

CHAPTER FOUR

VALIDATING A DRIVER TRAINING PROGRAM

Chapter 4

Validating a Driver Training Program

Evaluation is the means used to determine if the presented information was received and to what degree. It measures the effectiveness of the instructor, the materials, and the learning ability of the student. This chapter is directed toward the development of methods and processes for evaluating driver proficiency. The emphasis has been placed upon analyzing practical driving exercises rather than the development of a written examination. As a note, written examinations can be developed for testing the knowledge level of the student by selecting questions relevant to the process of emergency vehicle operations.

A review of the driver training curriculums of many different agencies from across the nation indicates that many similarities exist in the practical exercises as well as the paper-pencil type of examinations. Most often a disparity is found in the criteria for passing or failing student drivers. Information has been compiled in this chapter regarding testing, evaluation, and validation that will assist program training administrators driver training instructors with providing a method of instruction and evaluation that is uniform, as well as valid.

While many driver training programs use exercises to evaluate driver proficiency (serpentine, evasive lane change, or T-driveway exercises, etc.) the standards for passing these exercises differ considerably. In many instances, the driving methods taught vary, which in and of itself accounts for the absence of uniform standards. In other instances, agency goals for training differ as well as the unique demands of population, weather, terrain, etc. In any case, efforts to test and evaluate trainees should be conducted only after establishing that tests are indeed valid.

In reviewing methods of evaluating driver proficiency it is apparent that there is much commonality in the exercises that are used to determine an individual's driving skills. While many agencies use the same exercises to evaluate driving skills, their standards vary. As a result, there is a need to establish a valid process for determining driver proficiency. The following is an outline of a process training agencies may use in establishing criteria for determining driver proficiency.

Determine Test Specifications

Driving instructors should develop a series of test specifications. Minimally, these specifications should include:

1. skills to be measured
2. exercises used to measure a specific skill

3. scoring process to be used

For example, assume you have decided to measure backing skills. The question will be what kind of exercise is to be used to measure backing skills. In this case, let's say you have decided to use a serpentine maneuver, wherein the student must back a vehicle through a series of cones.

The next step will be to determine how to score the exercise; i.e., is the driving instructor going to score the number of cones struck, the amount of time a student takes to complete the exercise, steering technique, and vehicle control? Or, on the other hand, will striking any cone result in failure, as any collision with static objects can be considered a performance failure. When developing test specifications consideration must be given to the concept of "test validity."

Selection of Test Groups

The next step is to identify two groups of drivers: a group of excellent drivers and a group of poor or substandard drivers. Some factors to consider in selecting individuals for each group are: prior driving records, collisions in which the driver was at fault, supervisor assessments of driving skills.

It is important that the samples (the excellent group and the poor group) be of sufficient size to enhance the reliability of the results. It is recommended that a minimum sample of sixty officers be used: thirty in the excellent group, and thirty in the substandard group.

Conducting the Test

Generally, it is difficult to test sixty people at once. Therefore, you may decide to divide the sample into several groups. If you do this, it is best to use mixed groups, excellent drivers mixed with substandard drivers. Note: most assembled groups of drivers randomly selected are inherently comprised of the excellent and substandard drivers. A representative sample of each group is the key to success. Other considerations are:

1. The drivers should not know the criteria by which they were chosen.
2. Neither the drivers nor the instructors grading the exercise should be informed as to who are the excellent drivers and who are the poor drivers.

The following are general guidelines:

1. Each driver should receive the same instructions in the same manner and should proceed through the testing process in exactly the same sequence.

2. If possible, multiple raters should be used to evaluate each driver during each exercise. This will be useful in determining the reliability of the evaluation process. It is important to keep in mind that a major factor which reduces test reliability is "scoring unreliability." A test score may be unreliable to the extent that the score is affected largely by the person doing the scoring. By comparing the scores of multiple raters, you can determine the degree of objectivity in the scoring process.
3. Two types of scores should be obtained:
 - a. Scores for each exercise
 - b. A composite score
4. When analyzing test scores it will be important to determine if specific exercises, as well as the total test processes, were able to discriminate between superior drivers and substandard drivers.
5. Following a familiarization session, the student should drive the exercise without any warm-up laps. The student should take as much time as needed to safely complete the exercise; but the time taken to complete the exercise is recorded. At the end of the training session the student again drives through the exercise for time. The difference between the first and last runs should show significant improvement, which assists in validating the exercise.

Test Analysis

After conducting the test, the results must be analyzed. This is accomplished by determining whether the test criteria for passing and failing discriminated between superior drivers and substandard drivers. Fundamentally, you need to answer the question: Did the good drivers perform consistently better than the poor drivers? Several forms of statistical analysis can assist in quantifying the degree of sub-test such as individual driving exercises and total test discrimination.

When analyzing test results, a concern which should be addressed is whether an individual can have one or more significant driving weakness(es) and still pass the total test. For example, assume a driver is unable to perform a backing exercise with minimal proficiency. Further, assume that this same individual performs poorly on all backing maneuvers, but is able to obtain a passing score for all other exercises and for the total test. The question arises: Should this person be allowed to pass the program? A driver who lacks driving skills in only one or two areas may be a hazard, especially if that skill area is a critical one.

In reviewing driving test criteria, it is found that students frequently pass the total test even though they failed to demonstrate minimal proficiency in a major skill area, i.e., backing, braking, cornering, or evasive responses. This is an area that demands extensive review. Allowing a student to receive a successful overall rating while overlooking deficiencies in critical areas could set the student up for failure in the real world surroundings of a critical incident. Additionally, if the student is subsequently involved in a crash where the deficiencies are exploited, citizen safety may be compromised and department civil liability may be applicable.

Follow-Up Tests

Time and circumstances permitting, when establishing the standards for testing, it is desirable to bring the same group of drivers back for retesting sixty days after the initial test. The theory here is that, if test scores are to be useful predictors of future levels of performance, then they should not vary appreciably over relatively short periods of time. This assumes, of course, that no significant changes have occurred in the individual's abilities.

Reasonably, we can expect slightly higher driver scores due to the fact that the drivers will be more familiar with the program. However, we are most concerned with whether the test criteria is still able to discriminate between good and poor drivers.

Pre-Tests

Although the emphasis of this chapter has been placed on standards for the practical aspects of emergency vehicle operations, each practical test should be preceded by a written test. Pre-tests are recommended for every program, as they should clearly establish the knowledge level of the student as being much higher after having taken the program. This further validates the rationale for training. As a side note to written examinations, no student should be allowed to successfully complete the program without ensuring that he knows the correct answer to every question, regardless of his final score. This is generally accomplished by providing a Test Review Session at the end of the program after the tests have been corrected.

Establishing a Minimum Level of Competency

Determine the point at which the test is able to discriminate between superior and inferior drivers, then set the minimum passing level for students of basic driver training. The level should be set at, or slightly lower than, the level which separated the superior and inferior drivers.

Since every program is different, no specific standard is recommended. However, attention must be given to the following suggestions:

1. The program should reflect a realistic environment to be encountered by the students when they are driving while on duty.
2. The time element to complete the exercise should provide a reasonable level of stress, without promulgating the "Pedal to the Metal" attitude.
3. Students should not be allowed to strike any cones and still pass the exercise. Indoctrinating a student that it is allowable to strike a foreign object while en route to a specific location without a maximum penalty, fosters a dangerous attitude that later reveals itself in a "real life" situation.
4. Emphasis should be placed on "as safe as possible" in place of "as fast as possible."
5. Passing scores do not have to reflect 100%. However, any noticeable mistakes must be corrected prior to allowing the student to pass.

If the standard does not allow for some student failure, it may be unrealistic. Some experts claim that a valid program/test should have an overall failure rate of 5% to 15% at a minimum. This is established after several programs and, although each class is different, the overall pass/fail rate should be relatively consistent.

An alternate view is that there should be very little actual failure for veteran officers. Indeed, some may have superior skills. However, emphasis should be primarily focused on performing essential law enforcement functions at an acceptable level and not who is the best driver.

Skills Evaluation

The instructor's task of identifying which exercises will develop or test a student's level of driving proficiency can oftentimes be a difficult one. The individual instructor needs to determine what the facility will allow and the objective of the training program and then establish a program that will meet the needs of that particular agency.

In an attempt to assist the instructor, an Exercise Selection Matrix system has been developed. (See figure next page.) The system is designed to indicate the exercises that can be used to develop a particular skill. Some of the exercises are designed to develop or test several skills and they can be established according to the facility constraints in accordance with the objective. As a suggestion, the objective should always be to establish exercises that will test the student's ability to drive in an environment that is as close as possible to his work conditions.

Validity

Test validity is the single most important variable in the development and implementation of a test. The term "validity" refers to the accuracy and usefulness of inferences which are drawn from test results. Or more simply, it is answering the question, "Does the test measure what it was intended to measure?"

In law enforcement driver training, it is imperative to design testing processes that assist in determining the level of driver proficiency. Therefore, tests which accomplish this purpose must be developed. Of equal importance is the concern that the conclusions drawn from these test results be accurate. By designing exercises that are reflective of the work environment, the instructor ensures that the tests not only will determine proficiency but will be "valid" in the sense that they measure skills applicable in real life circumstances.

If such tests in any way affect career opportunities or benefits, then they must measure what have come to be known as "essential" tasks, not marginal ones. Essential tasks are those that, if they were removed from the job duties, would significantly alter the nature of the job, i.e., patrol officer. Marginal tasks are those that, if removed or performed by another officer or employee, do not significantly alter the nature of the job.

Additionally, essential job tasks or functions should be listed in a description of the job itself, and they should be identified specifically as essential tasks. Driving a law enforcement vehicle is clearly an essential function. This function can have a series of subsidiary tasks along the lines of those found in the matrix.

Exercise Selection Matrix

Exercise Selection Matrix	Skills											
	1	2	3	4	5	6	7	8	9	a	b	c
Angle Parking	X		X						X	X		
Baird's Judgment	X		X		X		X			X		
Controlled Braking	X	X	X				X			X		
Dutton's Weave	X		X					X		X		
Evasive Steering	X	X	X		X		X			X		
Lane Change	X	X	X		X					X		
Left-Side Road Turn	X		X							X	X	
Maximum Braking	X	X	X				X			X		
90E Turn	X		X	X			X			X	X	
Off-Road Recovery	X		X							X		X
Parallel Parking	X		X					X	X	X		
Perpendicular Parking	X		X					X	X	X		
Perpendicular Parking	X		X					X	X	X		
Power U-Turn	X		X							X	X	
Reverse Serpentine	X		X					X		X		
Right Side Road Turn	X		X							X	X	
Serpentine	X		X							X		
Shuffle	X									X		
Skid Control	X		X			X	X					
T-Driveway	X		X					X		X	X	
Turnaround	X		X							X	X	
U-Turn	X		X				X			X	X	
Y-Turn	X		X				X			X	X	

1-Steering
 2-Braking
 3-Acceleration Control
 4-Cornering
 5-Lane Change

7-Skid Avoidance
 8-Backing
 9-Parking
 a-Vehicle Positioning
 b-Turning

6-Skid Control

c-Off-Road Recovery

Appraising Validity

Writers on the topic of validity frequently discuss different kinds of validity. They explore the differences between empirical and/or statistical validity and logical or curricular validity. Various professional organizations have distinguished four types of validity: *predictive*, *concurrent*, *construct*, and *content*. While some argue that these are different "types" of validity, others argue that these are just different methods of assessing validity. For all practical purposes, such discussions are of little concern to those engaged in driver training.

What is important, however, is to design tests in such a manner that they reflect the kinds of situations students will encounter on the job. Ultimately, they should prove that students who do well on the driving tests will also do well in various driving situations while on-duty (*predictive validity*). Statistically, the degree to which a test can predict success or failure can be established. To do this there is a need to correlate test performance to on-the-job performance.

Unfortunately, measures of on-the-job driving performances are frequently difficult to obtain. A problem thus arises. By failing to establish *predictive validity*, significant attacks on the question of validity can be made. However, all is not lost. What needs to be done is the exploration of alternative methods of establishing validity. The best alternative is to provide a foundation for *content validity*. This is not to suggest that developers of driving tests should not be concerned with after-training performance. Rather, it is suggested that inferences as to the degree of correspondence between behavior in the test situation, and later on-the-job behavior, are to be made without resort to direct comparisons.

Content validity, generally, refers to the degree to which a test represents the range of on-the-job behaviors one will be required to perform. For example, a comprehensive law enforcement driving test would not be *content valid* if it simply measured backing and steering skills, thus excluding cornering, speed control, and roadway position.

In many instances, establishing *content validity* is a logical process for which the rules are frequently ambiguous. Experts generally agree that there is no agreed-upon criterion for determining the extent to which a measure has attained *content validity*. Given this ambiguity, experts have noted that inevitably *content validity* rests mainly on appeals to reason regarding the adequacy with which important content has been sampled, and on the adequacy with which the content has been cast in the form of test items.

In spite of inherent ambiguity in establishing content validity, a useful approach is to ask veteran patrol officers and supervisors if exercises reflect on-the-job type demands. A compelling case that a test is content valid can be made when a vast majority of veteran officers can be documented as agreeing that the test is indeed reflective of job demands.

To establish *content validity* test developers should:

1. Define the total range of job-related behaviors that the test is to represent.
2. Define sampling procedures to determine which behaviors will be measured on the test. This step is necessary because it is frequently impractical to measure all of the on-the-job behaviors.
3. Develop and implement a representative test.
4. Document the comments of incumbent officers.
5. Do a follow-up analysis to determine whether or not the test does indeed distinguish between superior and poor drivers.

Qualities of a Good Test

When developing driving tests it is suggested that the following qualities should be strongly considered:

1. *Validity*: Does the test measure what it is intended to measure?
2. *Reliability*: How consistently does the test measure what it was intended to measure?
3. *Objectivity*: To what extent does personal judgment affect the scoring of a test and the interpretation of a student's response to a test situation?
4. *Discrimination*: Does the test distinguish between excellent drivers and poor drivers?
5. *Comprehensiveness*: Does the test include a representative sampling of on-the-job behaviors?

Reliability

Reliability is a measure of consistency. Does a test render the same results consistently and is it free from error? Reliability is critical because a test cannot be valid for any purpose if it is not reliable. It is necessary to have a basic understanding of the principle causes of test unreliability. Once these causes are understood, then we can take appropriate steps to minimize their influence.

Principal Causes of Test Unreliability

1. Scoring Unreliability

A test score may be unreliable in the sense that the score depends to a great extent upon the particular person who does the scoring.

2. Content Unreliability

As previously mentioned, a test usually consists of a sampling of the total on-the-job behaviors. The sample may be poor in that it is too small.

3. Temporal Unreliability

If test scores are to be useful, they must not vary appreciably over relatively short intervals of time. There is an assumption that no significant changes occur in an individual's abilities in a short interval of time.

Conclusion

A great deal of thought and consideration should be given to the standards by which driving students are evaluated, graded, or ranked. Standards must reflect the actual job demands and not what anyone thinks is "good" for officers to know. Failure to perform at an acceptable or passing level must be a clear indicator that the officer may be a danger to themselves or others.

The same thing applies to practical driving exercises. Any exercise must reflect activities that are essential to the safe and effective performance of the driving function. A driver training program should be able to refer to a matrix of exercises and needed skills. To the extent possible, exercises should "look like" activities that officers perform on a regular basis and are an important part of the job.

Postscript

While some would argue that an effective test fails some of the students, the more important issue is whether or not the test discriminates between safe and unsafe drivers. Indeed, some drivers will do better than others. In this sense, a percentage score will indicate a student's level of skill to some extent and provide a ranking of students where it makes sense to do so.

An agency having an effective recruitment and in-service driver training program might reasonably expect no failures in its basic or in-service driver training programs. Or, at a minimum, those failures that are experienced should be a function of physical, psychological, or attitudinal problems. A failure indicates that the observed driving behaviors are unsafe. The agency should consider reassignment out of driving until remediation is successful.