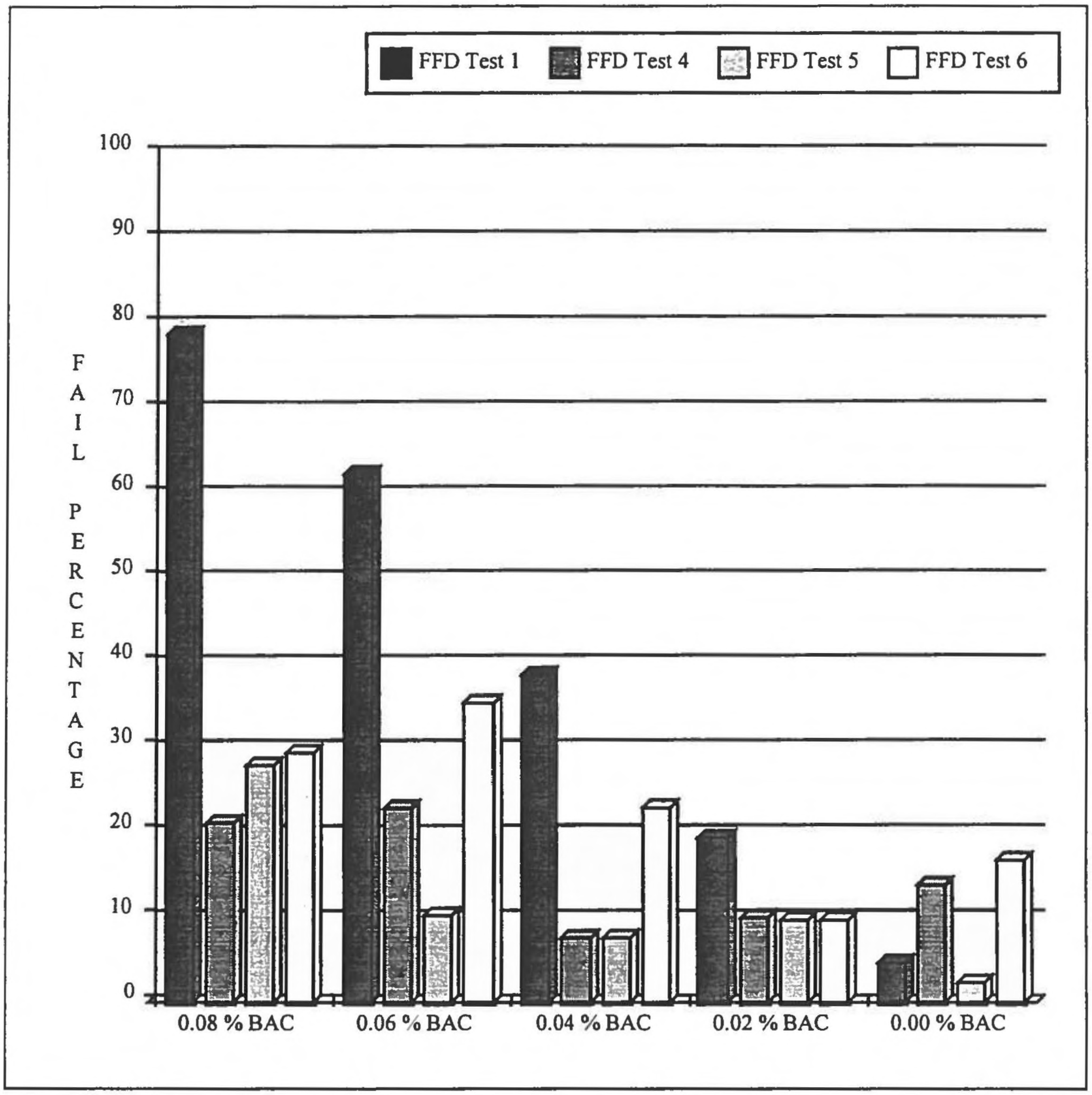


Figure 6-11. Fail Percentages at Various BACs for All FFD Tests

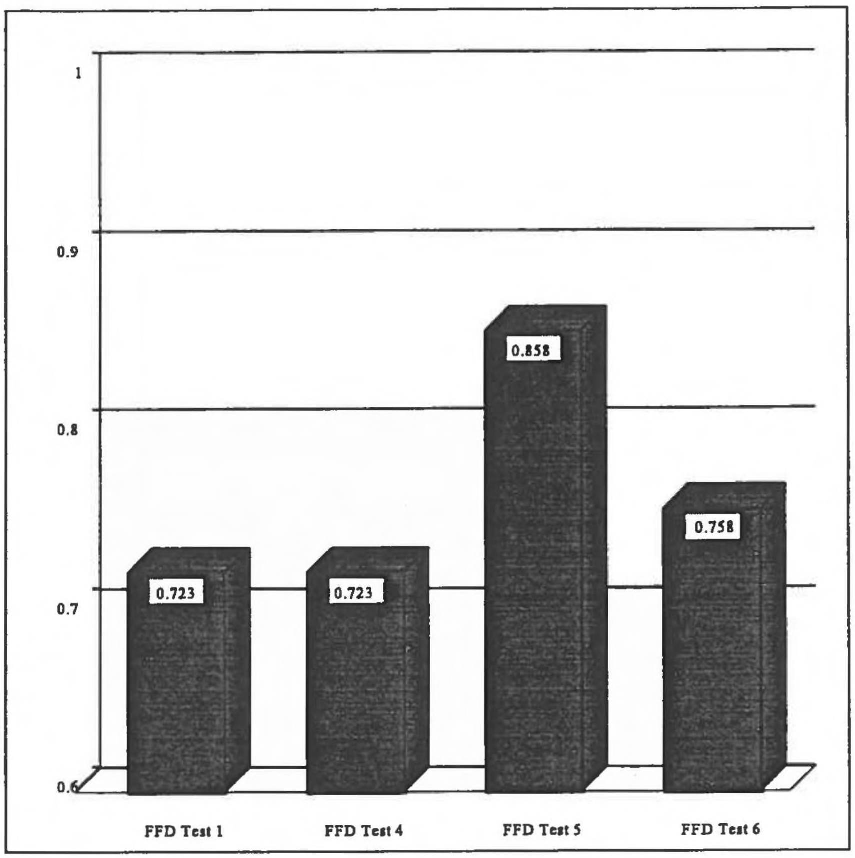


Figure 6-12. Consistency Indexes

	Hits	False Negatives	
BAC 0.08	48	10	58
	False Positives	Correct Rejection	
BAC < 0.08	82	204	286
	130	214	344
	1		

If impairment is defined as a BAC ≥ 0.04 (medium threshold), 276 test results were correct and 68 were incorrect. In 54 cases, the test passed subjects whose measured BAC were 0.04% of higher. In 14 cases, the test failed subjects whose BAC was below 0.04%.

	Hits	False Negatives	
BAC 0.04	98	54	152
	False Positives	Correct Rejection	
BAC < 0.04	14	178	192
	112	232	344

If impairment is defined as any amount of measurable alcohol (i.e., low threshold = 0.001 BAC or above), of the 344 tests results 258 were correct and 86 were incorrect. In 78 cases, the test passed subjects whose measured BAC were 0.001% or higher. In 8 cases, the test failed subjects whose BAC was 0.00%.

	Hits	False Negatives	
BAC 0.001	104	78	182
	False Positives	Correct Rejection	
BAC < 0.001	8	154	162
	112	232	344

These findings, specifically the findings that a lower BAC criterion decreases the proportion of False Negatives and increases the proportion of False Positives, strongly suggest that the test's False Positive errors at 0.08% are a function of the criterion rather than a misclassification by the test. That is, the test detected real changes from baseline at BACs below 0.08%.

Subjective Data. Subjects rated their levels of intoxication and impairment on a 0 to 100 mm VAS. Consistent with the alcohol data, subjects' mean rating of intoxication and impairment was lower in Treatment C than Treatment B (Figures 6-13 and 6-14). Subjects also rated test difficulty and their interest in each test on a 1 to 10 scale. Figures 6-15 and 6-16 report the average ratings. In general, females found physiological tests more difficult, and performance test less difficult, than their male counterparts. FFD Test 1 was rated as the most difficult overall by both males and females. Both males and females rated the physiological tests as less interesting than the performance tests.

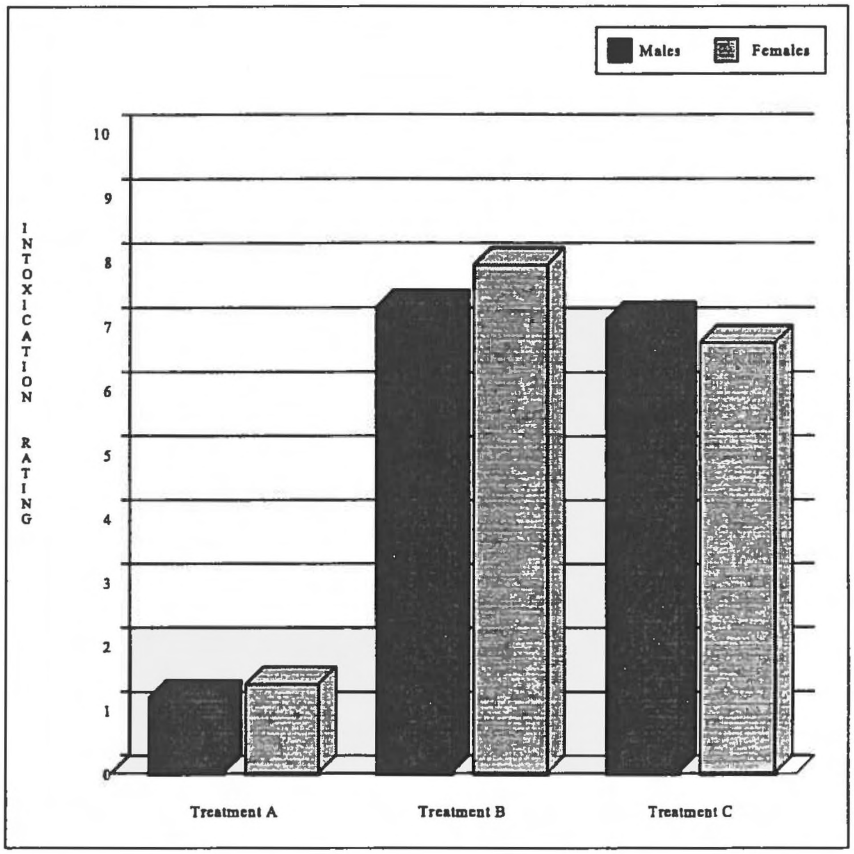


Figure 6-13. Subjects' Rating of Intoxication (Visual Analogue Scale)

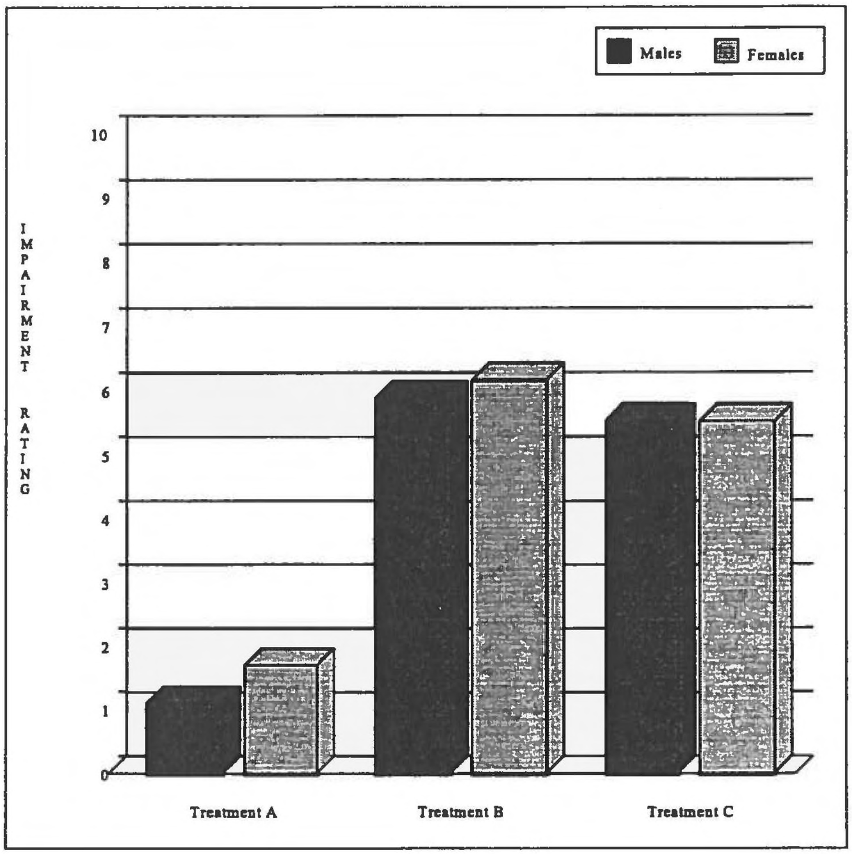


Figure 6-14. Subjects' Rating of Impairment (Visual Analogue Scale)

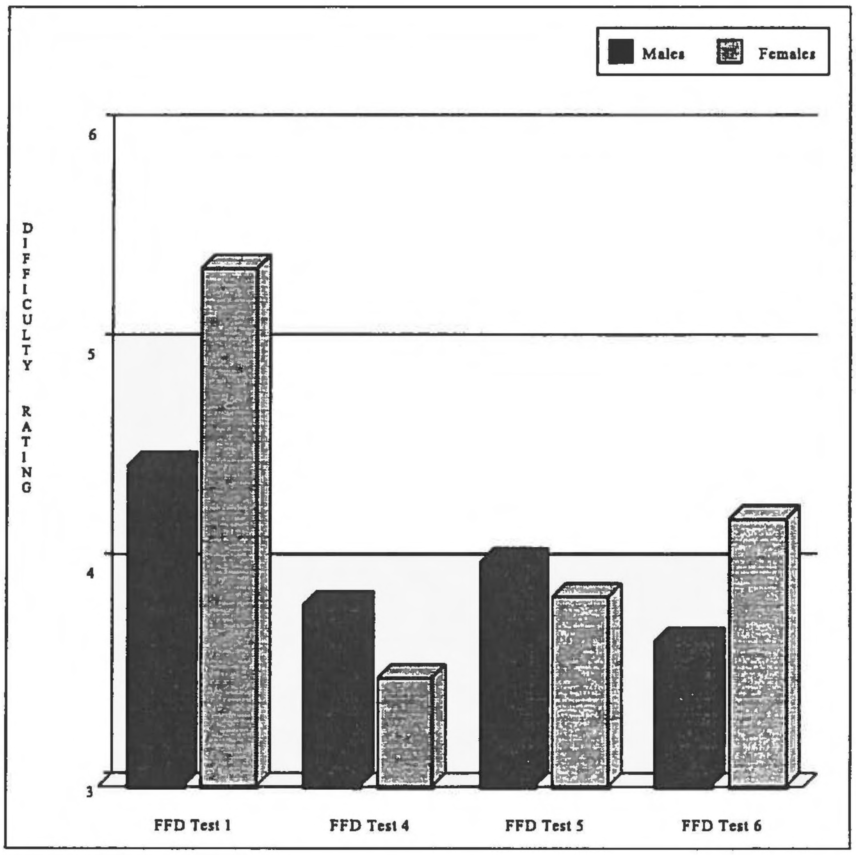


Figure 6-15. Difficulty Rating
(1 = Not at All Difficult, 10 = Very Difficult)

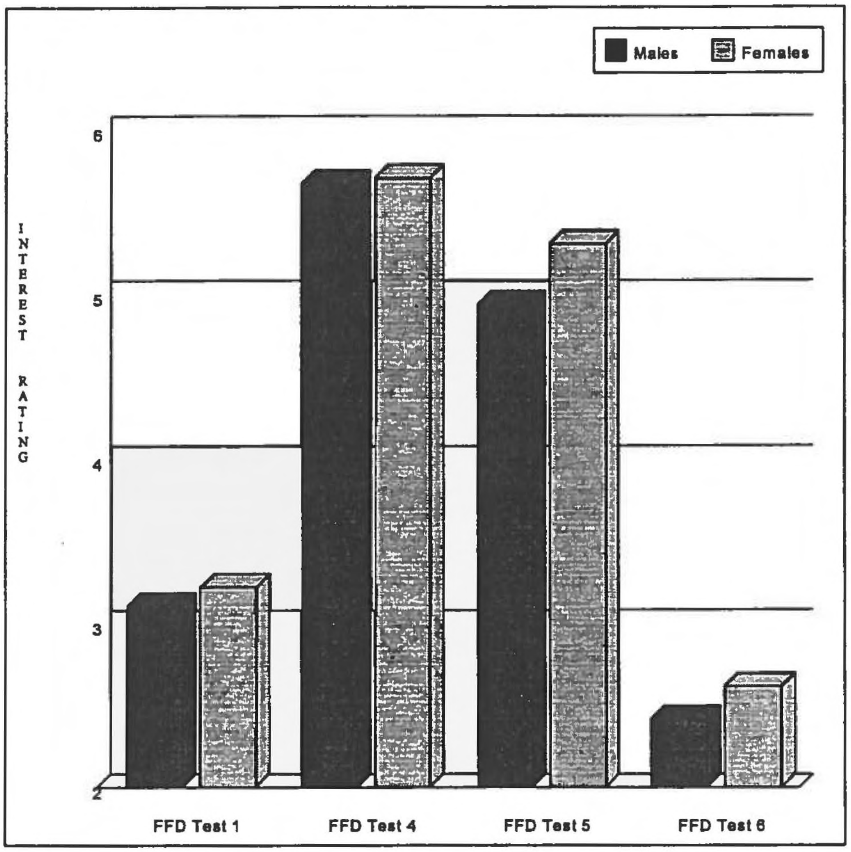


Figure 6-16. Interest Rating
(1 = Not at All Interesting, 10 = Very Interesting)

Summary of Results by FFD Test

FFD Test 1. As in the pilot study, FFD Test 1 yielded the most encouraging results. It failed the most subjects at Peak BACs, and linearly decreased the number of Fails as a function of BAC. It detected impairment in 79% of subjects at 0.08% BAC, 62.5% of subjects at 0.06% BAC, 38.46% of subjects at 0.04% BAC, 19.35% of subjects at 0.02% BAC. In addition, less than 5% of subjects were failed at 0.00% BAC, a relatively low False Positive rate. The reliability index was 0.73, indicating that a subject with a closely similar BAC will have the same Pass/Fail result 73% of the time.

Subject's ratings, however, indicated it was viewed as the most difficult test and the second least interesting, suggesting that FFD Test 1 may require careful introduction into the workplace if it is going to be accepted by employees.

FFD Test 4. Problems with FFD Test 4 had been noted in the pilot study: A counterintuitive, "perform as usual" approach to testing, a lengthy baselining period, and an automated data processing function that requires the system to be turned-on at all times. These requirements and conditions were addressed and governed procedures in the laboratory study.

Nonetheless, the results still indicate that FFD Test 4 fails subjects as a function of other than BAC alone. It failed only 21.31% of subjects at 0.08% BAC, 22.92% of subjects at 0.06% BAC, 7.69% of subjects at 0.04% BAC, 10% of subjects at 0.02% BAC. In addition, 13.69% of subjects were failed at 0.00% BAC. The reliability index was 0.723.

FFD Test 5. In the pilot study, this test was not administered correctly and no valid data were collected. In the laboratory study, steps were taken to insure that each test was administered in accordance with the vendor's specification, and no problems were encountered. It failed 27.87% of subjects at 0.08% BAC, 10.42% of subjects at 0.06% BAC, 7.69% of subjects at 0.04% BAC, 9.68% of subjects at 0.02% BAC. 2.40% of subjects were failed at 0.00% BAC.

When analyses were performed with the raw data rather than the pass/fail result, however, a greater sensitivity to BAC was found, indicating that the vendor's scoring algorithms might have been too lenient. The vendor appears to prefer to err on the side of avoiding False Positives, resulting in a bias toward False Negatives in their scoring of raw data. Note, however, that the scoring algorithm can be changed and that FFD Test 5 had the highest reliability index among the four tests, 0.858.

FFD Test 6. This test which was not included in the pilot, failed 29.51% of subjects at 0.08% BAC, 35.42% of subjects at 0.06% BAC, 23.08% of subjects at 0.04% BAC, 9.68% of subjects at 0.02% BAC. Surprisingly, 16.67% of subjects were failed at 0.00% BAC; thus differences between alcohol and no-alcohol sessions are not statistically significant. The reliability index was 0.758. The results suggest that further refinement is needed in this fitness-for-duty test.

CHAPTER SEVEN

INTERPRETATIONS, APPRAISALS, AND APPLICATIONS

Fitness-For-Duty Defined

When this project was proposed, it was expected that fitness-for-duty tests would be entirely performance-based. Some vendors, however, had developed tests based on measurement of physiological processes. For this reason, the definition of fitness-for-duty was expanded to encompass all technologies.

The Information Processing Model

Under this model, the processing of sensory information (perception, attention, memory, decision making, response selection and execution) is viewed as central to performance. Degraded central processes are associated with a deficit in performance. If a fitness-for-duty test measures central processes, it is expected to be a predictor of job performance. Although tests of physiological processes do not directly measure performance, the obtained measures are represented as reliably reflecting the presence of substances known to act on central processes.

The demonstration of a direct link between fitness-for-duty tests and the performance of the tasks required by all transit jobs by would be a difficult undertaking, as it would require subjecting transit employees to stressors and then comparing their test results to on-the-job performance. The link can be inferred by association if the link between 1) cognitive abilities and performance, 2) impairment and cognitive abilities, and 3) test results and impairment could be established. The link between impairment and degradation of cognitive abilities and processes is already well understood for alcohol.

The challenge for this study was to establish the link between cognitive abilities and on–the–job performance and the link between fitness-for-duty test results and impairment. The first link was established though interviews with transit employees to determine cognitive abilities and processes required to perform their jobs safely. The second link was established by dosing subjects with alcohol and then determining whether the fitness-for-duty tests could detect the resulting impairment. The association by inference was weakened because none of the proposed fitness-for-duty tests had been specifically designed for transit employees and it was not known whether the tests measured the cognitive abilities and processes of interest to the transit industry. This weakness was foreseen going into the project and was considered unavoidable. Fortunately, the cognitive abilities and processes found to be important in transit appeared to be similar to those that are likely to be important in other occupations for which the fitness-for-duty tests had been designed. Since the link between impairment and cognitive abilities and processes is the link least in question, and in fact, is almost established by definition, it was assumed in this study to be a strong link. Consequently, it was necessary only to determine whether fitness-for-duty tests detected the impairment of the dosed subjects.

Relationship of Fitness-For-Duty Tests and Job Skills

Twenty of the processes, abilities, and skills that were identified by transit personnel as important to safety were found to be common to all or nearly all of the seven categories of safety–sensitive transit jobs that were studied. These abilities, processes, and skills are candidates for fitness-for-duty testing:

- Memory
- Attention to detail
- Reaction time
- Reading comprehension
- Analytical ability
- Understand written communication
- Read instruments
- Communicate orally
- Willingness to instruct others
- Direct others
- Manual dexterity
- Eye-hand coordination
- Color vision
- Visual acuity
- Auditory acuity
- Peripheral vision
- Depth perception
- Willingness to follow procedures
- Willingness to work with others
- Ability to respond under stress

The job analysis which produced the above list was intended to provide a benchmark against which fitness-for-duty tests developed specifically for the transit industry could be evaluated. No such tests were located, but the list should be of interest to vendors who do undertake the development of industry–specific tests.

Fitness-For-Duty Test Research

Most of the processes, abilities, and skills found to be common to all or nearly all safety–sensitive transit jobs have been examined in research studies. Although only limited research data are available for commercially available tests, the concept of testing to examine performance is well established. The catalog of research and vendor fitness-for-duty tests that was assembled during this study should be of interest to transit agencies, researchers, and test developers.

Laboratory Findings for Fitness-For-Duty Tests

A total of six fitness-for-duty tests were examined, five in the pilot (Tests 1,2,3,4, and 5) and four in the study (Tests 1, 4, 5, and 6). Based on the results of both the pilot and the laboratory study, the following conclusions can be drawn:

1) There was great variability among the tests in detecting impairment. Some tests detected impairment as defined by the BAC criterion with a greater degree of sensitivity and reliability than others. The difference between tests may be attributable to theoretical underpinning, faulty measurement, or inaccurate scoring of the raw data.

   One system allowed subjects to re–try after test failure. Examination of the data revealed that subjects at a potentially impairing BAC, who initially failed the test, typically passed on re–trying. It appears that the initial failures accurately reflected impairment, but practice or intensified effort on a second set of trials enabled them then to pass.

   One system utilized a pass/fail criterion that appears to be too lenient; several subjects passed although their raw scores indicated significantly lower levels of performance compared to their baseline. Another system administered three out of eight possible tests. The difference in demands imposed on the subjects by the various possible test combinations may account for the obtained results.

2) The best performing test failed 79% of subjects at a 0.08% BAC, 62.5% of subjects at 0.06% BAC, 38.5% of subjects at 0.04% BAC, and 19.35% of subjects at 0.02% BAC, while failing 4.76% of subjects at 0.00% BAC. The reliability index (0.73) indicates that almost three-fourths of the subjects received the same Pass/Fail result at both alcohol sessions. Whether these results are acceptable will ultimately be determined by the transit industry itself.

3) In general, physiological tests yielded better results than performance tests. Subjects, however, rated physiological tests as more difficult and less interesting than performance tests. The best performing test, for example, was rated as the most

difficult and as the second least interesting. A few subjects commented on how uncomfortable taking this test was, raising potential employee acceptance issues.

An important distinction must be made at this point between "difficulty" and "discomfort." The physiological tests used in the study measured involuntary reflexes by requiring a subject to follow a light. By definition, such a task is not difficult, although it may be uncomfortable. It is possible that, in their rating of task difficulty, subjects misinterpreted the meaning of the question or, perhaps, generalized the negative connotation of "difficult" into their response.

In the final analysis, the feasibility of fitness-for-duty technology will depend on transit industry definitions and requirements. If a single test meets these requirements, the technology can be deemed feasible. Thus, the transit industry's definition of impairment is crucial. Based on the data obtained during the study, it appears that fitness-for-duty testing for the transit workplace may indeed be feasible.

Interpretation of Results from the Operational Perspective

Expectations for fitness-for-duty testing varied widely among transit agencies. Unlike other transportation industries, such as maritime, aviation, and railroad, few industry-mandated regulations cover operational practices and policies. Transit agencies, depending on local conditions, modes operated, labor contracts, managerial methods, and other factors, utilize various practices in conducting operations. Key operational factors and procedures are not uniform throughout the industry. As a result, some agencies were concerned with the cost of the proposed tests, while others were more concerned about the time required to test the employees.

Generally speaking, transit agencies would prefer fitness-for-duty tests that

- Require a test time per employee of 2 minutes or less
- Require little or no training of employees or administrators
- Use a baseline to avoid potential manipulation
- Test for involuntary responses to avoid potential manipulation
- Possess high "face validity" to encourage acceptance by employees
- Are inexpensive.

Clearly, these are ambitious and somewhat conflicting criteria. A fitness-for-duty test should not need to satisfy all of these criteria to be considered feasible from an operational perspective. Transit agencies interviewed during this study understand the necessary tradeoffs to achieve a practical fitness-for-duty test.

The transit agencies also understand that many issues regarding fitness-for-duty testing will have to be resolved by the agencies themselves, rather than by the vendors of the tests. All vendors contacted during this study purposely avoided getting involved in what they considered "local" decisions: who to test, when to test, what to do with an employee who fails a test, etc. The vendors focussed instead on satisfying tangible criteria such as accuracy, time to test, and cost.

The five examined systems vary widely in practical aspects, which may be important determinants of their potential use in a specific work environment. The dollar costs associated with the systems (equipment purchase, license agreement, supplies, maintenance, expected life) were not evaluated. Large differences were noted, however, in system hardware requirements, operations manuals, vendor support, baseline data requirements, and overall ease of use. For example, manuals range from lengthy, hard-to-use documents to concise, usable instructions. For some equipment, no manual was provided, but in-person and telephone support was readily available.

Transit agencies identified these and other issues that will need to be resolved for fitness-for-duty testing to be feasible from an operational perspective. Some of these issues must be addressed by vendors of fitness-for-duty tests, but most will require resolution by the transit agencies themselves. The most frequently cited issues are described below.

Frequency of Testing

Frequency of testing is expected to affect the effectiveness and cost of fitness-for-duty testing. While testing every employee prior to safety-sensitive duty is the option most often requested by transit agencies, this frequency may not be practical given the operational and administrative costs of conducting fitness-for-duty testing. One alternative could be to randomly select employees for testing, similar to the DOT's random drug and alcohol testing requirements.

Human Resources Procedures

Transit agencies expect that fitness-for-duty testing would require changes in their human resources policies and procedures, including:

- Attendance policies. Would employees be allocated additional sick or leave days if they are found unfit for duty? Would employees receive this leave on a paid or unpaid basis?
- Sick policies. Would employees be treated and categorized as sick if they were found unfit for duty?

- Workers' Compensation policies. (This may also impact or affect state Workers' Compensation laws.) Would employees who are found unfit for duty have Workers' Compensation claims if the result is due to workplace related stress?
- Vacation leave policies. Would employees be allowed or required to use vacation time if they are found to be unfit for duty?

All these issues need to be addressed by human resources professionals prior to instituting fitness-for-duty testing. These are clearly issues that need to be addressed by policy and operating practice modifications.

Collective Bargaining Agreements

Many issues that need to be addressed prior to implementing fitness-for-duty testing will impact collective bargaining agreements between most transit systems and their labor unions. A number of items in the agreements, including disciplinary practices and procedures, attendance policies, and work assignments may be impacted by fitness-for-duty testing. Changes during the collective bargaining process would be required to address this issue.

Employer Response to "Unfit" Employee

One of the issues raised most frequently during the Vendor/Transit Symposium concerned actions an employer might take when an employee is found unfit for duty. The transit agencies interviewed were unanimous in agreeing that "unfit" employees should not be allowed to work, but some transit agencies would prefer to allow them to withdraw from duty without further action. Other agencies would prefer to subject an "unfit" employee to follow-on tests or disciplinary actions. The employer response is further complicated because the fitness-for-duty test may not be specific to a stressor, or the transit agency may prefer not to know the stressor. One further complication results from the irony that an "unfit" superior employee may still be capable of higher performance than a "fit" inferior employee, since the fitness-for-duty tests are based on individual baselines.

Personnel Requirements

Regardless of the vendor test used, it is likely that an increased number of employees will be found unfit for duty in comparison to the present system, which relies, in large part, on supervisor observations. This will require that additional employees be available to perform safety-sensitive jobs. The problem is exacerbated by the short notice that would be required to call in additional employees.

Employee Reporting for Work Requirements

Transit agencies provide a specified amount of time at the start of a shift for employees to prepare for work, including clocking in and dressing. Additional time will need to be allocated for fitness-for-duty testing if testing is to be performed at the beginning of the shift. This will require adjusting reporting times or work schedules and will, in many cases, require modification of the collective bargaining agreement.

Legal Issues

Fitness-for-duty testing will cause impaired employees to lose time, money, and jobs; and some of those employees can be expected to mount a legal challenge. A survey of 1,238 employee assistance professionals (EAPs) found that litigation was considered as one of 10 most important drug program issues. One in 12 EAPs had been sued regarding some aspect of drug testing (Backer, 1989). Legal actions taken in alcohol or drug impaired driving cases give insights to how fitness-for-duty testing may be challenged. Criminal defense and civil suits often argue the following:

- The test or instrument (breath test equipment, field sobriety tests, drug recognition methodology) has not been shown to be scientifically valid and reliable.
- Records do not substantiate the accuracy (calibration, maintenance) of measurement instruments.
- Proper procedures were not followed in obtaining an evidential test.
- The individual was denied due process (implied consent, Miranda).
- Testing personnel were not qualified.

Similar challenges can be expected to fitness-for-duty testing. In addition, the following legal challenges may be made:

- Do the tests divulge non-job related information about the employee? If the answer to this question is yes, then the transit authority will need to develop some policy for monitoring this information and insuring that it is not misused.
- Will the use of the testing methodologies result in a disparate impact among the employee groups being tested? If the product for some reason tends to identify a higher positive result (unfit employees) for members of statutorily protected groups, this could result in legal complications for the testing agency.

Some states and/or municipalities limit the type of testing that employers can perform on employees, as well as the follow up actions that can be taken as a result of obtaining this information. This is analogous to the situation in which some states that prohibit random substance abuse testing of employees required Federal preemption under the DOT drug and alcohol rules. Federal preemption may not be available for fitness-for-duty testing.

Financial Impact of Testing

While all the transit agencies interviewed believe enhanced safety is important, there was significant concern about the overall cost of fitness-for-duty testing. While a fitness-for-duty testing program could share some resources (such as a substance abuse staff) with other ongoing programs at transit agencies, many elements would represent new costs to the agencies.

Integrity of Fitness-For-Duty Testing Equipment

Several transit agencies, especially those with multiple reporting locations, were concerned about the durability and security of fitness-for-duty testing equipment. This concern was exacerbated for those products that lend themselves to vandalism or abusive treatment.

Acceptability to Workforce/Unions

Acceptability to employees and their unions has several elements. First, the fitness-for-duty test would have to be scientifically valid, permitted by regulations, accommodated by applicable collective bargaining agreements and respect matters of interest to the employees, such as privacy and dignity. Second, because of the consequences of failing to pass the tests, employees would need to see the relationship between the test and the job, suggesting that tests with high "face validity" would be more likely to be accepted by employees.

Logistics

If employees report to multiple work sites and if expensive apparatus and trained testing personnel are required and are available only at one or a few sites, problems of transporting employees and timeliness of testing may interfere with test implementation. The timing of testing often is crucial to valid results, and it may be inappropriate if the employee's work location is distant from the test site. This consideration cannot be lightly dismissed, as illustrated by the difficulties in litigating impaired driving cases for which a breath test machine or blood drawing capability was not immediately available to the suspect. In the same vein, a requirement for special or highly controlled conditions for test equipment (lighting, temperature, noise, privacy) may rule out otherwise acceptable tests.

Costs

Although this study asked vendors to project the cost to acquire their testing equipment, it did not ask for data on "operating" costs or the total "capital" cost (cost per unit times number of units required). Transit agencies will need to consider a variety of costs:

- Employee wages while training, establishing a baseline, taking the test, and doing whatever follow up is required for failure to pass the test
- Administrator wages for scheduling and supervising tests, assessing test results, training employees, and maintaining records
- Professional services (medical officer, EAP representative) required to assess test results, conduct evaluations, and plan follow up actions
- Equipment maintenance to maintain and repair the equipment, maintain service logs, provide supplies, and calibrate and certify (if required) the equipment.

Benefits

Costs are balanced by benefits, of course; and many of the potential benefits of fitness-for-duty testing extend beyond reducing accidents. Drug and alcohol abusers are absent from work more, use more sick leave, use more medical benefits, and have more difficulty getting along with their co-workers. Co-workers, subordinates, and supervisors of chemically dependent employees are also likely to be less productive on the job. Medical benefit usage by dependents of alcohol and drug abusers is higher than other workers' dependents (Jones and Vischi, 1979; Holder, 1987). Transit agencies considering using fitness-for-duty tests will need to balance the benefits of those tests against the costs of implementing them.

Interaction of Fitness-For-Duty Testing with Other Forms of Testing

Fitness-for-duty tests have been suggested as a substitute for, rather than as a complement to, several other types of tests, as described below.

DOT Mandated Biochemical Drug and Alcohol Tests. Transit agencies are concerned with the costs, and other logistics, of biochemical testing for drug and alcohol consumption and have inquired about using fitness-for-duty testing in place of, or as a complement to, biochemical testing. DOT rules specifically require biochemical testing, however, and do not allow fitness-for-duty testing to be used as a substitute. The DOT rules require transit agencies to administer "reasonable suspicion" tests under certain circumstances. Presumably, an employee found to be impaired on a fitness-for-duty test would be subject to "reasonable suspicion" testing unless a stressor other than alcohol or drugs could be identified.

Pre-Employment Testing. Several transit agencies inquired whether fitness-for-duty tests could be used to select among applicants to fill a position. According to the definition of fitness-for-duty testing used in this study, fitness-for-tests are to be used to detect *changes* in an employee's ability to perform his job, not to assess his performance against other employees. The presumption is that a capable employee has been hired and that the fitness-for-duty test is to determine whether the employee reporting for duty on a particular day is the "same" employee that was hired.

Specific Stressor Detection. During the Vendor/Transit Symposium, several transit agencies expressed an interest in detecting low levels of alcohol consumption (at approximately the .02 BAC level). While several vendors are designing their tests to detect the impairment that results at this level, one vendor urged transit agencies that are interested in only one specific stressor to consider a stressor-specific test rather than a multi-stressor fitness-for-duty test. In the case of low-level alcohol detection, that test might be a passive alcohol sensor. A passive alcohol sensor samples ambient air around an employee and provides a yes/no or pass/fail decision based on a threshold selected by the administrator. These instruments are certified for preliminary assessments by law enforcement officers. Of course, a passive alcohol sensor (or similar device for alcohol or other stressor of great interest) could be combined with a fitness-for-duty test to provide greater protection against false negatives or to compensate for a deficiency in the fitness-for-duty test. The latter application may be crucial, as none of the fitness-for-duty tests examined could reliably detect alcohol consumption or impairment at levels as low as this example.

CHAPTER EIGHT

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

Overall

Fitness-for-duty testing is not yet feasible in the transit industry, based on this evaluation of scientific and operational perspectives. But, based on the same evaluation, the research team is optimistic that many of the remaining gaps between transit industry needs and the ability of vendors of fitness-for-duty tests to satisfy those needs can be narrowed, if not closed, in the near future. No inherent obstacles to fitness-for-duty testing in transit were identified.

Many individual elements comprise these general conclusions. These elements are categorized below.

Specific

The development of fitness-for-duty tests for transit is hampered by a lack of agreement on definition and expectations.

There is a wide variety of expectations among transit agencies and other transit professionals as to what constitutes fitness-for-duty testing and what fitness-for-duty testing can accomplish, ranging from detecting all impairments to detecting threshold levels of a particular stressor. This lack of agreement discourages vendors from designing tests for a transit market. This lack of agreement is not unique to transit and is typical of newly emerging technologies.

A variety of fitness-for-duty tests are in development and a selection of fitness-for-duty tests will be available for transit.

Over a dozen vendors contacted during this study expressed an interest in developing fitness-for-duty tests for transit. Vendors invested significant time and resources into assisting with this study. The fitness-for-duty tests identified during this study suggest that a variety of tests will be available for transit.

Current fitness-for-duty tests have not yet been demonstrated to be sufficiently sensitive to be used in transit.

At best, fitness-for-duty tests detected about 79% of impaired subjects at a .08 blood alcohol concentration (BAC). The ability of fitness-for-duty tests to detect impairment drops with lower BACs. The nature of transit operations suggests that fitness-for-duty tests will need to detect impairment at the equivalent of .02 to .04 BAC in 95% to 100% of impaired employees with very few false positives (i.e., non-impaired employees testing as impaired). Although a significant gap appears to remain between fitness-for-duty test capabilities and transit needs, two factors should first be addressed:

- The actual impact of low alcohol levels on an individual's physiological and performance-based functioning

- The influence of the test criteria used to define Pass and Failure on detection rate.

When assessing the effectiveness of a fitness-for-duty test in achieving the transit community's desire for sensitivity to very low BACs, it is important to note that these amounts may not, in every individual, have a detrimental effect on that individual. Evaluation of a test's sensitivity is based on the assumption that some level of impairment always occurs in response to alcohol present in the amounts used in this study. Differences in an individual's sensitivity could mean that those individuals classified as "false negatives" may, in fact, have not actually been impaired. This is an important issue that must be resolved in order to ensure that the transit industry's requirements are realistic.

An additional factor to consider is the influence of the the test criteria used to define Pass and Failure on detection rate. There is an inevitable tradeoff and choice to be made between:

1) A very stringent test criterion that results in the detection of most, if not all, impaired individuals but at the cost of "failing" some unimpaired individuals,

versus

2) A less stringent criterion that rarely yields a false "fail" at the cost of allowing some unknown number of impaired persons to proceed to work.

The selection of an appropriate impairment criterion is a difficult task, which will be unique to the work environment in which fitness-for-duty is to be assessed. In particular, what constitutes an acceptable false positive rate is an important question that must be addressed by the transit community. The best performing test in this study, using the current test criterion, showed a false positive rate of approximately 5%. Whether this level of false positives is acceptable in the transit environment is unknown. Attempts to improve the detection level of a test will probably increase the false positive rate, as well. Consequently, selection of the impairment criterion is a critically important effort in the fitness-for-duty arena. Once it is accomplished, the development of tests utilizing state-of-the-art, computer-based tests can be expected to proceed rapidly.

Given these two factors, it does appear that at least one of the tests evaluated in this study shows great promise for use in the transit environment. However, since only one stressor was used, additional testing will be required to assess sensitivity to other stressors. This testing also should be used as an opportunity to manipulate the impairment criterion to determine its effect on sensitivity and false positive levels. The promising results of this study suggest that this additional testing is warranted.

Transit agency policies, work practices, and collective bargaining agreements will need to be changed to accommodate fitness-for-duty testing.

Fitness-for-duty testing will lengthen the time required for employees to prepare for duty. Removing an impaired employee from duty will require replacement labor on short notice. Follow-up actions against an impaired employee must be negotiated within applicable collective bargaining agreements, particularly with regard to work assignments, discipline, and eligibility for rehabilitation.

Technologies not examined in the clinical testing may also have promise.

Clinical testing in this study was limited by budget and time constraints. Vendor tests not included in the clinical testing clinical testing also show promise.

Whether currently available fitness-for-duty tests will meet transit objectives depends on the impairment criteria set by the agency.

A total of six fitness-for-duty test systems were examined in a pilot and subsequent laboratory experiment using alcohol as the stressor. Approximately 79% of subjects at BACs of 0.08% were identified by one of the tests. If transit operations require the detection of 0.02 - 0.04% BACs with an accuracy near 100% and few false positives, however, that objective can be better achieved with a PBT, either alone or in conjunction with a multi-stressor fitness-for-duty test. Unresolved cost, training time, testing time, and legal issues suggest that further work must be done to determine if current fitness-for-duty tests would be workable from an operational perspective.

Achieving feasibility will require efforts by vendors and transit industry.

Vendors need to increase the sensitivity of their tests, shorten training and testing times, and reduce costs. The transit industry needs to formulate a set of common expectations for fitness-for-duty testing to provide design criteria for vendors.

Recommendations

Scientific evaluation of fitness-for-duty tests should be expanded.

The clinical tests in this study were limited to a single stressor within a limited age group. Future clinical tests should address:

- The ability of fitness-for-duty tests to detect impairment from other stressors of interest to transit, such as drugs and fatigue
- The tradeoffs between false negatives and false positives as vendors adjust the failure thresholds within their tests.

A limited demonstration project might be feasible at this time as a means of encouraging the development of necessary policies.

The encouraging results obtained with FFD Test 1 suggest that a limited demonstration project might be useful as a means of working through the various policy issues that must be resolved in order to introduce fitness-for-duty testing into the transit environment. Such a demonstration must be initiated with a clear understanding of the goals of such a demonstration. As the first recommendation stated, the sensitivity of these tests to other stressors is not known. In addition, such a demonstration project carries some risk in that it could create operational problems for participating transit agencies and encourage resistance among employees if care is not taken to constrain the impact of the demonstration on all participants, especially those negatively impacted by the testing. In spite of these reservations, a demonstration project does offer a potentially useful means for addressing logistical and policy issues that must be resolved if fitness-for-duty testing to be implemented.

Vendors should be encouraged to develop fitness-for-duty tests for use in transit.

Transit agencies and organizations should continue to fund research projects addressing fitness-for-duty testing. Design specifications for fitness-for-duty tests in transit should be developed. Care must be taken not to exclude any technology or approach, but functional criteria such as sensitivity, time to train, and time to test can and should be developed. These recommendations could be implemented through several activities:

- An occasional conference, similar to the Vendor/Transit Symposium that was part of this study, to allow vendors, transit agencies, and researchers to assess the current state of fitness-for-duty test development
- A task force to develop design criteria for fitness-for-duty tests, based on the specific needs of the transit industry. This could take the form of an iterative survey where a large selection of transit agencies is polled as to their needs,

those needs are compiled and consolidated, and then the transit agencies are asked to rank and/or rate the needs to develop criteria. This process would be continued until general consensus is reached.

- A survey to poll transit agencies as to their needs. This would include compiling a list of the needs identified by the initial survey and obtaining rankings/ratings of the identified needs from transit professionals. This process could be continued, in conjunction with the conference and task force efforts, until a consensus is reached.

CHAPTER NINE

REFERENCES

Burns, Marcelline and Moskowitz, Herbert (1980) Methods for Estimating Blood Alcohol Concentration.

DOT-HS-8-01999, National Highway Traffic Safety Administration, U.S. Department of Transportation.

Cahalan, Don, Cisin, Ira H., and Crossley, Helen (1969) American Drinking Practices. New Brunswick: Rutgers Center of Alcohol Studies.

Ellis, George M. Jr., Mann, Marian A., Judson, Barbara A., Schramm, N. Ted, and Tashchian, Agop (1985) Excretion patterns of cannabinoid metabolites after last use in a group of chronic users. Clin. Pharmacol. Ther., Vol. 38, No. 5, 572-578.

Leal, Charles (1984) Admissibility of biochemical urinalysis testing results for the purpose of detecting marijuana use. Wake Forest Law Review, Vol. 20, pp. 391-412.

Moskowitz, H. and Robinson, C. (1988) Effects of Low Doses of Alcohol on Driving-related Skills: A Review of the Evidence. DOT HS 807 280, NHTSA, U.S. Dept. of Transportation, Washington, D.C.

Bittner, A., Toquam, J., Forslund, C., & Fleming, T. (1993). TCRP Project Fitness for Duty in the Transit Work Place: Task 2. Seattle, WA: Battelle Human Affairs Research Center.

Burka v. N.Y.C. Transit Authority, 85 Civ. 5751 (RPPO and Fa v. N.Y.C. Transit Authority, 89 Civ. 6536 [RPP], slip op. at 13-14 (S.D.N.Y. June 5, 1990).

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette and L. M. Hough (Eds.). Handbook of industrial and organizational psychology, Volume 2 (2nd ed). Palo Alto, CA: Consulting Psychologists Press, Inc.

"Implementation Guidelines for Anti-Drug Program in Mass Transit," Federal Transit Administration, March, 1989.

_________ (1990) Healthy People 2000, National Health Promotion and Disease Prevention Objectives. Washington: U.S. Dept. of Health and Human Services, Public Health Services, pp. 164-183.

_________ (1991) The Drug Recognition Technician School, Student Manual, 1991 Edition. U.S. Dept. of Transportation, NHTSA.

Axel, H. (1989) Characteristics of firms with drug testing programs. In Gust, S.W. & Walsh, J.M. (eds.) Drugs in the Workplace: Research and Evaluation Data, NIDA Research Monograph 91. Rockville: NIDA. pp.219-226.

Backer, T.E. (1989) Characteristics of firms with drug testing programs. In Gust, S.W. & Walsh, J.M. (eds.) Drugs in the Workplace: Research and Evaluation Data, NIDA Research Monograph 91. Rockville: NIDA. pp.227-244.

Bittner, A.C.,Jr., Carter, R.C., Kennedy, R.S., Harbeson, M.M. & Krause, M. (1986) Performance evaluation tests for environmental research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, pp. 683-708.

Burns, M. (1987) Sobriety tests for the presence of drugs. Alcohol, Drugs and Driving, 3(1), pp. 25-29.

Burns, M. & Moskowitz, H. (1977) Psychophysical Tests for DWI Arrest., DOT-HS-802-424, U.S. Dept. Of Transportation, NHTSA.

Burns, M. & Wilkinson C. (1990) The laboratory study of drug-related performance changes. Journal of Occupational Medicine, 32(4), pp. 320-326.

Campbell, D.T. & Stanley, J.C. (1963) Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally.

Carter, R.C. & Krause, M. (1986) Beware the reliability of slope scores for individuals. Human Factors, 28(6), pp. 673-683.

Cronbach, L.J. (1949) Essentials of Psychological Testing. New York: Harper & Bros.

Damos, D.L., Bittner, A.C.,Jr., Kennedy, R.S., & Harbeson, M.M. (1981) Effects of extended practice on dual-task tracking performance. Human Factors, 23(5), pp. 627-631.

Fleishman, E.A. & Hemple, W.E.,Jr. (1955) The relation between abilities and improvement with practice in a visual discriminatory task. Journal of Experimental Psychology, 49, pp. 301-312.

Green, D.M. & Swets, J.A. (1966) Signal Detection Theory and Psychophysics. New York: Wiley.

Guilford, J.P. & Fruchter, B. (1978) Fundamental Statistics in Psychology and Education. New York: McGraw-Hill.

Goodman, A.G., Goodman, L.S. & Gilman, A. (1980) The Pharmacological Basis of Therapeutics. New York: MacMillan.

Gust, S.W. & Walsh, J.M. (1989) Research on the prevalence, impact, and treatment of drug abuse in the workplace. In Gust, S.W. & Walsh, J.M. (eds) Drugs in the Workplace: Research and Evaluation Data, NIDA Research Monograph 91. Rockville: NIDA. pp. 3-13.

Jex, H.R. (1967) Two applications of a critical-instability task to secondary work load research. IEEE Transactions on Human Factors in Electronics, HFE-8, pp. 279-282.

Jex, H.R., McDonnell, J.D. & Phatak, A.V. (1966) A "critical" tracking task for manual control research. IEEE Transactions on Human Factors in Electronics, HFE-7, pp. 138-145.

Kennedy, R.S., Bittner, A.C., Jr., Harbeson, M.M. & Jones, M.B. (1981) Perspectives in performance evaluation tests for environmental research (PETER): Collected papers. NBDL-80R004. New Orleans: Dept. of the Navy, Naval Biodynamics Laboratory.

Lund, A.K., Preusser, D.F. & Williams, A.F. (1988) Drug use by tractor-trailer drivers. Journal of Forensic Sciences, 33(3), pp. 648-661.

Moskowitz, H. and Robinson C. (1988) Effects of Low Doses of Alcohol on Driving-Related Skills: A Review of the Evidence. Dot-HS 807 280, NHTSA, U.S. Department of Transportation.

Moskowitz, H., Burns, M.M. & Williams, A.F. (1985) Skills performance at low blood alcohol levels. Journal of Studies on Alcohol 46(6).

Norman, D.A. (1964) A comparison of data obtained under different false-alarm rates. Psychological Review, 71, pp. 243-246.

Norman, D.A. (1976) Memory and Attention: An Introduction to Human Information Processing. New York: Wiley.

Richardson, J.T.E. (1972) Non-parametric indexes of sensitivity and response bias. Psychological Bulletin, 78, pp. 429-432.

Smith, P.J. (1981) The use of Sternberg's memory-scanning paradigm in assessing effects of chemical exposure. Human Factors, 23(6), pp. 701-708.

Sutcliffe, J.P. (1980) On the relationship of reliability of statistical power. Psychological Bulletin, 88, pp. 509-515.

Swets, J.A., Tanner, W.P., Jr. & Birsdall, T.G. (1961) Decision processes in perception. Psychological Review, 68, pp. 301-340.

Winer, B.J. (1971) Statistical Principles in Experimental Design. New York: McGraw-Hill.