Measuring the Outcomes of Driver Training: University of Southern California Driver Performance Test

Margaret Hubbard Jones, Traffic Safety Center, University of Southern California

A new test for evaluating driver performance is described. The objective was to develop a test that is reliable, valid, and feasible for routine administration in the high schools. This test requires 30 min and is scored by a trained coder. The scoring is simplified to permit the coder to focus on observing and judging driver behavior. Standards for performance are learned by the coders during a 40-h training program. The test requires driver interaction with moderately heavy traffic and is intended to test the limits of driver performance. Intercoder agreement was about 80 percent, even though there were different seating positions. Some changes in scoring assignments and training are expected to improve the reliability of the test. This pilot study was carried out with 197 students at the end of their driver training course. A number of part scores and subtotals were used so that faults could be diagnosed. In general, these novice drivers did poorly in visual scanning. However, there was a wide range in driver proficiency. A hypothesis, the validation of which needs a large-sample study, is proposed to test the reason for the poor scanning.

The effectiveness of driver training courses has suffered because these courses lack a reliable, valid, and standardized measure of student performance in the normal traffic environment. Without such a measuring device, neither alternative programs nor individual students can be evaluated. Hence, curricular decisions continue to be made on the basis of personal preference, and course effectiveness is not achieved. Not only is there a pressing need for a reliable and valid test for beginning students, but there is also a need for evaluating experienced drivers because of an increasing concern for promoting driver training programs for the elderly, minorities, and the handicapped.

There have been many performance tests devised in the past, and these include (a) the McGlade test (6), which has a strong instructor bias that makes it unreliable; (b) the Michigan State BETSS test (2, 3), which is too costly to be practical for routine testing and does not have route criteria; (c) the Rockwell test (9), which does not evaluate traffic interaction; (d) the Quenault test (8), which is only in the pilot-test stage and would be much too expensive for routine use; (e) the German behavioral survey (1), which is too time-consuming and not validated; and (f) the California Department of Motor Vehicles test, which is a cost-effective experimental road test in which the route criteria are available but the scoring is not reliably controlled. Thus, there is still no test that can be used satisfactorily by a school district or a research team to evaluate large numbers of drivers. Because there is no acceptable criterion of real-world performance other than accidents, large numbers of students must be tested to validate a program. A valid intermediate criterion would make the task of developing effective curricula simpler, quicker, and less costly. However, such an intermediate criterion must be validated by using a large-sample study.

The objective of this study was to develop a test that is feasible, reliable, and valid for use in a school program. If the test is to be feasible, it must be as short as possible, which makes it consonant with the other goals. Therefore, the upper limit of time was set at 30 min. If a test is to be reliable, psychometric rules must be followed, the most critical of which is the requirement for a large number of independent observations. If a test is to be valid, a minimal requirement is that the items tap the most important aspects of safe driving behavior rather than those that have been the most convenient to test in the past. There are two further requirements: (a) The test must be simple to score, so that the coder can devote more attention to observing the driver's behavior than to scoring the results, and (b) the test must be safe for the public and the driver-education automobile. The last requirement is difficult to meet because high levels of skills, rather than minimal competencies, must be tested in moderately heavy traffic to assess driver capabilities. Thus, driver interactions with traffic are required so that the skills necessary for safety can be assessed.

DRIVER PERFORMANCE TEST

The University of Southern California (USC) driver performance test was devised to meet the previously mentioned needs; thus it is different from previous tests. The content of the USC test is based on the Safe Performance Curriculum (4), and those elements that are the last in a skill series and require on-road testing in traffic were emphasized. Because of the time limit, a selection of elements was required. (Because the testing is based on sampling the individual's skills or behaviors, a representative sample should be chosen.) Automobile control skills were not tested because these skills cannot be safely tested on a public street. These skills should be tested on a range that provides for a thorough, demanding, and safe test at a reasonable cost. Freeway and rural driving tests were eliminated for four reasons: (a) These tests require an inordinate amount of time, for even a brief excursion, and yield only one or two measures, which does not yield a reliable index; (b) the traffic volume on freeways varies according to time of day, and thus it is impossible to provide for comparable conditions for all cases; (c) freeway merges could be used in the absence of significant traffic, but they do not challenge the driver; and (d) it is unsafe to take novice drivers on crowded freeways. The final list of items includes those necessary for safe driver interaction with normal traffic, and these are based on the content of the Safe Performance Curriculum and the literature on accident causation. For example, a test for properly observing while backing up was included because backing up accounts for a significant number of pedestrian fatalities. As it must be, that test is highly standardized and leaves little room for variation from one location to another or individual foibles in scoring.

The USC test begins on a route that is constructed according to rigid specifications, which are contained in the route construction manual. These requirements include a certain number of specified kinds of intersections, streets, and traffic densities, which are based on counts. These requirements are shown in Figure 1, and the required maneuvers are shown in Figure 2. Once the route is determined, it is marked on a route map (Fig-

ure 3). This map is also the score sheet. There are three advantages to this method: (a) By following the route, the coder knows precisely where to mark the score, hence he or she can attend to the driver's behavior; (b) the coder is then set to attend to the next behavior; and (c) the scoring process is very simple.

Because the driving instructor is normally responsible for the safety of the vehicle, he or she cannot be the coder. The coder is a trained person who sits in the middle rear seat with a clipboard on his or her lap and attends to nothing but the driver's performance. However, the instructor is now assigned two important scoring duties, which are marked on his or her route map: (a) All hazards that are encountered are recorded and circled or crossed out, depending on whether the student handled the hazard properly; (b) every instance in which the instructor must take control, either physically or verbally, is also recorded. These instructor duties have only recently been assigned because they are part of the instructor's normal task. However, in the results reported here, these duties were assigned to the coders.

The coders are teacher's aides who are selected because they have certain minimal skills, but, in California, they are untrained. This type of person is selected after considering trainability, availability for a peculiar schedule, reliability of attendance, and cost. One requirement was that the coders have no special training in driving or driving instruction. This requirement promotes the aim of training the coders to produce identical judgments, which can best be accomplished if there is no diversity of entrenched opinions.

Figure 1. Sample of a route layout form.

School Marina High School Route: Inbound Outbound Circular

City Huntington Beach, CA.

TRAFFIC DENSITY	CHARACTERISTICS	MANEUVER	MINIMUM NUMBER	CHECK
High	Signalized	Left Turn	2	111
High or Medium	Signalized	Left Turn	2	1
Low	Uncontrolled	Left Turn	2	1111
High	Signalized	Right Turn	2	111
High or Medium	Signalized	Right Turn	2	11
Low	Uncontrolled	Right Turn	2	11
High or Medium	Signalized	Through	2	1111
Low	(Blind) Uncontrolled	Through	2	11
MIDBLOCK ENTERING AND LEAVING TRAFFIC				
High/Medium	Uncontrolled for High/Ned, to Low	Left Turn H/M to L	1	1
High/Medium	Uncontrolled for High/Med. to Low	Right Turn H/M to L	2	111
Medium	Uncontrolled or controlled by stop sign for Low to Medium	Left Turn L to M	2	1111
High/Medium	Uncontrolled or controlled by stop sign for Low to High/Med.	Right Turn L to H/M	1	1
MIDBLOCK LANE CHANGE AND TURNABOUT				
High		Lane Change	4	11X1
High or Medium		Lane Change	4	11X1 1
Low		Three-point Turnabout	2	11

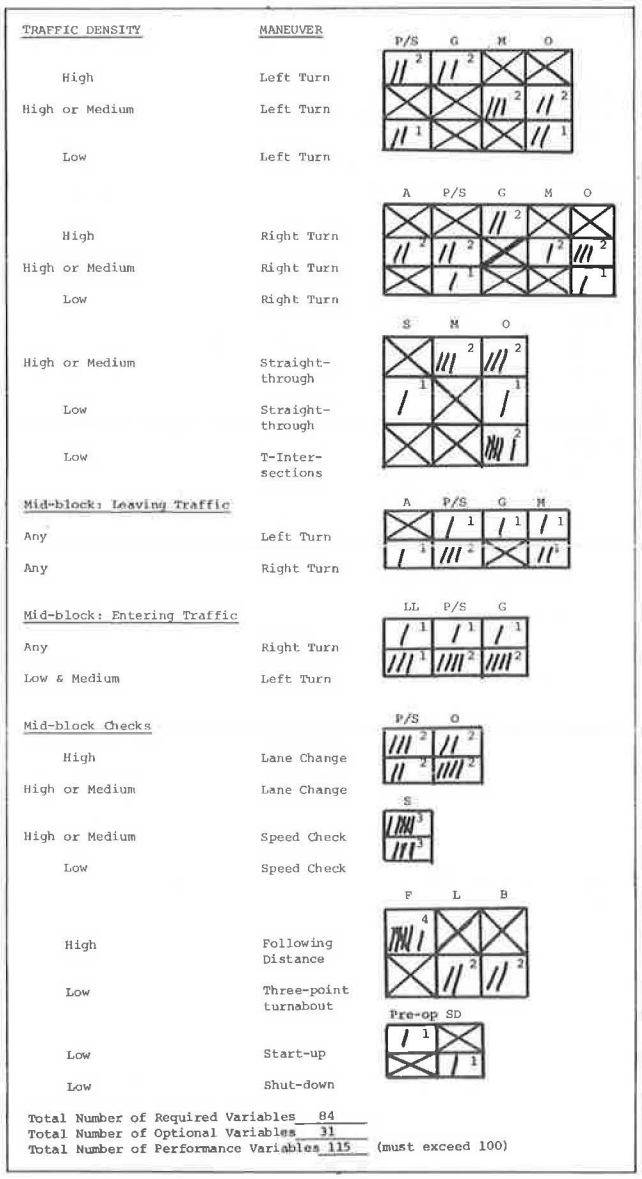
Driving time (experienced driver): 20½ minutes Date: 16 December 1975
Length of route 7.1 miles
Amount of time to lay out route 16 hours
Name of route constructor J. Wood

However, the coders must be intelligent, alert, and observant.

The training of the coders is guided by the training manual and materials, which include videotapes for practice scoring. The training requires 2 weeks; a little less than half of the time is spent in the classroom, and the remainder is spent in the automobile. The training in the automobile includes comparisons among coders scoring the same driver and also a comparison of each coder with the course instructor. Coders are not accepted if they fail to achieve the required intercoder reliability. The standard performance for each maneuver is described, and the coders are required to learn the criteria thoroughly.

The driver behavior to be rated at any one time is but a part of the total driving maneuver. This is done to enable the coder to attend closely to accurate performance, to compare it with the standard, and to mark the score sheet. Individual judgment is minimized as far as possible to maximize standardization among the coders. Thus, the coder may be directed to observe either eye movements (seen in the rear-view

Figure 2. Sample of a performance variables check sheet.



mirror) or path and speed, but never both; or the coder may be required to watch speed through one uncontrolled intersection and for the next intersection watch only eye movements. Only in this way can necessary singleness of attention be achieved. In the course of the test route there are more than 100 observations; therefore, each aspect of the driving behavior is rated a number of times. The aim is to randomly sample the critical skills over the entire duration of the drive. For example, it is not necessary to sample all the part behaviors involved in a left turn at the same instant; in fact, it is preferable to sample these behaviors independently. Thus, any sizable unbiased sample will be representative of the driver's usual behavior. The aim here is to obtain a large number of independent measures.

The test is a criterion-referenced test, which is in line with modern educational theory. The purpose is not selection of applicants, but an assessment of the success of the program in bringing all students to a standard performance criterion for safe driving in normal traffic. For this reason, the standards of performance are standards for all drivers. An attempt to devise a standard that takes into account experience and that measures on a sliding scale what is good for a student driver against what is good for a driver with 2 years of experience would be unreliable and defeat the purpose of the test. Because the standard of performance is absolute, experienced drivers can be tested as readily as novices; thus, the differences in performance between the two experience groups can be stated in behavioral terms. The further advantage of this test permits diagnosis of deficiencies in individual training and particular curricula.

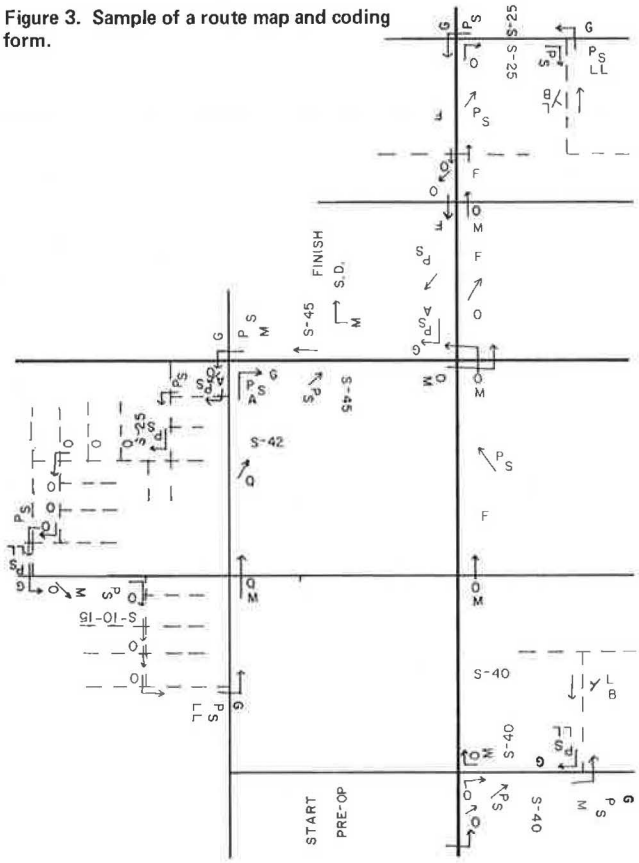


Figure 3. Sample of a route map and coding form.

CODER AGREEMENT

The crucial question is how well the coders agree in their judgments. Table 1 gives the intercoder agreement for all pairs. The overall mean agreement varies considerably by category, which is dependent on the complexity of the judgment, the precision of the critieria given the coders, and the coders' levels of skill. The more critical categories of speed and observation are in respectable agreement: 78 and 93 percent for speed and 84 percent for observation. The judgment of hazards by the coders was poor. The probable reasons are inadequate intuition about what a hazard is and, since these are unexpected events, concentration of attention on other things. This problem has subsequently been solved by assigning this coding function to the driving instructor because he or she attends to hazards and knows what constitutes a hazard. The category of instructor control has also been assigned to the driving instructor because the instructor knows when he or she has taken control. The other category with less than desired agreement is gap and following in which the major problem was where to code following distance. (Following distance is a problem because neither it nor gap acceptance can be scheduled as the other events are scheduled.) Extra training has since been added for this point. Overall, the mean percentage agreement was approximately 80, which is satisfactory under the comparison circumstances. The agreement would be higher if the coders could sit in the same position; of the two different seating positions, one is noticeably worse for observing eye movements and better for observing path and limit line. The positions are also quite different in terms of seeing the hazards, a difficulty that is no longer present. The transfer of the ratings of hazard and instructor control to the instructor and the improvement in following and gap training should improve the agreement still further.

An analysis of rating performance by coders 1 and 2 is as follows:

Performance	Coder 1	Coder 2	Total
Correct	547	565	1112
Wrong	326	308	634
Missed	104	108	212
Total	977	981	1958

Table 1. Mean percentage agreement by coder pairs after training.

Variable	5 and 1	5 and 2	3 and 1	3 and 2	1 and 2	Mean
Approach path and limit line	77	79	83	82	84	81
Speed for turns and lane change	75	75	76	86	80	78
Gap and following	80	60	66	60	67	67
Speed through mirror	90	88	95	95	95	93
	77	81	88	86	90	84
Location, backing-up, preoperation, and shutdown	88	87	84	74	94	85
Hazard	64	36	42	33	37	42
Instructor control	89	73	75	70	62	74
All variables	78	76	80	78	82	79

Table 2. Analysis of codings missed by coders 1 and 2.

Coder 1	Coder 2	Total	Following	Gap	Other
Missed	Correct	27	11	13	3
Missed	Wrong	5	0	0	5
	Subtotal	32	11	13	8
Correct	Missed	27	1	9	17
Wrong	Missed	9	1	0	8
	Subtotal	36	2	9	25
Missed	Missed	72	25	23	24
Total		140	38	45	57

Table 3. Means and standard deviations for USC driver performance test scores.

Variable	Original Test (N = 194) Mean	S.D.
Test duration, minutes	27.0	3.6
Observing score, percentage correct	51.6	16.3
Gap score	83.2	18.8
Mirror score	36.2	23.9
Observing subtotal	53.7	14.1
Path score	80.3	12.1
Speed for turns and lane change	67.5	21.1
Speed for through intersections	74.7	21.7
Control subtotal	73.1	13.4
Limit-line score	76.0	26.4
Approach score	52.6	30.9
Following score	93.0	18.1
Judgment subtotal	74.2	16.4
Y-turn location	45.9	27.1
Y-turn backing-up	44.1	41.7
Preoperation	42.8	49.3
Shutdown	92.6	26.0
Miscellaneous subtotal	52.5	19.4
Total score for all variables	65.0	9.8
Hazard score	69.8	31.5
Instructor control, N	2.7	2.4

There is no significant difference between the coders in the assignment of right, wrong, or missed, as indicated by the ϕ coefficient (0.0212). There is also no significant effect of coder, as indicated by X^2 (0.8024). Analysis of the codes missed is given in Table 2, and in one-half of the cases, the coders agreed. In many of these cases (two-thirds), the called-for behavior did not occur (no traffic to provide for gap or following distance). The coding instructions required that the symbol be underlined to indicate the observations that were missed under these circumstances. In the other third of the cases, there was a difference in judgment regarding whether the traffic was close enough for the called-for situation. Further, in a task requiring vigilance over a long period of time, it is inevitable that some blocking of attention will occur. The coders were instructed to use the code missing under such circumstances. In addition, for these comparison runs, one coder sat in the right rear seat, which was poor for observing some maneuvers. For short coders, it was difficult to observe speed. The short coder missed 16 of the speed codings that the other coder was able to code. However, there were no codings missed in the opposite direction. This situation only applies to the comparison coding and not to the data for students. The analysis of variance on student performance data is significant for the coder $p < 0.02$). The coders themselves do not differ significantly by direct comparison, which may mean that a coder might have been assigned to better students or that the coder had a tendency toward giving the student the benefit of the doubt. For the latter to be evident, a large sample of judgments should be used.

TEST RELIABILITY

A small sample of students (N = 67) was retested approximately 2 weeks after the original test. The correlations were not impressive (0.30 to 0.40). The technical problems with the test-retest technique as a measure of reliability in this particular situation are as follows:

1. The route test might have become familiar to some students, especially those who have better spatial memories.
2. The test itself represents a large proportion of the student's total driving time (20 percent), and this additional driving time could promote a substantial learning. It is well known that different abilities come into play at different stages in any learning process and that people learn at different rates. Hence, the extra training will change rankings significantly.
3. The complexity of the driving task implies that many individual and environmental factors will vary from day to day. Thus, the test may be theoretically reliable, but the driver-vehicle-environment interaction may differ from day to day. This difference reflects the current concept of varying individual risk. The use of an odd-even technique is also not suitable, because the test is not intended to be homogeneous and there are not enough measures of a single type of behavior to permit splitting the test into two minitests.

Another attempt to measure the reliability of the test will be made by using experienced drivers, who are unlikely to gather more knowledge by taking a course such as this. Since reliability sets the limit for validity, validity coefficients may shed some light on the problem.

RESULTS

The pilot-test data, reported here, were obtained from 194 driver education students in a moderately large California school district. The students were tested immediately after completing their on-road training. All students who completed the program in midsemester were tested. Table 3 gives the means and standard deviations for the USC driver performance test scores. The test duration was 27.0 min (the maximum duration was set at 30 min). The scores (which are given as percentage correct) are far from perfect: They vary between 36 and 93 percent. Thus, it is apparent that the diverse skills are unevenly learned at this point. Of even greater interest are the sizable standard deviations that indicate the diversity of skills among students who are the same age and have approximately the same amount of training and experience (5). Thus, driver education programs should have individually tailored instruction to be the most cost-effective.

The intercorrelations among the subcategories are typically low, indicating that there are different complexes of abilities. However, the two speed measures were significantly related ($r = 0.43$, $p < 0.001$). The following were also found

1. Students scan intersections only about half the time and check their mirrors even less; they are flying blind a good deal of the time and tend to stare straight ahead with unmoving eyes.
2. For three-point turnabouts, students will frequently choose an unsafe location.
3. Students generally do not look back while they are backing up in a turnabout situation.

The last two items can be easily learned within the limits of the present driver education programs, if they are given more emphasis.

A number of these students were retested 2 weeks later. Most behaviors showed improvement, particularly those for speed when turning. Originally the speed was too slow for most students coming out of a turn, but occasionally the speed was too fast going into a turn. The limit-line observance deteriorated, but moved toward the norm for the few experienced drivers tested.

An analysis of variance was made for route, estimated traffic density (which varied somewhat with the time of day but was controlled within limits), and coder for each subtotal score. The routes are not significantly different, and the small variation in traffic does not matter, given the constraints of the original route construction. Although instructors do not differ, coder differences were statistically significant, as mentioned above.

DISCUSSION OF RESULTS

The USC driver performance test appears to have met the objectives of the study in terms of time limit, code reliability, safety, and ease of scoring. In addition, it permits a breakdown into part scores that represent the various facets of this complex psychomotor and judgmental task. The validation of the test must await a long and costly study of a large sample of students and their first-year accident records, which is currently under way in Georgia. Further work with experienced drivers is also planned.

The objective description now available of typical performance patterns of new driver education graduates leads to the proposal of an important hypothesis that is currently testable on a small sample. The failure of students to scan adequately or often requires consideration. It is well known that the driver is often overloaded with information, which engenders considerable stress. Probably the most useful model of driving is as a two-function task that requires divided attention between automobile control and visual scanning. However, if automobile control is uncertain, then it receives priority because without control scanning will be useless. This concept is attested to by Mourant's and Rockwell's description of the eye movements of novice drivers (7). The hypothesis proposed here is that these students cannot adequately scan because they have not yet developed sufficiently sure and automatic responses to automobile control. Until they achieve this control, they cannot be taught to scan properly. If proper scanning were taught under those conditions, it would overload the system to the point where driving behavior would break down. This hypothesis could be tested immediately by training student drivers to scan using a control group design. This research can be done on a relatively small scale with several groups who would receive training in various amounts. It may prove to be that adequate scanning behavior will be the true test that determines when a person can drive safely alone.

ACKNOWLEDGMENTS

This research was performed under a contract with the National Highway Traffic Safety Administration. The opinions expressed are mine and do not necessarily reflect those of the National Highway Traffic Safety Administration.

REFERENCES

1. B. Biehl, G. H. Fischer, H. Häcker, D. Klebelsberg, and U. Seydel. A Comparison of the Factor Loading Matrices of Two Driver Behavior Investigations. Accident Analysis and Prevention, Vol. 7, 1975, pp. 161-178.
2. T. W. Forbes, R. V. Nolan, F. L. Schmidt, F. E. Vanosdall, D. L. Smith, G. S. Burnett, J. W. Counts, W. H. Greenwood, F. J. Graber, R. H. Johnson, J. W. Lounsburg, J. E. Schlick, and F. W. Reuter. Driver Performance Measurement Research. Department of Psychology and Highway Traffic Safety Center, Michigan State Univ.; Final Rept., National Highway Traffic Safety Administration, 1973.
3. T. W. Forbes, R. O. Nolan, F. L. Schmidt, and F. E. Vanosdall. Driver Performance Measurement Based on Dynamic Driver Behavior Patterns in Rural, Urban, Suburban, and Freeway Traffic. Accident Analysis and Prevention, Vol. 7, 1975, pp. 257-280.
4. Safe Performance Curriculum. Human Resources Research Organization, Alexandria, Va., 2 Vols., Final Rev., no date.
5. M. H. Jones. The California Driver Training Evaluation Study. School of Engineering and Applied Science, Univ. of California, Los Angeles, 1973.
6. F. McGlade. Test Driving Performance: Developmental and Validation Techniques. Highway Research News, No. 5, 1963, pp. 13-22.
7. R. R. Mourant and T. H. Rockwell. Strategies of Visual Search by Novice and Experienced Drivers. Human Factors, Vol. 14, 1972, pp. 325-335.
8. S. W. Quenault and C. F. Harvey. Convicted and Nonconvicted Drivers: Values of Drive Indices. Road Research Laboratory, Crowthorne, England, Rept. LR395, 1971.
9. The Development and Validation of Attitude, Knowledge, and Performance Tests for Evaluating Driver Education Curricula. Ohio Department of Education, Columbus, 1973.

Publication of this paper sponsored by Committee on Driver Education.