SELECTION OF LOCOMOTIVE ENGINEER TRAINEES

U.S. Department
of Transportation

**Federal Railroad
Administration**

TEST MANUAL

**Office of Research and
Development
Washington, D.C. 20590**

Merri-Ann Cooper
Charles Uhlman
Lynn Frost

University Research Corporation
7200 Wisconsin Avenue
Bethesda, MD 20814

October 1993

TABLE OF CONTENTS

LIST OF EXHIBITS

I. BACKGROUND

A. PURPOSE AND GOALS OF THE STUDY

The major objective of this study was the development of a battery of tests for predicting an applicant's potential success for the job of locomotive engineer. The locomotive engineer's job is very demanding and requires skills, knowledge, and aptitudes for the effective and reliable handling of the train. The consequences of error in the engineer's job are high, both in loss of life and destruction of equipment. Improper train handling can lead to delays, cargo damage, derailment, and collusions.

The Federal Railroad Administration (FRA) has recognized the importance of improving the procedures used for selecting and promoting locomotive engineers. In addition to safety concerns, improved selection procedures can help increase the productivity and capabilities of workers, reduce the threats of discrimination and subjectivity in hiring and promotions, and facilitate engineer training by effectively prescreening applicants for entry into training programs.

University Research Corporation (URC) was contracted by the FRA to construct a battery of tests for selecting potential locomotive engineers. The psychological, educational, and psychometric literature suggests that cognitive tests are effective predictors of performance in a wide variety of jobs. Previous research also indicates that certain cognitive abilities, such as general reasoning, are important underlying determinants of performance on the locomotive engineer's job. This manual briefly summarizes how cognitive selection test were developed and validated, and describes how to administer and score the tests.

B. JOB ANALYSIS

In order to develop a battery of tests for locomotive engineers a job analysis was conducted. Job analysis is a systematic procedure for determining the behaviors required for effective job performance, and the knowledges, skills, and the abilities (KSAs) that underlie the behaviors. Once the job's critical behaviors and KSAs are identified, selection tests can be developed to measure the key job components.

The job analysis for the locomotive engineer's job involved seven steps: 1) review of existing research on locomotive engineer job requirements, for the purpose of constructing a preliminary list of work behaviors or tasks; 2) a site visit to one of the participating railroads, Union Pacific; 3) review of the task list by subject matter experts; 4) ratings of the tasks by Union Pacific engineers; 5) identification of the required KSAs by Union Pacific engineers; 6) review of the task list by additional participating railroads, (Amtrak, Burlington Northern, Conrail, and Santa Fe) including a site visit to Amtrak; and) identification of required KSAs by the other participating railroads. This thorough and comprehensive job analysis procedure resulted in a list of critical KSAs that were considered appropriate for subsequent test development. These requirements included reading, memorizing, understanding oral instructions, decision making, attention, and conscientiousness.

C. TEST DEVELOPMENT

Test development proceeded directly from the job analysis results and was designed to meet several objectives. The tests were intended to be as practical as possible, to facilitate their use by the railroads. Hence, the tests were developed to be easy to administer, objective, easy to score, require minimal equipment, and amenable to both group and individual administration. To the extent possible, an attempt was made to make the tests face valid, i.e., make them look job relevant.

The following six cognitive ability tests were developed to measure the important KSAs for the selection and promotion of locomotive engineers: Memory, Reading Comprehension, Perception, Listening, Logical Reasoning, and Dichotic Listening. The tests were initially pretested with a sample of engineers and road foremen to identify confusing instructions and items. A second sample of college students was tested to help determine appropriate time limits and to evaluate the technical quality of the tests. Statistical analysis of the pretest results, along with comparison of the test battery to other selection measures, indicated that the six cognitive ability tests were comparable to other tests that have proven valid for selecting applicants across a wide variety of jobs in many different organizations.

D. TEST RELIABILITY AND VALIDITY

Internal consistency analyses on the tests indicated that they had high reliability; coefficient alpha estimates ranged from .77 to .98 (see Exhibit 1). Furthermore, the tests correlated with one another in a meaningful, interpretable pattern (see Exhibit 2). These analyses further substantiated the technical adequacy of the predictor battery.

The next step in the project involved determining if the tests predict engineers' job performance. This step is called validation. Validation is demonstrated by a statistical relationship between tests scores and ratings of job performance. In other words, a test is valid to the extent that tests scores predict job performance ratings.

The method used for validation in this study involved having a sample of engineers (from Amtrak; Burlington Northern; Canadian Pacific; Chicago Northwestern; Conrail; CSX; Duluth, Missabe and Iron Range; Elgin, Joliet and Eastern; Norfolk Southern; Union; and Union Pacific), take the tests and be rated on their job performance. URC project staff administered the tests at each of the companies to small groups of engineers. Later the engineers were rated on their job performance using a specially developed rating form.

The relationship between the scores on the tests and the job performance ratings were statistically examined. No significant relationship was found between them which would be necessary for validating the tests.

We hypothesized that the use of subjective ratings across varying conditions caused these results. A second validation study was conducted using a more objective measure of job performance, performance on a simulator.

Burlington Northern offered to have several classes of engineer trainees take the predictor tests. The trainees were also evaluated on three separate simulator runs, and on two end of training multiple-choice written knowledge tests involving general operating rules and air brakes.

EXHIBIT 1

INTERNAL CONSISTENCY ESTIMATES OF RELIABILITY FOR THE PREDICTOR TESTS

Test	Coefficient Alpha
Dichotic Listening	.98
Logical Reasoning	.88
Memory	.85
Listening	.77
Reading Comprehension	.82

EXHIBIT 2

INTER-TEST CORRELATIONS

Test	Memory	Reading	Perception	Listening	Logical Reasoning	Dichotic Listening
Memory	1.00					
Reading	.44	1.00				
Perception	.23	.05	1.00			
Listening	.31	.50	-.01	1.00		
Logical Reasoning	.49	.70	.14	.57	1.00	
Dichotic Listening	.20	.11	.11	.09	.18	1.00

Statistical analyses were performed on data from 97 engineer trainees to assess the relationship between the six cognitive ability tests, the two multiple-choice training tests, and scores on the three simulator runs. The major results are summarized as follows:

- Scores on the cognitive ability tests were not significantly related to simulator performance.
- Scores on three of the cognitive ability tests - Reading, Logical Reasoning, and Dichotic Listening - were significantly related to performance on the two training exams.
- Performance on the training tests were significantly corrected with simulator performance.

Exhibit 3 displays the correlations between the cognitive ability tests with the training tests and simulator performance. Exhibit 4 displays the correlations between the training tests and the simulator runs.

What implications do these result have for the usefulness of the ability tests? First, three of the tests - Reading, Logical Reasoning, and Dichotic Listening - predict scores on training tests. This finding is consistent with extensive previous research demonstrating that cognitive abilities are important predictors of training success. Because training of locomotive engineers is necessary for safe and efficient train handling, legally mandated, and is time consuming and expensive, it would benefit the railroads to predetermine who is not likely to pass training. The three cognitive ability tests do indeed identify applicants who are most likely to successfully complete training.

EXHIBIT 3

CORRELATIONS BETWEEN COGNITIVE TESTS, SIMULATOR RUNS, AND TRAINING TESTS

	Simulator			Training Tests	
Test	Run 1	Run 2	Run 3	General Rules	Airbrake
Memory	.00	.11	.02	.16	.17
Reading	.12	.04	.04	.30**	.25*
Perception	.04	.02	.03	.09	.09
Listening	.05	.06	-.03	.17	.16
Logical Reasoning	.06	.13	.06	.29**	.28
Dichotic Listening	-.09	.06	-.12	.23*	.00

* $p < .01$
** $p < .001$

EXHIBIT 4

CORRELATIONS AMONG THE SCORES ON THE SIMULATOR RUNS AND TRAINING TESTS

	Simulator			Training Tests	
Test	Run 1	Run 2	Run 3	General Rules	Airbrake
Simulator Run 1	1.00				
Simulator Run 2	.27*	1.00			
Simulator Run 3	.36**	.12	1.00		
General Rules	.03	.41**	.27*	1.00	
Airbrake	.20	.37**	.34**	.68**	1.00

* $p < .01$
** $p < .001$

II. TEST PREPARATION

A. EQUIPMENT AND MATERIALS

The following materials are necessary for administering the Reading, Logical Reasoning, and Dichotic Listening tests:

1) Administration Manual

2) Special Answer Sheet for Dichotic Listening

3) Answer Sheet for Reading and Logical Reasoning (we used an optically scannable answer sheet during the project)

4) Reading Test

5) Logical Reasoning Test

6) Pencils

7) Headsets for Dichotic Listening Test

8) Two (2) AA Batteries for Cassette Players

9) Dichotic Listening Test Tapes

Prior to the testing session, the administrator should thoroughly read the manual and be familiar with all of it's contents. The administrator should ensure that the headsets for the Dichotic Listening test are in proper working condition, and that all Dichotic Listening test tapes and are rewound to their beginning. Be sure to bring extra tests, answer sheets, pencils with erasers, and AA batteries to the testing session.

III. TEST ADMINISTRATION

A. GENERAL INTRODUCTION

Prior to reading the specific instructions for each tests, it is advisable to help put the examinees at ease by briefly providing some background information. For example, the administrator should introduce himself or herself, mention which tests are being given and roughly how long the session will last. The examiner may wish to state the purpose of testing and give a brief description of how the tests were developed and what they are designed to measure. At this time, the administrator should hand out the answer sheet(s), pencils, and a background form.

B. TEST INSTRUCTIONS

1. READING TEST (Hand Out Reading Test and Answer Sheet)

Administrator, read the following aloud:

DO NOT OPEN THIS TEST UNTIL I INSTRUCT YOU TO DO SO.

This is a test of your reading ability. On the following pages are a number of passages. Read each passage carefully, then read the questions following the passage. Choose the best answer for each of the questions. Blacken the circle which corresponds to your answer on the answer form. Mark only one answer per question. You may refer to the passage while answering the questions. You will have 40 minutes to complete this test.

OPEN TO THE FIRST PAGE OF THE READING TEST. READ THE FIRST EXAMPLE AS I READ IT.

Example:

Mountains and hilly terrain often involve sharp curves which must be negotiated at low speeds. Many of the tight curves on railroads were constructed using nineteenth century technology. Explosives, hand labor, and draft animals were used to construct much of the American rail roadbed. Gradually, the railroads are reducing grades, easing tight curves, and increasing the dimensions of tunnels. At the top of Maxwell Pass on the mountains that rise out of Piedmont, a number of curves were eliminated only a few years ago by modern construction equipment. Despite these recent efforts however, many of the original restrictions on land profiles still exist.

According to the passage, which of the following statements is correct?

a. Tight curves on railroads were a result of 20th century technology

b. No new railroads have been build since 1910.

c. Nothing has been done to correct the problem of tight curves.

d. A train's speed is limited by the grades and curves it encounters.

The correct answer is "d". Answer "a" is incorrect because the passage states that tight curves were a result of nineteenth century technology. Answer "b" is incorrect because the building of new railroads is not mentioned in the passage. Answer "c" is incorrect because efforts have been made to correct the problem of tight curves.

BEGIN TIMING. AFTER 40 MINUTES, COLLECT THE TESTS AND ANSWER SHEET.

2. LOGICAL REASONING (Hand Out Logic Test and Answer Sheet)

Administrator, read the following aloud:

DO NOT OPEN THIS TEST UNTIL I INSTRUCT YOU TO DO SO.

This is a test of logical reasoning. You will be presented with a list of facts. Then look at each conclusion. Only one conclusion is correct for each set of facts. All other conclusions are definitely wrong or there is not enough information given to tell if they are correct or incorrect. For each question, select the one correct conclusion. Blacken the space which corresponds to your answer on the answer form. You will have 30 minutes for this test.

OPEN TO THE FIRST PAGE OF THE LOGICAL REASONING TEST.

READ THE EXAMPLE ITEM.

Example:

Facts:

All signal flags indicate caution.

All blue flags are signal flags.

Which conclusion is correct?

a. No blue flags indicate caution.

b. All blue flags indicate caution.

c. Green flags indicate caution.

d. Some signal flags are red.

In the example above, the first conclusion "a" is incorrect because blue flags are signal flags, and therefore do indicate caution. For both conclusions "c" and "d" there is not enough information presented in the facts section to decide whether or not these conclusions are correct and incorrect. Therefore, the correct conclusion is "b".

Administrator:

DO YOU HAVE ANY QUESTIONS? YOU MAY TURN TO THE PAGE AND BEGIN TESTING.

BEGIN TIMING. AFTER 30 MINUTES, COLLECT THE TESTS AND ANSWER SHEET.

3. DICHOTIC LISTENING TEST (Hand Out Tape Recorders and Answer Sheet)

Administrator, read the following aloud:

THE INSTRUCTIONS FOR THE NEXT TEST ARE RECORDED ON TAPE. WHEN YOU PUT ON THE HEADPHONES, CHECK TO MAKE SURE THE LEFT EARPIECE IS ON YOUR LEFT EAR AND THE RIGHT EARPIECE IS ON YOUR RIGHT EAR. PUT ON THE HEADPHONES NOW AND PRESS "START."

BEGIN TIMING. AFTER 30 MINUTES, COLLECT TAPE RECORDERS, HEADSETS, AND ANSWER SHEETS

IV. TEST SCORING

A. TEST KEYS

On the following pages you will find the answers keys for the Reading, Logical Reasoning, and Dichotic Listening tests. Obviously, it is important that these keys remain strictly confidential. In addition, the special answer sheet for Dichotic Listening Test is also included. Following the answer keys, are the instructions for determining raw scores, converting raw scores to standard scores, and then applying cut scores and making selection decisions.

CORRECT ANSWERS FOR READING TEST

1.	d	22.	c
2.	d	23.	d
3.	d	24.	c
4.	c	25.	unscored
5.	d	26.	a
6.	b	27.	c
7.	a	28.	a
8.	a	29.	a
9.	c	30.	c
10.	b	31.	b
11.	c	32.	c
12.	a	33.	b
13.	c	34.	c
14.	d	35.	c
15.	a	36.	unscored
16.	c	37.	c
17.	b	38.	b
18.	c	39.	c
19.	b	40.	b
20.	d	41.	d
21.	a	42.	c

Scoring: Examiner's raw score equals number of correct answers

CORRECT ANSWERS FOR LOGICAL REASONING TEST

1.	b	16.	d
2.	c	17.	b
3.	b	18.	unscored
4.	c	19.	b
5.	d	20.	b
6.	c	21.	a
7.	a	22.	d
8.	b	23.	a
9.	b	24.	c
10.	a	25.	b
11.	b	26.	d
12.	c	27.	b
13.	b	28.	b
14.	c	29.	d
15.	d	30.	d

Scoring: Examiner's raw score equals number of correct answers

CORRECT ANSWERS FOR DICHOTIC LISTENING TEST*

1.	2	1	4	3	0
2.	2	1	3	9	0
3.	1	9	8	4	6
4.	6	4	3	5	7
5.	8	2	7	6	5
6.	9	1	7	3	5
7.	4	8	7	5	7
8.	4	3	5	7	2
9.	6	3	0	9	7
10.	2	8	6	0	5
11.	7	6	3	2	4
12.	5	1	9	3	7
13.	1	7	8	5	4
14.	5	8	0	6	2
15.	7	9	2	0	3
16.	5	2	6	5	8
17.	1	0	9	4	3
18.	4	6	2	0	8
19.	3	6	0	9	2
20.	5	0	1	8	9
21.	8	1	5	4	2
22.	4	9	3	1	7
23.	9	0	5	8	1
24.	8	2	6	0	4
25.	2	6	3	9	0
26.	3	9	7	1	4
27.	2	5	4	1	8
28.	9	0	3	4	1
29.	3	4	9	0	1
30.	8	0	2	6	4
31.	2	9	0	6	3
32.	9	8	1	0	5
33.	2	4	5	1	8
34.	7	1	3	9	4
35.	1	8	5	0	9

* Refer to page 17 for specific scoring instructions.

DICHOTIC LISTENING TEST ANSWER SHEET

1. ☐ ☐ ☐ ☐ ☐
2. ☐ ☐ ☐ ☐ ☐
3. ☐ ☐ ☐ ☐ ☐
4. ☐ ☐ ☐ ☐ ☐
5. ☐ ☐ ☐ ☐ ☐
6. ☐ ☐ ☐ ☐ ☐
7. ☐ ☐ ☐ ☐ ☐
8. ☐ ☐ ☐ ☐ ☐
9. ☐ ☐ ☐ ☐ ☐
10. ☐ ☐ ☐ ☐ ☐
11. ☐ ☐ ☐ ☐ ☐
12. ☐ ☐ ☐ ☐ ☐
13. ☐ ☐ ☐ ☐ ☐
14. ☐ ☐ ☐ ☐ ☐
15. ☐ ☐ ☐ ☐ ☐
16. ☐ ☐ ☐ ☐ ☐
17. ☐ ☐ ☐ ☐ ☐
18. ☐ ☐ ☐ ☐ ☐
19. ☐ ☐ ☐ ☐ ☐
20. ☐ ☐ ☐ ☐ ☐
21. ☐ ☐ ☐ ☐ ☐
22. ☐ ☐ ☐ ☐ ☐
23. ☐ ☐ ☐ ☐ ☐
24. ☐ ☐ ☐ ☐ ☐
25. ☐ ☐ ☐ ☐ ☐
26. ☐ ☐ ☐ ☐ ☐
27. ☐ ☐ ☐ ☐ ☐
28. ☐ ☐ ☐ ☐ ☐
29. ☐ ☐ ☐ ☐ ☐
30. ☐ ☐ ☐ ☐ ☐
31. ☐ ☐ ☐ ☐ ☐
32. ☐ ☐ ☐ ☐ ☐
33. ☐ ☐ ☐ ☐ ☐
34. ☐ ☐ ☐ ☐ ☐
35. ☐ ☐ ☐ ☐ ☐

B. DETERMINING RAW SCORES

For both the Reading and Logical Reasoning tests, the examinee's raw score is simply the number of correct responses. Write the total number of correct answers in the top right corner of the examiner's answer sheet. For the Dichotic Listening test, scoring is somewhat more complex. The Dichotic Listening test consists of 35 items; each item contains 5 possible responses. Raw scores for any particular item therefore can range from 0 to 5, with raw scores on the entire test ranging from 0 to 175.

For any particular item, an examinee's raw score depends on the sequence of answers. An examinee is given one point for each number that corresponds to the number on the answer key. The numbers must appear in the proper sequence regardless of whether there are other numbers between the correct numbers. Consider only the responses that appear in the 5 lines provided on the answer sheet; sometimes, an examinee will respond to an item with more than 5 numbers. For example, consider the initial item on the Dichotic Listening test. The correct sequence of numbers is given below, along with 7 hypothetical responses and the scoring rationale for each response.

Correct Answer:	2	1	4	3	0			
Examinees' Responses							Score	Rationale
a.	2	1	4	3	0		5	All five digits are given in the current sequence.
b.	2	1	4	9	3	0	4	Four of the five digits are given (2, 1, 4, 3) in the correct sequence. The examinee does not receive credit for the 0, because it does not appear in the lines provided on the score sheet.
c.	5	3	2	1	4		3	The examinee receives one point each for the 2, 1, and 4. No point is awarded for the 3, because it does not appear in proper sequence after the 2, 1, and 4.
d.	2	6	4	7	3		3	One point each is awarded for the 2, 4, and 3. No points are given for the 6 and 7.
e.	5	7	9	8	6		0	None of the digits match the correct answer, so zero points are given.
f.	5	3	9	8	2		1	A single points is given for the 2. No points are given for the 3 because it is not proper sequenced order.

g.	2	1	_	_	_	2	One point each is given for the 2 and the 1.

Write the examinee's raw score to the left of each item. The examinees raw score for the entire test is the sum of the raw scores for the thirty-five (35) items.

C. CONVERTING RAW TO STANDARD SCORES

Raw scores are useful in comparing examinees' performances on any single test, but are limited in what they can tell us about overall performance on multiple tests. The goal of this study is to derive a single composite score, based equally on all 3 test scores that is indicative of the subjects overall performance. Examinees can then be meaningfully compared in terms of overall test performance.

One method for deriving a single composite score would be to simply sum the raw scores from all three tests. However, if the standard deniations of the test scores differ, then the test with the greater standard deviation will be more heavily weighted in terms of the composite score. One can combine this problem of combining raw scores by converting the raw scores to standard or z scores. Standard scores covert all scores to the same scale before summing them so that each test contractors equally to the composite score.

Statisticians often avoid the problems of combining raw scores by first converting to standard scores, or z scores. Standardized scores convert all raw scores to the same scale before summing them, so that each test contributes equally to the final composite. Raw score are converted to standard scores via the following equatio:

$$z = (x_i - u) \div sd$$

Where z = the standard score

$$sd = \sqrt{\frac{\text{sum } (x_i - u)^2}{n - 1}}$$

x_i = the raw score

u = arithmetic average

sd = the standard deviation

n = number of test scores

Based on the data we collected from a sample of locomotive engineers, we have estimates of u and sd for each of the 3 cognitive tests. Substituting these into the previous equation we obtain:

z Reading $= (x_i - 28.66) \div 5.55$

z Logical Reasoning $= (x_i - 17.96) \div 6.12$

z Dichotic Listening $= (x_i - 165.25) \div 30.49$

Subsequently, the overall composite score can be computed by adding the standardize scores of the 3 tests. Hence,

Overall Composite Score = z Reading + z Logical Reasoning + z Dichotic Listening

For example, suppose an individual obtained raw scores of 35 for Reading, 25 for Logical Reasoning, and 150 for Dichotic Listening. The individual standardized scores are computes as follows:

z Reading = (35 - 28.66) ÷ 5.55 = 1.14

z Logical Reasoning = (25 - 17.96) ÷ 6.12 = 1.15

z Dichotic Listening = (150 - 165.25) ÷ 30.49 = -0.50

The individuals composite score is therefore 1.79 = 1.14 + 1.15 + (-0.50)

D. CUT SCORES AND DECISION MAKING

The purpose of a cut or minimum passing score is to determine a score which eliminates from consideration those applicants unlikely to perform will in training and on the job. The data we used for setting the cut score was derived from the concurrent validation sample. Expectancy tables were developed using the battery scores and the scores on the air brakes and General Rules Tests. We tried to identify a cut score which differentiated between those who are more and less likely to pass the tests.

Relatively few trainees fail either test. We, therefore, decided to use a cut score which would be conservative, eliminating those performing least well on the selection tests. The cut score we recommend is -2.0 using this cut score, only 4% (4 of 97) of those who passed the selection battery would fail the Air Brake Test, compared to 24% (4 of 17) of those who fail the battery. The figures on the General Rules Test are 7% (7 of 97) for those who passed the battery compared to 29% (5 of 17) who failed the battery.

We strongly recommend that railroads, either individually, or in consortium studies, undertake additional research on the cut score. The score we proposed should be considered provisional until such data are available.