The Written Driver-Licensing Examination as a Technique in Driver Selection

JOHN A. CONLEY and WARREN J. HUFFMAN, Safety and Driver Education Laboratory, University of Illinois

•PUBLIC LAW 89-564, enacted by Congress on September 9, 1966, and known as the Highway Safety Act, has great significance for the State of Illinois and its motorists. It states (5):

> Each State shall have a highway safety program approved by the Secretary, designed to reduce traffic accidents and deaths, injuries, and property damage resulting therefrom. Such programs shall be in accordance with uniform standards promulgated by the Secretary. Such uniform standards shall be expressed in terms of performance criteria. Such uniform standards shall be promulgated by the Secretary so as to improve driver performance (including, but not limited to, driver education, driver testing to determine proficiency to operate motor vehicles, driver examinations, . . . and driver licensing) . . . .

In February 1967, tentative standards for driver licensing were proposed by the National Highway Safety Agency (4). These standards require that drivers be reexamined at least every 4 years for visual acuity and knowledge of signs, signals, and laws of the road, and that the examination on the laws of the road be written. Although the written examination is only one part of the initial driver selection, it is a major part of the reexamination. Thus, it must be capable of discriminating between an individual who knows and understands the "Illinois Rules of the Road," and one who has made only a cursory examination of it or who has not read it or who does not understand it.

In Illinois, as in most other states, the written examination has never undergone a rigorous scientific analysis to determine whether the questions are valid, are reliable, or discriminate between an individual who knows and understands the "Illinois Rules of the Road" and one who does not. Perhaps an individual has been granted a driver's license despite his lack of knowledge regarding driving rules because the instrument or examination questionnaire did not measure what it was expected to measure. The items or questions may have been too few and too easy to represent a fair test of comprehension of the material.

Thus, it was not sufficient merely to construct a new examination to test the knowledge of a new amount of material. Sufficient statistical treatment had to accompany this task so that the examination could be revised on the basis of scientific analyses to become an improved instrument capable of accomplishing the function designated to it. Also, unlike its predecessor, the new examination had to be designed so that follow-up analyses were possible.

The only other study available in this area was done at the University of North Carolina in 1959 by Campbell, who pointed out the need for research in test analysis (1): "There are few jurisdictions in which driver license examinations are subject to analysis. As long as no systematic program of test analysis is in effect, driver license tests will likely remain static at their present level, which is often inadequate." He further supports the idea that today an unreliable written test is an unnecessary inadequacy in any licensing program because specific methods of improving test reliability are available.

Paper sponsored by Committee on Road User Characteristics.

Campbell's report is not an experimental study of an actual attempt to improve a written driver-licensing examination, although it contains some information found in a small pilot study. It is rather a message urging states to use a method of item analysis to improve their tests. Most of his report deals with the purposes and methodology behind such a study, and, as such, it will be referred to again in later sections. The study reported in this paper is a pioneering effort to put into effect those practices advocated by Campbell, and is probably a unique study in the United States today. The lack of literature in this area indicates that studies of this type either have not yet been attempted or have not been reported and published.

PURPOSE OF STUDY

Highway accidents, injuries, and fatalities are increasing, and apparently disrespect for traffic laws and regulations is also increasing. In the past decade, particularly, many attempts have been to improve the safety of the highways and the motor vehicles, to improve and increase instruction in driver education, to improve law enforcement, and to provide additional legislative support for increased safety on the nation's highways. However, little has been done to improve the driver. Driving a motor vehicle is thought by many to be a right instead of a privilege, to be granted to them irrespective of their ability to drive safely and efficiently or of their knowledge of the rules for safe driving. The computer, which has been put to work in nearly every facet of American life, has not been used in driver licensing for analysis purposes! There have been reforms in basic curriculum materials, but there have been no accompanying analyses. This study is an attempt to overcome this deficiency and to obtain for Illinois an improved written driver-licensing examination that may serve as a pioneer effort toward uniform testing in the United States.

The general objective of this study was to obtain a valid and reliable instrument capable of testing a person's knowledge and understanding of the material in the "Illinois Rules of the Road" issued in December 1967, and to include sufficient statistical analyses and resulting revisions to provide a sound basis for future revisions of the "Illinois Rules of the Road" by indicating those areas that are most often misunderstood. Future analyses of the data obtained in this study could lead to the discovery of areas where special emphasis needs to be given to particular subgroups of the driving population according to age and sex. This could provide an excellent educational base for further emphasis in driver education, refresher courses, and driver-improvement programs in the state. The data will also allow a further study of the records of those taking high school driver-education courses, those taking commercial driver-education courses, and those taking no formal driver-education courses of any kind. Other items such as the number of years of completed education, rural or urban place of residence, and the applicant's expectation on the examination will provide additional research data that could be used as a general educational instrument for all drivers in the state. This, in turn, may improve the driving record of the state.

DESIGN OF THE STUDY

Limitations of the Study

To construct and analyze a new written driver-licensing examination required that a number of factors be considered in planning the research design. Rarely, if ever, is it possible to structure a statewide setting in such a manner as to satisfy fully the requirements of the classical experimental approach. For example, the researcher is restricted in his sample because he cannot wait to take a completely random sample from the applicants of a whole year. In order to stay within a reasonable time schedule, he must take the sample from a sample of the population that takes the test on certain days during the course of the year rather than from the population as a whole. Moreover, to secure cooperation of officials of the Driver License Division of the Office of the Secretary of State, he must conduct the research within the framework of existing work loads, schedules, and routines. As many variables as possible were controlled by the experimental design within reasonable limits of manageability.

The applicants included represent one variable. Because selection was partially based on the desire and availability of the individual to take the examination during the weeks used to collect data, there is no way to be sure that the applicants had the same basic abilities as those who chose to take their examinations at a different time. Although the selections were random as far as the total available population of test scores was concerned, there may have been bias within the population because of the abilities of the persons available to take the examination at those times of the year. For example, some high school driver-education classes may not have been completed in time for these students to be included among the applicants.

Because a double-blind technique could not be applied, the Hawthorne effect was still present. Also, there was no control concerning the number of attempts an applicant had previously made to pass the examination. Thus, he may have received the same form of the examination a second time not only by chance but also by necessity.

Other limitations are based on the nature of the study. More complete analyses as well as a longer study covering several years were not possible because of the time and money available. Item analyses according to age, sex, educational level attained, place of residence, driving experience, and driver education source would have given a more complete picture of the examination. Also, a long-term study comparing examination results with accident and violation reports would be valuable.

Application of the results of this study is limited to the "Illinois Rules of the Road" as of December 1967 and, of course, to the State of Illinois itself.

Fundamental Assumptions of the Study

Because it is impossible in a study of this type to control all the variables involved, the following assumptions concerning the information and administration of the testing instrument had to be made:

1. That the questions used were a valid sample of the "Illinois Rules of the Road";
2. That the weighting of the questions by the board of experts was valid according to the criteria used;
3. That all forms of the test were parallel forms;
4. That the tests were administered fairly and equally to all applicants;
5. That all applicants were given sufficient time to complete the test without undue time pressure;
6. That the applicants for the drivers' licenses had at least an eighth-grade reading skill;
7. That the sample of tests was a valid one; and
8. That the scoring of the tests was accurate.

Test Instrument

The written driver-licensing examination used by the State of Illinois before this study consisted of 3 forms, each having 20 true-false questions and multiple-choice items with 3 choices per item. These tests have been severely criticized as being too easy and nondiscriminating. An analysis of 464 of these tests revealed the following:

1. Nearly 87 percent of the applicants passed the test.
2. An 18 percent scoring error existed in marking the test.
3. Of the 20 items, 8 were passed by over 95 percent of the applicants.
4. Only one item was passed by less than 80 percent of the applicants.
5. No item met the suggested 70-30 pass-fail discrimination criteria.
6. For 4 of the 20 items, only 1 of the 3 choices was chosen by 2 percent of the applicants, thus indicating that the answer was too obvious.
7. An additional 12 items degenerated to mere true-false questions because one of the choices was not being selected by at least 2 percent of the applicants.
8. Of the 20 questions, 13 failed the minimal point biserial discrimination index of 0.30.
9. Only 3 items out of the 20 were considered to be good questions based on the suggested criteria of this study.

Some of the 20 questions were repeated on 2 or all 3 of the forms of the test. Each question was given equal weight in the total score. In general, the old test instrument, although in use for many years, was recognized as an inadequate instrument. The new test instrument attempted to overcome these difficulties. Five parallel forms of the examination were constructed with no overlap of questions so that applicants who were required to take the examination more than once in order to obtain their drivers' licenses would have a relatively small chance of repeating the same examination. Also, reexamination of all drivers at least every 4 years is required by the federal government and at least every 9 years by the State of Illinois, and the use of 5 forms of the examination instead of 3, as previously used, makes the reexamination more valid. The form used during the first data collection period of the study had 50 multiple-choice items with 4 choices per item. The form used in the second data collection period had 35, and the final form had only 30 items.

The original 250 test items were constructed so that each area in the "Illinois Rules of the Road" (December 1967)from which the questions were taken was represented in proportion to the amount of information it contained. Each of the 5 forms was constructed to continue this proportionality. Following the analysis of the data collected during the first period, the forms were revised to contain 35 questions, each determined to be the best by means of the statistical treatment described later. As they were finally revised, each of the 5 forms consisted of 30 multiple-choice questions with 4 choices per question.

Each form of the examination was subjected to a readability test, described later, so that the reading ease of the questions was equivalent to that expected of someone with a ninth- or tenth-grade education. This level of reading difficulty was chosen because almost 82 percent of the adult population of Illinois has completed grade eight, and this percentage is increasing steadily (3). The completion of a grade in school, although it does not mean an equivalent reading skill, is the best indicator available of an individual's ability to read and comprehend printed matter.

In addition to the scoring questions, there were 14 nonscoring questions that obtained demographic information, and were also used as criteria in the item analysis, described later, in the attempt to find an acceptable external criterion. The 14 questions related to the following: date of birth; sex; number of years of formal education; college degree obtained; number of years of driving experience; expected test score; place of residence, rural or urban; average number of miles driven per year; rating of driving skill; where previously licensed, if so; high school driver-education course completed; commercial driver-education course completed; number of previous attempts to obtain Illinois driver's license; and driver-improvement course attended.

Selection of Subjects

The subjects for the study were chosen at random from all those who applied for Illinois drivers' licenses during the 2 periods of data collection. Only those who could not read and, therefore, were given an oral examination and those who took an unusually long time to complete the examination were automatically exempt from selection. Because of large volumes at certain times, some of the large stations could not give the examination to every applicant. Also, because of a variety of problems, the 3 Chicago examination stations were not included in the sampling.

Each examination station (Fig. 1) was assigned a quota for each test form based on the station's proportional share of all the examinations given in the state during 1965, 1966, and 1967. It was originally intended that 10,000 answer sheets, 2,000 for each form, be obtained in each testing period. Only 6,773 were collected for the first period, however, and 7,022 for the second period, because of the lack of data from Chicago, where 2,970 answer sheets were expected, and the inevitable loss of some answer sheets because of coding difficulties, incompleteness of answers, or missing data from some of the stations. This represents a net retrieval downstate of 96.34 percent for the first period and 99.89 percent for the second period.

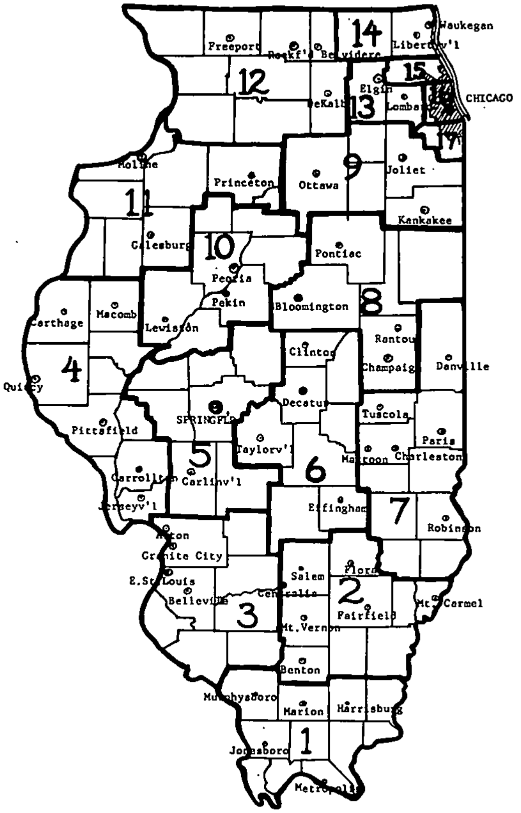

Figure 1. Driver-licensing examination stations where data were collected during two periods.

Data Collection Process

First Period—The preliminary first draft of 5 forms of the written driver-licensing examination was completed in January 1968. These forms were then distributed to a committee for revision and criticism in regard to content and structure. Final revisions were completed and the forms sent to the Driver License Division of the Office of the Secretary of State for publishing and distribution at the end of April 1968. On June 6, 1968, the question and answer sheets for all 5 test forms were distributed to supervisors in all 17 examination districts in Illinois. Specific instructions were given to the supervisors at this time regarding the techniques to be used in the collection of the data. The supervisors, in turn, distributed the allotted number of the examinations to examiners in 59 stations throughout Illinois. This examination was given to applicants for Illinois drivers' licenses. The collection of the data took place during the period of June 17 to July 31, 1968. The time required for each station to collect its share of the data varied from 1 to 6 weeks. The examination answer sheets were sorted at each station by test form number and sent on completion to the Driver License Division in Springfield, Illinois. The authors then collected all the answer sheets of examinations written in downstate Illinois during the testing period and checked them to see if each examination station had turned in its allotted number of answer sheets for each of the 5 forms. Then the answer sheets were scanned electronically for completeness of data or duplication. The rejected answer sheets were cleaned and coded as completely as possible by the investigator and then resubmitted for machine scoring. The subsamples for each form of the examination were then subjected to an item analysis, to a key selector program to determine the effectiveness of selected external criteria, and to multiple-regression computations by computer as a basis for revision of the test items and item choices.

Second Period—Each of the 5 forms was revised and then distributed to the district supervisors on August 29, 1968, and the entire procedure just described was repeated during the second data collection period from September through November, 1968.

Statistical Treatment

The investigator or an assistant hand-checked each answer sheet that was rejected by the electronic scanner for reasons of incomplete identification, incomplete demographic data, incomplete question responses, multiple marking of answers, extraneous marks, or requirement of special coding. Answer sheets collected during the second period were all hand-coded for the examination station source. The answers were put on Digitek answer sheets and transferred automatically to keypunch cards, thus avoiding errors. The cards were then sorted by test form number in preparation for analysis.

An item analysis program, devised by the Measurement and Research Division of the University of Illinois, produced the following information for each test form:

1. Total raw score of each subject;
2. Mean score;
3. Standard error and standard deviation;
4. Proportion passing each item;
5. Proportion selecting each response for each item;
6. Items that are extremely easy or extremely difficult, i.e., over 90 percent passing or less than 10 percent passing;
7. Kuder-Richardson test of reliability at three levels: 14, 20, 21;
8. Coefficient of discrimination;
9. Validity coefficient;
10. Distribution of total test form scores and the total score;
11. Point biserial correlation coefficient for each item;
12. Plot of reliability and validity indexes for each question; and
13. Point biserial correlation coefficent for each demographic criterion and the total score.

The information obtained by the item analysis was used to revise the 5 forms of the examination based on the overall picture of an item according to the following criteria:

1. There were no illegal responses to an item, i.e., multiple answers;
2. The proportion passing as many items as possible was between 40 and 70 percent;
3. At least 2 percent of the subjects selected each of the item choices;
4. Any items designated too easy or too difficult were reviewed and replaced or revised;
5. The coefficient of discrimination was about 0.96; and
6. Item score-test score point biserial correlation coefficients for each item were at least 0.30 whenever possible.

The item analysis program used has the added advantage of initially testing each of the external or demographic criteria by means of a total score-criterion score point biserial correlation to see if any other criterion besides total score is suitable for the item analysis. It then automatically does the total test score-item score analysis. If another criterion is indicated as being suitable as the item analysis base, the data can then be resubmitted on that basis. This program has the advantage of indicating the highest possible criterion to use in the item analysis.

In order to obtain another measure of the effect of the criteria on total test score variance, a multiple-regression program was used. This program chooses the criterion that accounts for the highest proportion of variance among the mean scores of each form and then accumulates each of the remaining unique variations so that the total amount of variation accounted for by all criteria, as well as each individual effect, is calculated.

ANALYSIS AND INTERPRETATION OF DATA COLLECTED DURING FIRST PERIOD

Intercorrelations Between Criterion Scores and Total Test Scores

Although the item score-total test score correlation technique is the one commonly used in item analysis, it only measures the internal consistency or homogeneity of the items. Because it is advantageous to use an external criterion to validate the test, a key selector program was first employed to develop the best items to maximize the predictability of the criterion. It compared 14 different predictors (the demographic criteria) on the basis of point biserial correlation coefficients for each of the 5 test forms. A point biserial correlation of at least 0.300 was arbitrarily chosen as the requirement for the use of any external criteria. No external criterion met this standard. The criteria with the most consistent point biserial differentials include years of education, college degree, licensed in another state, and high school driver-education course.

A further attempt was made to go beyond the key selector program of criterion score intercorrelations and to ascertain the effect of each criterion separately and as a cumulated whole. The product moment correlations and calculated multiple correlations were used to give the multiple r of all 14 independent variables predicting the dependent variable of total score. The cumulated effect of all 14 criteria was 0.39233998, 0.4322603, 0.3569309, 0.3554583, and 0.4138793 for Forms 1 through 5 respectively. Thus less than 20 percent of the score variance can be accounted for in any of the test forms by the criteria selected.

This result further corroborated the conclusions derived from the key selector program. However, it was decided to reword some of the criteria and to revise the format of the answer sheet for the second data collection period in another attempt to find at least one legitimate external criterion.

Because a suitable external criterion was not found, the total test scores were used in the item analysis in order to determine the questions with the greatest degree of discrimination, difficulty level, and suitability of foils.

TABLE 1

ITEM ANALYSIS BASED ON TOTAL SCORE CRITERION FOR FIVE FORMS USED DURING FIRST PERIOD

Question	Form I			Form II			Form III			Form IV			Form V		
	PBC	PSP	FNS	PBC	PSP	FNS	PBC	PSP	FNS	PBC	PSP	FNS	PBC	PSP	FNS
1	0.219	0.944	2	0.193	0.823	1	0.118	0.522		0.252	0.870		0.260	0.674	
2	0.274	0.423		0.247	0.835		0.203	0.601		0.278	0.345		0.032	0.503	
3	0.284	0.449		0.329	0.664		0.356	0.549		0.207	0.650		0.360	0.533	
4	0.307	0.317		0.294	0.924	1	0.324	0.448		0.408	0.796		0.406	0.807	
5	0.368	0.653		0.273	0.526		0.271	0.589		0.184	0.668		0.167	0.241	
6	0.249	0.246		0.067	0.612		0.176	0.372		0.323	0.540		0.221	0.461	
7	0.327	0.656		0.381	0.662		0.086	0.653	1	0.177	0.536		0.527	0.903	
8	0.397	0.531		0.228	0.303		0.315	0.541		0.333	0.647	1	0.212	0.346	
9	0.349	0.873	1	0.199	0.376		0.271	0.336		0.364	0.916	1	0.173	0.406	
10	0.379	0.938	2	0.377	0.791		0.356	0.851		0.328	0.795	1	0.609	0.905	
11	0.182	0.521		0.324	0.790	1	0.418	0.708	1	0.370	0.811	1	0.437	0.755	
12	0.454	0.691		0.094	0.330		0.363	0.821		0.461	0.859		0.482	0.776	
13	0.097	0.129		0.187	0.224		0.280	0.599		0.359	0.698		0.376	0.700	1
14	0.350	0.366		0.304	0.689		0.288	0.564		0.376	0.841	1	0.324	0.489	
15	0.263	0.338		-0.020	0.144		0.330	0.503		0.305	0.541		0.494	0.839	
16	0.350	0.770	1	0.383	0.567	1	0.377	0.886	1	0.242	0.619		0.243	0.359	
17	0.291	0.699		0.364	0.833	1	0.416	0.848		0.451	0.781		0.088	0.416	
18	0.344	0.882		0.407	0.788		0.457	0.736		0.456	0.783		0.335	0.637	
19	0.343	0.375		0.195	0.300		0.317	0.525		0.209	0.545		0.414	0.735	
20	0.380	0.889	1	0.273	0.629		0.113	0.226		0.355	0.567		0.265	0.623	
21	0.420	0.813	1	0.289	0.379		-0.110	0.039		0.099	0.457		0.132	0.677	
22	0.374	0.422		0.173	0.311		0.382	0.779	1	0.473	0.865		0.389	0.627	
23	0.355	0.559		0.126	0.251		0.322	0.722		0.267	0.423		0.396	0.661	
24	0.256	0.919	1	0.348	0.464		0.218	0.331		0.284	0.829	1	0.221	0.599	
25	0.308	0.285		0.123	0.159		0.278	0.341		0.350	0.653		0.487	0.727	
26	0.213	0.243		0.422	0.698		0.298	0.542		0.205	0.460		0.488	0.887	
27	0.341	0.803		0.323	0.901	1	0.411	0.810	1	0.457	0.883		0.523	0.864	
28	0.451	0.852	1	0.319	0.423		0.340	0.675		0.225	0.482		0.550	0.891	
29	-0.012	0.244		0.303	0.780		0.408	0.799		0.393	0.452		0.516	0.855	1
30	0.349	0.682		0.383	0.779		0.430	0.776		0.441	0.828		0.489	0.818	1
31	0.475	0.605		0.425	0.894	1	0.383	0.732		0.381	0.760	1	0.226	0.642	
32	0.203	0.234		0.349	0.668		0.412	0.692		0.448	0.676		0.220	0.358	
33	0.356	0.568		0.367	0.463		-0.086	0.227		0.200	0.518		0.497	0.719	
34	0.249	0.187		0.276	0.836	1	0.219	0.423		0.033	0.615	1	0.059	0.475	
35	0.155	0.408		0.125	0.315		0.302	0.406		0.171	0.342		0.193	0.320	
36	0.070	0.219		0.467	0.829		0.301	0.749		0.303	0.689		0.201	0.275	
37	0.420	0.792	1	0.249	0.300		0.315	0.533		0.399	0.516		0.354	0.702	
38	0.451	0.648		0.070	0.218		0.494	0.465		0.100	0.299		0.571	0.869	
39	0.393	0.673	1	0.309	0.560		0.422	0.418		0.560	0.787		0.462	0.814	
40	0.246	0.299		0.340	0.251		0.372	0.581		0.313	0.526		0.353	0.654	
41	0.227	0.295		0.371	0.808	1	0.411	0.472		0.127	0.322		0.485	0.848	
42	0.338	0.302		0.159	0.200		0.377	0.757		0.442	0.574		0.500	0.734	
43	0.377	0.907	1	0.339	0.553		0.383	0.590		0.234	0.263		0.378	0.537	
44	0.471	0.827		0.276	0.792		0.364	0.755	2	0.343	0.530		0.412	0.718	
45	0.400	0.444		0.225	0.251		0.215	0.275		0.336	0.833	1	0.131	0.382	
46	0.343	0.424		0.178	0.462		0.358	0.567		0.384	0.518		0.300	0.491	
47	0.252	0.436		0.323	0.658		0.163	0.178		0.267	0.366		0.295	0.469	
48	0.135	0.325		-0.044	0.037		-0.135	0.144		0.370	0.649		0.264	0.492	
49	0.273	0.813	1	0.355	0.569	2	0.343	0.446		0.347	0.898		0.493	0.810	
50	0.270	0.828		0.330	0.368	1	0.247	0.568		0.376	0.452		0.159	0.280	

Note: PBC is the point biserial coefficient of correlation, PSP is the proportion of subjects passing each item, and FNS is the number of foils for each item not meeting the predesignated standard.

Item Analysis

Table 1 gives the results of the item analysis for the 5 test forms based on a total test score criterion. In order to select items for the test forms for use during the second period, each item was judged on the point biserial coefficient of correlation and on the proportion of subjects passing. The items were placed in rank order and the top 35 questions were chosen on the basis of their point biserial coefficient. The acceptable range for the proportion passing an item was 40 to 70 percent. The items chosen for each form are shown in Figures 2 through 6.

Although the arbitrary standards of a 0.300 point biserial correlation coefficient and a 40 to 70 percent proportion passing range was hoped for, it was not possible to find sufficient questions that strictly met both standards. Therefore additional questions were chosen that first met the standard of at least a 0.300 point biserial coefficient. Next, questions were chosen that met the standard of between 40 and 70 percent passing. Where necessary questions were then selected that met neither standard but were as close as possible. An average of 6 questions for each form did not meet the point biserial coefficient standard; the lowest figure used was 0.232. An average of 19 questions for each form fell outside the proportion passing range of 40 to 70 percent; the lowest figure used was 25 percent and the highest was 94 percent. Many of the upper level proportion passing figures are inherently tied in with high point biserial correlation coefficients. This technique was considered acceptable because the original standards were very strict, and many other studies use acceptable point biserial coefficients as low as 0.200 and proportion passing ranges from 20 to 80 percent. Table 2 gives the results of this selection process.

In addition, if an item was chosen that had a foil that failed to have at least 2 percent of the subjects choose it, either it was replaced in its entirety or the affected foil was revised. This entailed 10 questions for Forms I and II, 5 questions for Form III, 8 questions for Form IV, and 3 questions for Form V.

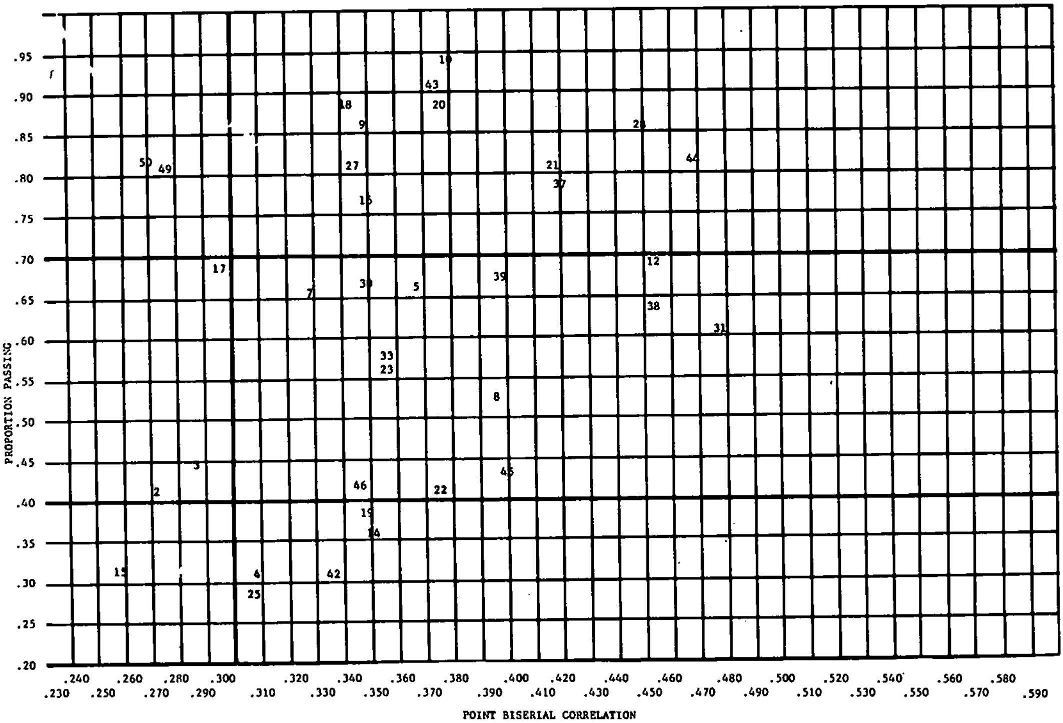

Figure 2. Distribution of questions for Form I used in second data collection period.

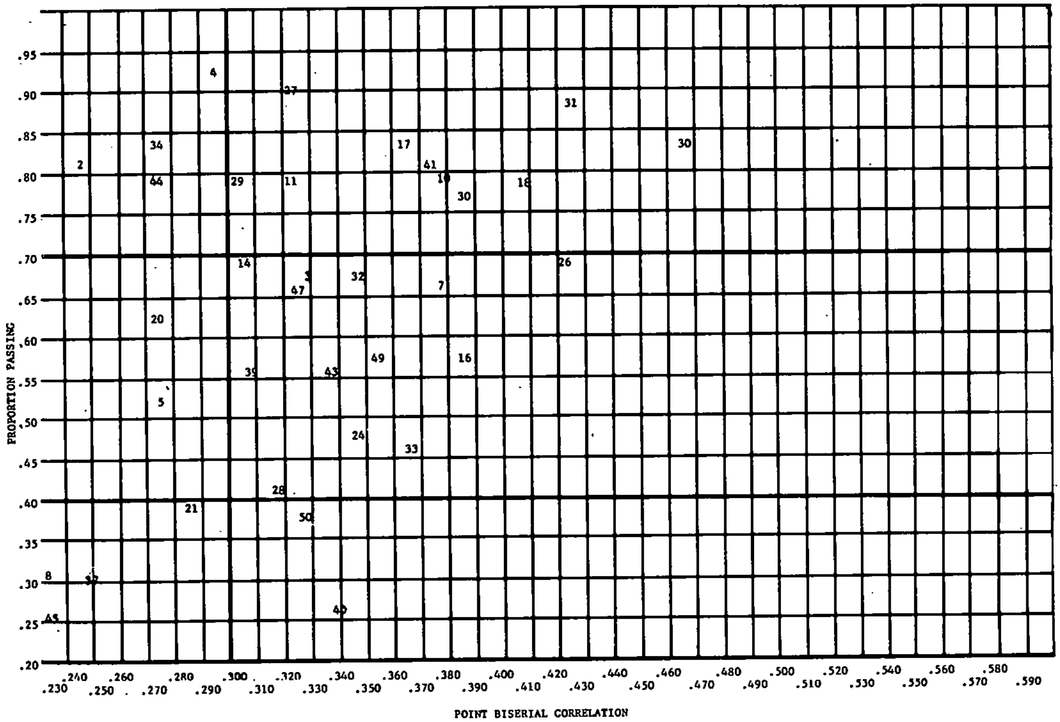

Figure 3. Distribution of questions for Form II used in second data collection period.

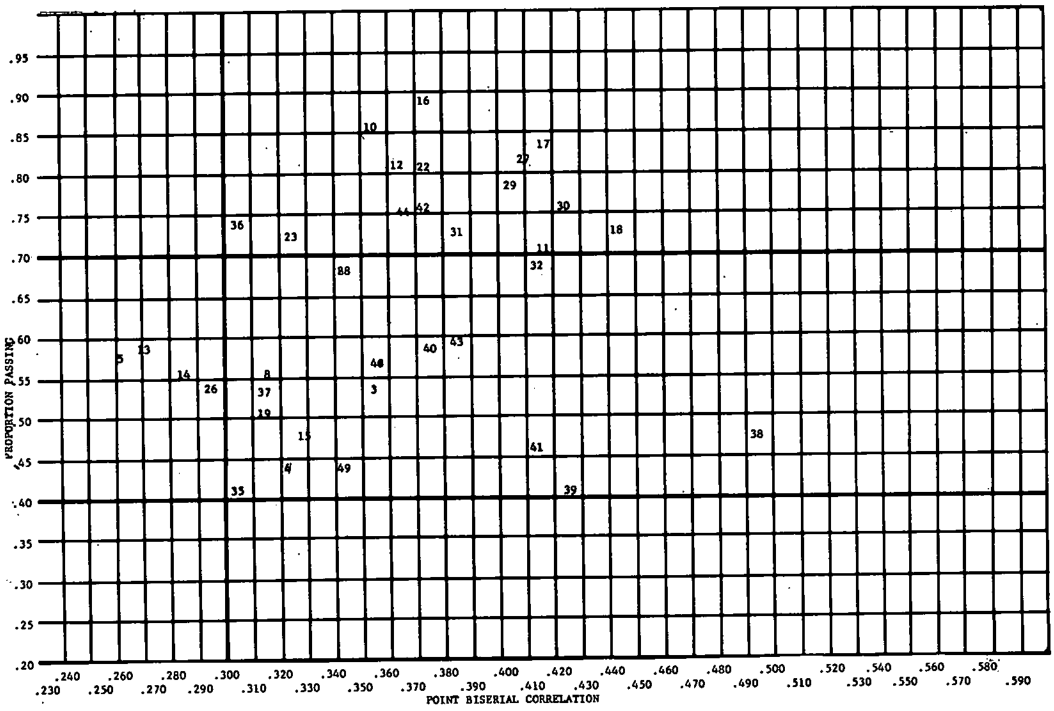

Figure 4. Distribution of questions for Form III used in second data collection period.

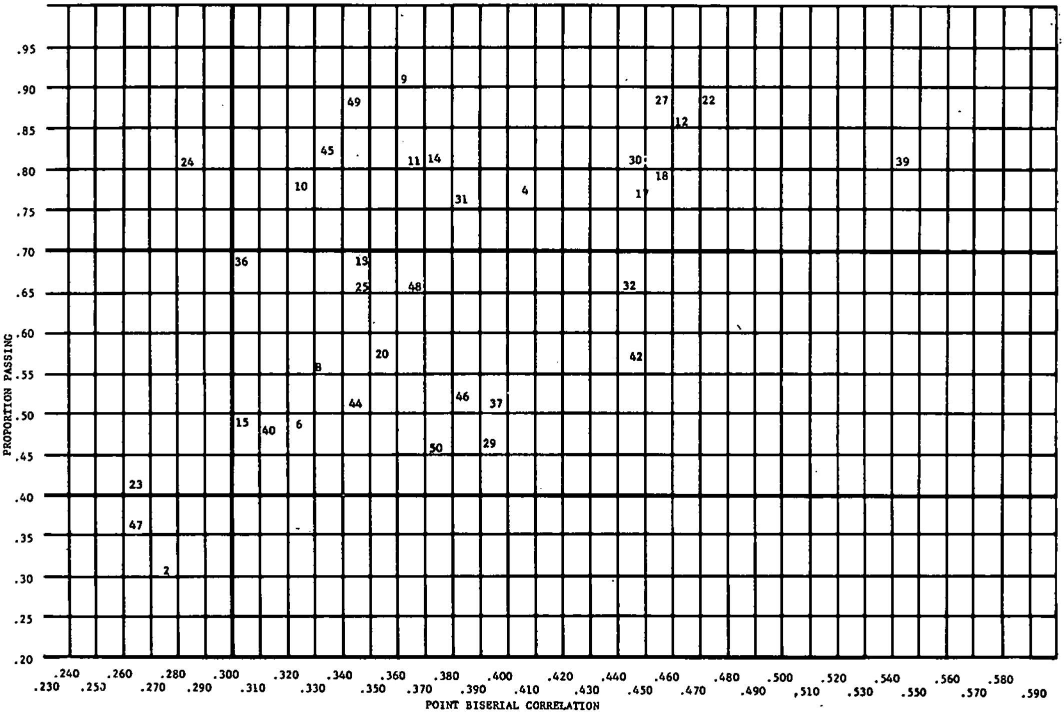

Figure 5. Distribution of questions for Form IV used in second data collection period.

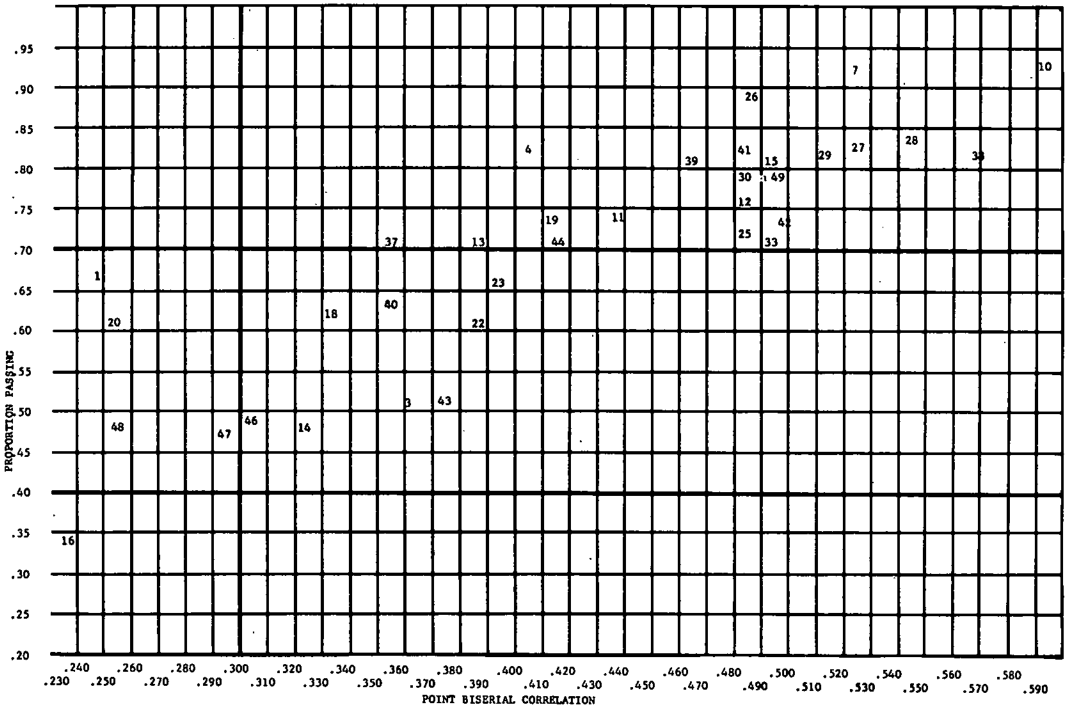

Figure 6. Distribution of questions for Form V used in second data collection period.

TABLE 2

NUMBER OF ITEMS MEETING SELECTION STANDARDS FOR INCLUSION IN TEST FORMS FOR SECOND PERIOD

Form	Biserial Coefficient Standard	Proportion Passing Standard	Both Standards	Neither Standard
I	16	3	13	3
II	12	2	13	8
III	15	4	16	0
IV	15	1	16	3
V	22	4	8	1

TABLE 3

MEAN SCORES, STANDARD DEVIATIONS, AND NUMBER OF SUBJECTS FOR FIRST PERIOD

Form	Mean Score	Standard Deviation	No. of Subjects
I	27.75	6.59	1,391
II	26.99	5.93	1,390
III	27.59	6.90	1,369
IV	30.03	6.70	1,329
V	28.13	6.72	1,314

Note: Total score possible was 50.

Statistical Data of Total Scores

The mean scores, standard deviations, and number of subjects for each form of the examination are given in Table 3.

Although 10,000 answer sheets, 2,000 for each of the 5 forms, were anticipated to make up the total number of subjects, only 6,793 valid answer sheets were returned. No returns were included from the 3 Chicago examination stations. Of those from the downstate examination stations less than 2 percent were incomplete. After necessary coding and cleaning operations were performed on the answer sheets, 96.5 percent of all downstate answer sheets were returned and were acceptable for statistical treatment. The differences shown in the numbers for each test form are caused by the incomplete returns on Form IV from DeKalb and on Form V from DeKalb and Freeport, and by some answer sheet rejections because of failure to complete the examination or to fill in at least part of the criterion data. The total number of possible returns for each form from the 56 downstate stations was 1,406, had each returned its allotted number of answer sheets properly filled out.

ANALYSIS AND INTERPRETATION OF DATA COLLECTED DURING SECOND PERIOD

Intercorrelations Between Criterion Scores and Total Test Scores

Although the search for an external criterion using the arbitrary standard of a 0.300 point biserial correlation coefficient was futile for the first data collection period, another attempt was made with the data collected during the second period because it was thought that the revision of the test instrument and the answer sheet might have a beneficial effect.

The same statistical program with the same point biserial correlation coefficient was used as in the first period. None of the criteria met the desired standard of 0.300 point biserial coefficient. It is of interest though to note some of the trends of each of the 14 demographic criteria used. Although only the criterion based on the applicant being previously licensed in another state showed point biserial coefficients approaching the required 0.300 (0.383, 0.289, 0.341, 0.294, 0.254 for Forms I through V respectively), certain patterns of consistency may be indicative of a need for further study.

The criterion of age based on date of birth indicated consistent negative correlations for those born from 1860 to 1929 and those born since 1950. The best positive correlations were for those drivers born in the 1930's and 1940's and who are thus between the ages of 19 and 38.

The point biserial correlation coefficients for male applicants were consistently higher than those for females.

Education level proved to be very inconsistent. Those with 0 to 1 year of formal education completed consistently had negative correlations. For those with 2 to 7 or 8 years of education, the correlation was positive and generally increasing in size. For those with 8 to 10 years of education, there was a consistent negative correlation with total test score. The highest correlation occurred for those with 11 years of education, perhaps indicating the influence of the high school driver-education courses usually given during the ninth or tenth grade. The coefficients remained positive but smaller in magnitude until the 17- through 19-year level, at which point they turned negative again. This would indicate that those with graduate degrees scored poorly compared with college undergraduates. In conjunction with this, those who had completed college scored consistently better than those who had not obtained a college degree.

Driving experience had a marked trend, related both to the number of years of experience and to the annual mileage driven. Those who had been driving from 2 to 39 years scored best. Those who had driven only 1 year or who had yet to drive and those who had driven 40 or more years, with few exceptions, had negative correlations between the criterion score and the total test score, thus indicating lower test scores. This is the same pattern that was evident for the age criterion. Similarly, those who drove infrequently, less than 2,500 miles per year, and those who drove over 50,000 miles per year scored poorly in comparison with the moderate and average drivers.

When asked to indicate the score they expected to achieve on the examination out of a possible total of 35, those who predicted scores higher than 24 had positive correlations and those who predicted scores lower had negative correlations ranging from -0.010 to -0.183. The same trend in sign and magnitude was exemplified for people who rated their driving skills. Those who thought that their driving skills were above average, exceptional, or in the top 10 percent in the state tended to have low positive correlations. Those who lacked confidence in their driving skills had negative correlations. The meaningfulness of negative correlations is debatable, especially when they are as low as the ones shown in this analysis. Although the magnitude is more important than the sign of the correlations, the consistency shown in these variables may be noteworthy. Certainly for use in predicting, these correlations are meaningless.

The external or demographic criterion with the best possibilities to be used to validate the test is the one indicating that the applicant had previously been licensed in another state. As indicated earlier, these people consistently showed relatively high positive correlation coefficients between the criterion score and total test score. This may indicate some carry-over value in the training, experience gathered while driving in another state, or better studying of the "Illinois Rules of the Road" because of apprehensiveness about being an out-of-stater. Another possible explanation may be that they had not as yet been contaminated with misinformation about the easiness of the written examination.

Those who had attempted the examination for drivers' licenses before scored consistently poorer than those who were attempting to obtain their licenses for the first time. Although there was some difficulty in the interpretation of this criterion, it seems to indicate that repeaters, in the field of driver licensing or elsewhere, have poorer knowledge than the general population.

Those applicants who had taken high school driver-education courses consistently scored better than those who had not. This was not the case for those who had completed training in commercial driver-education courses. Here the pattern was inconsistent as 3 out of the 5 forms showed a negative correlation between taking the course and total test score.

In 4 out of 5 forms, drivers who had been required to attend driver-improvement courses scored worse than those who had not attended such courses. This is to be expected because those required to take such courses usually have poor driving records or have had their licenses suspended or revoked for some offense. These people may tend to have a poorer attitude toward the importance of learning and understanding the materials in the "Rules of the Road" or they may lack the ability to learn such material.

The place of residence consistently showed that urban dwellers scored higher than rural dwellers. Undoubtedly the experience gained while driving in diverse urban settings has a beneficial effect on the driver's knowledge.

Another attempt was made to ascertain the effectiveness of the external criterion scores by means of a multiple-regression technique. As with the first period data, the cumulated effects of the 14 criteria to predict the total test score and to account for the variance in the scores were low. The cumulated effect of all 14 criteria was 0.3975900, 0.4169196, 0.4359103, 0.4017046, and 0.4228875 for Forms I through V respectively. Although higher than in the case of the first period data, less than 20 percent of the score variance can be accounted for in any of the test forms by the criteria selected.

Many of these criteria have cross influences, and further study needs to be done in the area of finding a suitable external criterion to be used to further validate the examination. Because none was found in this study, the item analysis was again performed on the basis of total test score.

Item Analysis

Table 4 gives the results of the item analysis for the second period data. From the 35 items, the top 30 were selected to constitute the final form of the examination. Again it was not possible to find sufficient questions that strictly met both major standards. Therefore, additional items were chosen that first met the standard of at least a 0.300 point biserial correlation coefficient. Next, questions were chosen that met the standard of between 40 and 70 percent passing. Where necessary, questions were then selected that met neither standard but were as close as possible. On Forms I, II, and III, there were 4 items each that did not meet the point biserial correlation coefficient standard. The lowest figure used was 0.270. On Forms I and II, the proportion passing standard was lowered to 28 and 36 respectively for 3 items on each form. On all 5 forms the proportion passing standard had to be extended to include higher figures associated with the higher point biserial coefficients. The extension ranged from 85 to 92 percent and involved from 14 to 19 questions. Figures 7 through 11 show the distribution of questions for the final revision. Table 5 gives the results of this selection process.

TABLE 4

ITEM ANALYSIS BASED ON TOTAL SCORE CRITERION FOR FIVE FORMS USED DURING SECOND PERIOD

Question	Form I			Form II			Form III			Form IV			Form V		
	PBC	PSP	FNS	PBC	PSP	FNS	PBC	PSP	FNS	PBC	PSP	FNS	PBC	PSP	FNS
1	0.291	0.527		0.355	0.849	1	0.280	0.692		0.319	0.376		0.319	0.600	
2	0.288	0.430		0.398	0.662		0.265	0.563		0.513	0.797		0.284	0.574	
3	0.288	0.410		0.374	0.788		0.368	0.581		0.223	0.768		0.395	0.811	
4	0.362	0.671		0.203	0.542		0.369	0.531		0.391	0.731		0.561	0.870	
5	0.186	0.233		0.380	0.699		0.366	0.858		0.419	0.791		0.551	0.879	
6	0.373	0.565		0.161	0.290		0.376	0.703		0.389	0.784		0.421	0.761	
7	0.387	0.832	1	0.418	0.744		0.418	0.804		0.432	0.774		0.508	0.783	
8	0.399	0.914	1	0.411	0.637		0.218	0.554		0.514	0.817		0.479	0.759	
9	0.240	0.387		0.304	0.844		0.356	0.525		0.435	0.718		0.370	0.519	
10	0.232	0.291		0.393	0.602		0.263	0.518		0.486	0.898		0.495	0.786	
11	0.188	0.287		0.378	0.803		0.214	0.753		0.384	0.544		0.166	0.388	
12	0.404	0.814		0.459	0.806		0.445	0.828		0.494	0.806		0.337	0.692	
13	0.359	0.719		0.135	0.627		0.448	0.757		0.565	0.823		0.477	0.748	
14	0.343	0.897		0.322	0.354		0.280	0.500		0.420	0.651		0.227	0.534	
15	0.324	0.364		0.329	0.446		0.350	0.636		0.554	0.847		0.397	0.646	
16	0.439	0.890	1	0.453	0.657		0.387	0.742		0.316	0.498		0.367	0.672	
17	0.353	0.801		0.473	0.875		0.441	0.823		0.543	0.807		0.446	0.703	
18	0.374	0.456		0.298	0.377		0.487	0.792		0.351	0.646		0.575	0.867	1
19	0.381	0.538		0.362	0.773		0.445	0.713		0.551	0.885		0.562	0.855	
20	0.270	0.290		0.343	0.744		0.492	0.680		0.350	0.490		0.589	0.879	
21	0.398	0.791		0.422	0.871	1	0.282	0.575		0.498	0.793		0.312	0.461	
22	0.472	0.809		0.409	0.641		0.387	0.810		0.289	0.672		0.562	0.825	
23	0.369	0.687		0.392	0.454		0.350	0.710		0.465	0.635		0.518	0.724	
24	0.501	0.672		0.377	0.876	1	0.313	0.410		0.380	0.662		0.407	0.685	
25	0.342	0.569		0.517	0.764		0.275	0.782		0.379	0.515		0.584	0.861	
26	0.401	0.809		0.324	0.444		0.276	0.494		0.559	0.794		0.473	0.809	
27	0.453	0.643		0.348	0.545		0.391	0.487		0.354	0.521		0.376	0.681	
28	0.375	0.683	1	0.360	0.506		0.340	0.433		0.458	0.573		0.431	0.841	
29	0.303	0.295		0.368	0.770		0.434	0.621		0.341	0.558		0.479	0.752	
30	0.391	0.829		0.369	0.618		0.362	0.527		0.443	0.812		0.352	0.566	
31	0.476	0.820		0.321	0.758		0.493	0.769		0.442	0.636		0.469	0.753	
32	0.391	0.457		0.262	0.253		0.459	0.583		0.187	0.367		0.268	0.444	
33	0.380	0.417		0.425	0.702		0.391	0.723		0.469	0.650		0.395	0.518	
34	0.249	0.515		0.271	0.373		0.479	0.588		0.462	0.867		0.304	0.484	
35	0.385	0.790		0.287	0.372		0.374	0.559		0.383	0.452		0.500	0.826	

Note: PBC is the point biserial coefficient of correlation, PSP is the proportion of subjects passing each item, and FNS is the number of foils for each item not meeting the predesignated standard.

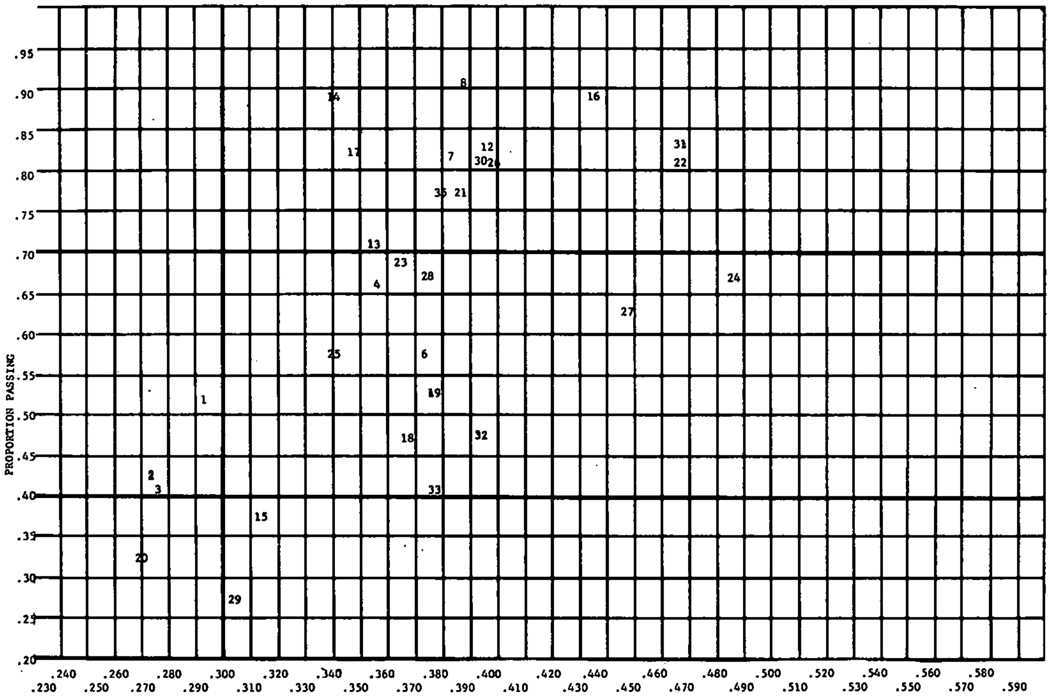

Figure 7. Distribution of questions for Form I as finally revised.

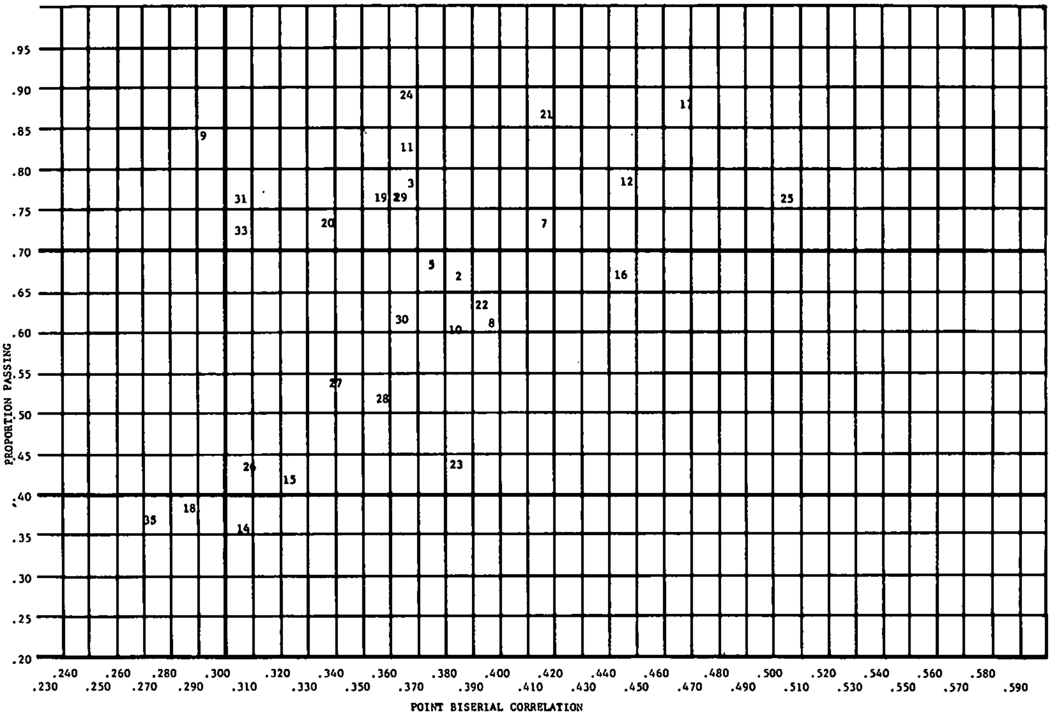

Figure 8. Distribution of questions for Form II as finally revised.

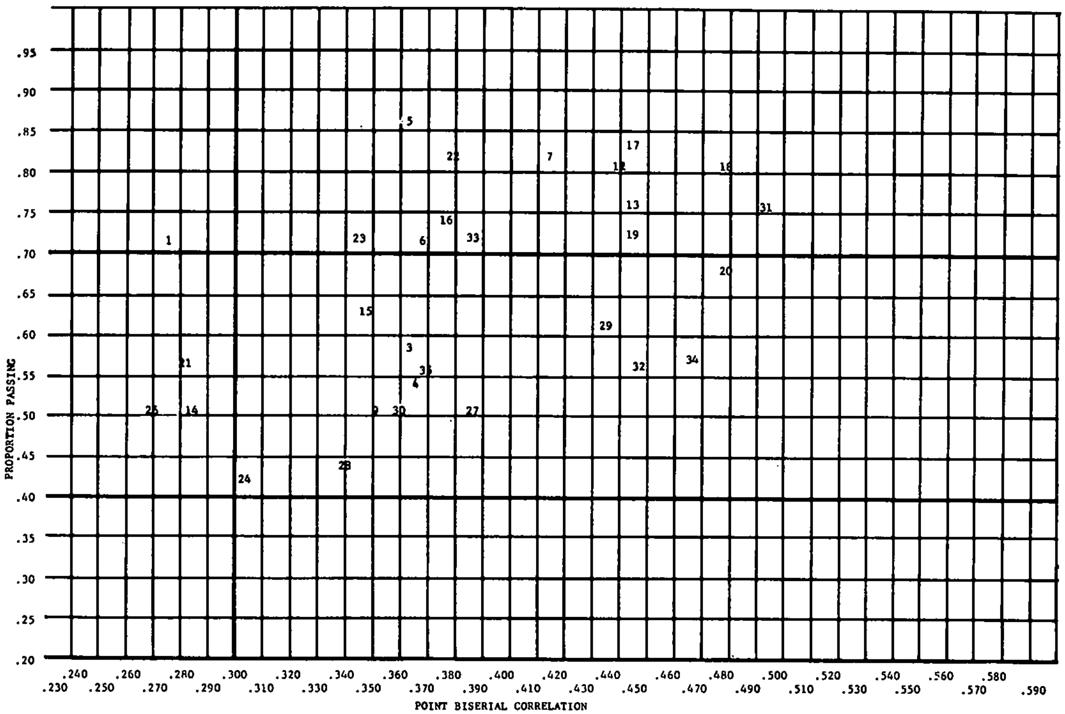

Figure 9. Distribution of questions for Form III as finally revised.

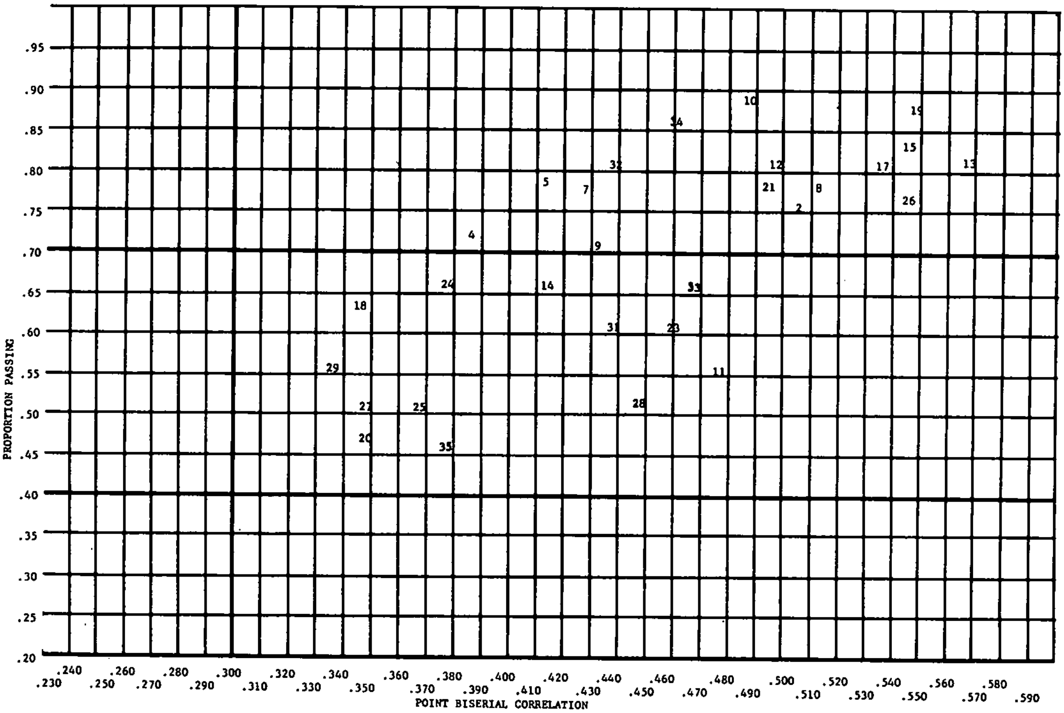

Figure 10. Distribution of questions for Form IV as finally revised.

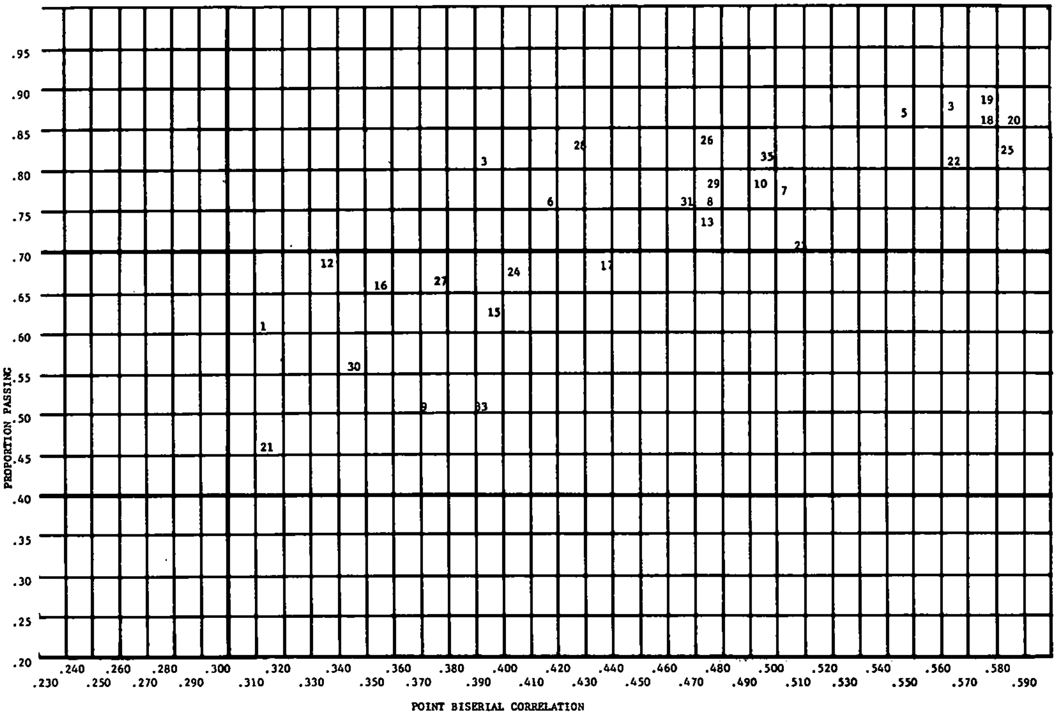

Figure 11. Distribution of questions for Form V as finally revised.

In addition, if an item was chosen that had foils that did not have at least 2 percent of the subjects choose it, either it was replaced in its entirety or the affected foil was revised. This involved 4 questions in Form I, 3 questions in Form II, and 1 question in Form V. In Forms III and IV, no questions were involved that required foil changes.

It should be pointed out that many of the questions showed point biserial correlation coefficients that were extremely high, in many cases in excess of 0.500, which is considered to be in the extreme upper regions. Normally 0.400 is considered exceptional. With the exceedingly high point biserial coefficients, however, there is also a correspondingly high proportion of subjects who passed the item. Thus, although discrimination is high in these cases, difficulty is somewhat low.

Statistical Data of Total Scores

Table 6 gives the mean score, standard deviations, and number of subjects for each of the 5 test forms.

TABLE 5

NUMBER OF ITEMS MEETING SELECTION STANDARDS FOR INCLUSION IN FINAL TEST FORMS

Form	Biserial Coefficient Standard	Proportion Passing Standard	Both Standards	Neither Standard
I	15	3	11	1
II	15	0	12	3
III	13	3	13	1
IV	17	0	13	0
V	19	0	11	0

TABLE 6

MEAN SCORES, STANDARD DEVIATIONS, AND NUMBER OF SUBJECTS FOR SECOND PERIOD

Form	Mean Score	Standard Deviation	No. of Subjects
I	21.10	5.39	1,430
II	22.36	5.88	1,405
III	21.23	5.35	1,420
IV	23.24	6.21	1,386
V	24.56	6.28	1,411

Note: Total score possible was 35.

As indicated earlier for the first data collection period, the number of subjects taking the examination was less than the anticipated 10,000 (2,000 for each of the 5 forms) because of the lack of returns from the Chicago examination stations. Thus, from downstate Illinois, the maximum number of answer sheets possible for each form of the examination was 1,406. Except for the stations in Marion and DeKalb, where no answer sheets were returned, in Macomb, where Form IV was missing, and in Clinton, where Forms IV and V were missing, all answer sheets were returned. Unlike the first data collection period, the second period had very few answer sheets that had to be discarded because of incompleteness or illegal responses. Thus the effective return rate was 97.7 percent. The differences shown in the numbers for Forms I, II, and III are based entirely on the rejection rate of returned answer sheets. Forms IV and V had a higher rate of incomplete returns and thus would be expected to have lower numbers of subjects.

Because comparisons of scores achieved on each form of the examination need to be made and because the desired equivalence of forms could not be tested in this study because of different populations as well as the different instruments, the means need to be adjusted by the use of standard scores. We can assume that the populations used were randomly selected and that the forms of the examination were equal because we made categories of questions to test specific points and then randomly assigned them to each form. Therefore, it is valid to compare the raw scores on the same base by using standard scores. The differences in the means given in Table 6 are all well within one standard deviation of each other. For an arbitrary raw score of 25 out of the possible 35, the corresponding standard scores for Forms I through V of the test are 0.72, 0.40, 0.70, 0.28, and 0.07.

Readability

Because most writers stress that the difficulty of reading the test should not be a factor in determining the chance of passing the test, research into readability was included. Although this form of readability analysis makes no attempt to adjust for the special readability problems of minority groups, it does attempt to provide a lowest common denominator suitable to the majority of the population under study. Flesch (2) states that testing readability includes the reading ease and the human interest of the passage. He advocates using 100-word sample passages for testing purposes.

Reading ease measures the length of the sentences and words. The longer they are, the more difficult it is to read. The average sentence length (2) is calculated by dividing the number of words in all the samples by the number of sentences in all the samples. The average word length is found by dividing the number of words in a sample into the number of syllables multiplied by 100. Then the formula to find the reading ease score is as follows:

$$206.835 - (\text{average sentence length})\,(1.015) + \\ (\text{number of syllables per word})\,(0.846)$$

For the desired level of eighth-grade education, a reading ease score of 60 to 70 is accepted, made up of a syllable-per-100-words score of 147 and an average sentence length of 17 words.

A readability study was performed on each of the 5 test forms using the formula developed by Flesch. For each test form the 2 criteria were computed on the basis of three samples, each consisting of 3 questions, for a total of 15 random samples. Computation was simplified by using the Farr-Jenkins Tables (2) to find the Flesch reading ease score for each passage. The results of these computations are given in Table 7. Although reading difficulty varied from question to question, each test form except the second had very similar results. The difficulty of Form II was rated as fairly easy while the other 4 forms were rated as fairly difficult. Thus Form II is about the seventh-grade level and the other forms are about the tenth- to twelfth-grade levels. When 15 samples were calculated together, the overall reading level was tenth grade. Sampling variation may account in part for the discrepancy of the second form because the mean test scores do not indicate that this is the easiest form of the examination; further study, therefore, may be warranted on a question-by-question basis to determine

TABLE 7

READABILITY OF FINAL TEST FORMS

Form	Passage Questions	Average Sentence Length	Average Word Length	Reading Ease Score	Difficulty Level	Grade Level
I	6, 7, 8	22.33	169	41	Difficult	College
	12, 13, 14, 15	19.75	165	48	Difficult	12
	23, 24, 25	12.67	134	81	Easy	6
	All	18.22	156	56	Fairly difficult	11
II	4, 5, 6	14.67	147	70	Fairly easy	7
	12, 13, 14	16.33	142	71	Fairly easy	7
	24, 25, 26	16.33	134	77	Fairly easy	7
	All	15.78	141	71	Fairly easy	7
III	10, 11, 12	21.00	157	53	Fairly difficult	11
	16, 17, 18	13.00	156	62	Standard	9
	27, 28, 29	15.00	168	50	Fairly difficult	12
	All	16.33	160	55	Fairly difficult	11
IV	1, 2, 3	12.67	167	53	Fairly difficult	11
	9, 10,11	17.67	160	54	Fairly difficult	11
	20, 21, 22	29.67	148	52	Fairly difficult	12
	All	20.00	158	53	Fairly difficult	11
	3, 4, 5	15.00	152	63	Standard	9
	8, 9, 10	15.33	161	56	Fairly difficult	11
	28, 29, 30	26.67	163	43	Difficult	College
	All	19.00	159	53	Fairly difficult	11
All		17.87	155	58	Fairly difficult	10

correlation between proportion of subjects passing the question and its reading difficulty. Although the examinations do not meet the eighth-grade level desired, they do not seem to be too far out of line.

CONCLUSIONS

1. Because data were not obtainable from the 3 examination stations in Chicago, all conclusions are valid only for downstate Illinois.
2. Although the search for an appropriate external criterion on which the item analysis could be based was futile, it did succeed in eliminating several criteria previously considered important. It also indicated a few criteria that warrant further study.
3. The item analyses show that the final form of the examination is one that can discriminate between individuals who do know and understand the "Illinois Rules of the Road" and those who do not.
4. The mean scores of the 5 test forms may be significantly different statistically and, therefore, adjustments need to be made on the passing score for each form. However, these differences are inherently tied to the extremely high number of subjects (7,038) used, and the practical differences are questionable.
5. The results of the multiple-regression analysis of the criteria scores indicate the need for further research in the area of significant external criteria. Because less than 20 percent of the test score variance is accounted for by the 14 criteria, the remaining +80 percent needs further study in the hope of finding at least some criteria to account for the major portion of this quantity.
6. The readability level of the examinations is a little high. Because most of the language used in the questions is taken from the "Illinois Rules of the Road," the readability level of this source of information is also in question.

RECOMMENDATIONS

Based on the experience of this study, the authors feel that certain recommendations should prove useful to anyone interested in conducting a similar investigation. Also, these recommendations form a foundation for further study in this area.

1. Further analysis of the influence of the reading difficulty of the questions used is required. A question-by-question readability testing needs to be done accompanied by a correlation analysis with the proportion passing each item.

2. Because only a small sample of questions was used in the readability analysis and because different combinations of questions were used for each test form, readability testing of the total examination for each form needs to be done and then compared to the mean scores for each test form.

3. Because the readability of the examination questions is closely allied to the readability of the "Illinois Rules of the Road" and because the "Illinois Rules of the Road" is the predominant source of information concerning driving in Illinois, readability testing of this booklet is warranted.

4. Further analysis needs to be done on the examinations of applicants who were previously licensed in another state. This criterion proved to be the most significant of the 14 chosen for this project. By examining the mean scores for each state and the items that were answered significantly better by out-of-state applicants, and by comparing this information with the driver-licensing program in those states that show marked superiority, perhaps significant additions can be made to the Illinois program.

5. Analysis of the questions most frequently missed by those who have completed high school driver-education programs could be of great assistance in revising the Illinois driver-education curriculum.

6. A continuing search for a valid external criterion for the item analyses seems warranted. Further analysis using factors such as reading ability, intelligence quotient, or some other academic indicator might prove worthwhile.

7. For the benefit of special educational opportunities for various segments of the population within the state, an analysis of item difficulty by driver-licensing examination stations may supply information on areas of educational needs for the people in the district.

8. In order to establish the degree to which the 5 test forms are equivalent, the 5 forms need to be given to the same population, and an analysis of variance performed on the results.

9. Tighter controls are needed during the collection of data. With over 400 examiners giving the instructions and presiding over 7,000 examinations, variations in filling out the answer sheets and in the interpretation of questions are inevitable. Problems of incomplete data need to be minimized. Selection of subjects to take the examination needs to be standardized. The problems encountered in obtaining returns from all the collection points emphasize the difficulties in working with large politically oriented agencies. Wherever possible, the administration of examinations and the collection and treatment of data should be kept clear of political pressures and involvement.

10. In the interest of ascertaining the predictability of the instrument developed in this project as it relates to accidents and traffic violation situations, a long-term study should be done in which areas of difficulty as indicated by the marking of incorrect answers in the examination can be correlated with causal factors involved in the accidents and violations. At least a 3-year time period should be allowed for the accumulation of risk experience on the part of the driver.

REFERENCES

1. Campbell, B. J. The Improvement of Written Driver License Examination Through Test Analysis. Univ. of North Carolina, Chapel Hill, PhD dissertation, 1959.
2. Flesch, Rudolf. How To Test Readability. Harper and Brothers, New York, 1951.
3. Stanley, Julian C. Measurement in Today's Schools. Prentice-Hall, Englewood Cliffs, N. J., 1964.
4. Draft Highway Safety Program Standard No. 4.4.5. National Highway Safety Agency, U. S. Department of Commerce, Feb. 1967.
5. U. S. Public Law 89-564. Section 402, Sept. 9, 1966.