TRANSPORTATION RESEARCH

RECORD

No. 1421

Highway Operations,
Capacity, and Traffic Control

Traffic Control Devices, Visibility, and Traffic Signal Systems

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1993

Transportation Research Record 1421
ISSN 0361-1981
ISBN 0-309-05574-1
Price: $26.00

Subscriber Category
IVA highway operations, capacity, and traffic control

TRB Publications Staff
Director of Reports and Editorial Services: Nancy A. Ackerman
Associate Editor/Supervisor: Luanne Crayton
Associate Editors: Naomi Kassabian, Alison G. Tobias
Assistant Editors: Susan E. G. Brown, Norman Solomon
Office Manager: Phyllis D. Barber
Senior Production Assistant: Betty L. Hawkins

Printed in the United States of America

Transportation Research Record 1421

Contents

Foreword

The papers in this volume address some of the problems and issues facing urban traffic engineers as they grapple with the needs of an ever more complex traffic system. They were presented in various sessions at the 1993 TRB Annual Meeting and were sponsored by the TRB Committees on Traffic Control Devices, Visibility, and Traffic Signal Systems.

Readers with a specific interest in signalization will find papers related to delay and progression comparisons for leading and lagging left-turn signalization, guidelines for the selection of the most appropriate left-turn phasing and indication sequence, analysis of flashing signal operation, and use of simulation to determine the need for a traffic signal installation.

Traffic signal systems are addressed in a number of papers covering topics that include decision support systems for modification of signal control in an urban signal system, the optimal design of traffic detector layout and controller timing in terms of safety and operations, development of an enhanced delay estimation model, and development of different computer programs for network signal timing and optimization.

Traffic control devices and visibility are the focus of four papers. The reader will find papers related to the concepts of a separate category of traffic signs to control traffic in emergency situations, the adequacy of current federal Interstate vision standards for commercial motor vehicle operators, the validity of the current specifications for entrance angles for retroreflective traffic signs, and the development of a computer program to determine the actual angles at which traffic signs are seen by drivers.

Operational Comparison of Leading and Lagging Left Turns

JIM C. LEE, ROBERT H. WORTMAN, DAVID J. P. HOOK, AND MARK J. POPPE

Field studies were conducted in the Phoenix and Tucson metropolitan areas for the purpose of comparing the operational differences between leading and lagging left-turn signal phases. Delay studies were conducted in both areas at isolated intersections, and the influence on signal system progression was evaluated at three locations in the Phoenix area. In both areas, the lagging left-turn phases do not utilize overlaps even though phase overlaps were used with the leading left-turn operations. Intersections with protected-only as well as protected-plus-permissive left turns were included. The study involved a before-and-after analysis of the intersection as well as arterial operations. For the individual intersections, the field studies found that the intersection delay is significantly greater with the lagging left-turn operation. This finding was true for both of the metropolitan areas. In terms of the signal system progression, no significant differences were found in progression among the leading, lagging, and mixed operations.

In 1985 the city of Tucson, Arizona, initiated an effort to convert the protected left-turn signal phases from a leading to lagging operation. It was believed that the use of lagging left turns would improve intersection operations and network flows.

In order to provide uniformity in the area, Pima County converted from leading to lagging left-turn operations in 1987. On the basis of the Tucson experience, other jurisdictions in Arizona began to consider changing to the lagging left-turn phasing. Scottsdale, which is in the Phoenix metropolitan area, converted their protected left-turn phasing to a lagging operation in early 1989.

It should be noted that most intersections in Arizona with a protected left-turn phase also have a permitted phase that allows motorists to turn left through gaps in opposing traffic. At intersections with permitted/protected phasing, simultaneous lagging left-turn arrows are used to avoid trapping motorists who have pulled into the intersection while waiting to turn.

The concept of the lagging green left-turn interval is not new. Neither is the question whether leading or lagging left turns are preferable, as shown by the discussion in the 1965 edition of the *Traffic Engineering Handbook* (*1*). It is noted that the use of either leading or lagging green should be approached with extreme caution because a motorist who is receiving the shorter green might not realize it since the driver sees opposing traffic flowing freely. In addition, some authorities believe that the leading green is probably less hazardous than the lagging green because motorists in opposing directions would generally be starting from a stopped position. Nevertheless, other authorities favor the lagging green because they believe that the left-turn capacity is increased.

The potential for the lagging left turn's being more hazardous as mentioned above refers to what is sometimes called the "trap" of lagging left turns.

The simultaneous dual-lag operation is utilized by the city of Tucson, which has had the most experience within the state with lagging left-turn operations. In 1984 Tucson conducted an experiment on 22nd Street from Tucson Boulevard to Kolb Road. In this study, it was found that converting from leading to lagging operation reduced delay, fuel consumption, emissions, and accidents (memorandum from Joel D. Valdez, City Manager, to Mayor and Council, May 10, 1985).

In a study based on a simulation model called TEXAS, Machemehl and Mechler investigated various left-turn sequence patterns at an isolated intersection. They reported no significant difference in delay between leading and lagging turn phases (*2*).

The literature does not support the current phasing practices within the state, particularly the apparent need for standardization of either leading or lagging operation within the various governmental jurisdictions. Conversely, the literature generally recommends that the decision for leading versus lagging operation be based on conditions at the specific intersection and the opportunity to provide the best progression.

In this paper the studies and findings relative to the effect of leading versus lagging left turns on intersection delay and signal system progression are described. Studies were conducted in both the Phoenix area and Pima County. The following studies were performed:

- Phoenix Area Intersection Delay Study
  - –Leading versus lagging,
  - –Leading versus combination;
- Pima County Signal Operation Analysis
  - –Vehicle arrival,
  - –Vehicle delay,
  - –Cycle length;
- Phoenix Area Travel Time Study
  - –Leading versus lagging,
  - –Leading versus combination.

At seven intersections in Glendale, Tempe, and Mesa, intersection delay with leading left turns was compared with that with lagging left turns. At the one intersection studied in Mesa, the only after condition involved a leading left turn in

Lee Engineering, Inc., Suite 310, 2701 E. Camelback, Phoenix, Ariz. 85016.

one direction and a lagging left turn in the opposing direction. The delay in the Phoenix area was obtained by counting the queued vehicles in 15-sec increments. The Pima County intersection delay was obtained using time-lapse photography with both leading and lagging left turns at nine locations.

Signal system progression was evaluated in Glendale, Tempe, and Mesa. Four conditions in Glendale and Tempe were examined: existing timing (all leading), optimized all-leading timing, optimized all-lagging timing, and optimized combination. The signal progression evaluation in Mesa consisted of evaluating the existing leading operation and combination of leading left eastbound and lagging left westbound. A minimum of 2 weeks was provided between the before-and-after delay studies. It was the intent of the researchers to allow enough time for drivers to become familiar with the new phasing but to keep this time short so volume and vehicle mix were not substantially changed. An evaluation of the accidents for leading and lagging operation was also conducted but is not included in this paper.

PHOENIX AREA INTERSECTION DELAY STUDY

Intersection stopped-time delay studies were conducted to evaluate the difference in performance between leading and lagging left-turn arrow operation. One intersection was studied to evaluate the difference between leading and combination-leading-and-lagging operation.

Leading Versus Lagging Operation

A paired comparison was made between the average delay per vehicle in the leading condition and the average delay per vehicle in the lagging condition. Six intersections were used in the analysis, and manual stopped-time delay studies were conducted at each intersection before any signal timing changes associated with this research were made. In the before condition, each of the six intersections operated with leading left turns. Five of the six intersections operated with protected/permissive left-turn phasing and third-car actuation on all approaches. The intersection of 48th Street and Broadway operated with protected-only left turns and first-car actuation on the northbound and southbound approaches and protected/permissive left-turn phasing with third-car actuation on the eastbound and westbound approaches.

Manual stopped-time delay studies were conducted at each intersection with lagging operation. All approaches that were protected/permissive in the leading condition were permissive/protected in the lagging condition. The two protected-only approaches remained protected only in the lagging operation.

Results

A before-and-after difference in the average stopped-time delay per approach vehicle was calculated for each intersection. A difference was calculated for left-turn vehicles, through or right-turn vehicles, and total intersection approach vehicles. The percent change in delay from the before to the after condition was also calculated. The results of the Phoenix area intersection analysis of leading versus lagging left-turn operation are presented in Table 1.

Average stopped-time delay per left-turn approach vehicle increased in the after condition at four of the six intersections studied. The largest change occurred at 51st Avenue and Northern, where delay increased by 139 percent for left-turn vehicles. The intersection of 48th Street and Southern measured essentially no change for left-turn vehicle delay with the conversion to lagging left turns, whereas the intersection of 48th Street and Broadway registered a 5 percent decrease in delay for left-turn vehicles in the after condition.

Average delay per through or right-turn approach vehicle increased at five of the six intersections studied. The largest increase occurred at 48th Street and Southern, with 129 percent more delay for through or right-turn vehicles in the after condition. The intersection of 51st Avenue and Northern was the only one that registered a decrease in delay for through or right-turn vehicles in the after condition. Delay decreased approximately 16 percent at this location.

Average delay per total approach vehicles also showed increases in the after condition at the same five intersections, though the changes were not as drastic when total intersection approach vehicles were considered. The large increase in through or right-turn delay at 48th Street and Southern was partially offset by no change in left-turn delay. However, this intersection still registered the largest increase (85 percent) in total intersection delay with the conversion to a lagging operation. The intersection of 51st Avenue and Northern was the only location that registered an overall improvement in the after condition, with a decrease in total intersection delay of approximately 4 percent.

Three statistical tests were performed: difference by intersection left-turn movements, difference by intersection through or right-turn movements, and difference by total intersection delay. In each case, the statistical test performed was a paired *t*-test using the difference for each pair as one observed value. A mean of the difference was then calculated. The null hypothesis for the each test was that the difference between the before and after conditions is equal to zero. A two-tail test was performed at a 95 percent level of confidence.

The results of the paired data analysis are also presented in Table 1. On the basis of this analysis, it is concluded that left-turn delay and total intersection delay are significantly greater for the lagging left-turn operation.

Leading Versus Combination Operation

Two delay studies were performed at the intersection of Southern Avenue and Stewart in Mesa to compare the difference in delay for a leading operation and a combination-leading-and-lagging operation. Southern Avenue is an east-west arterial street and Stewart is a local collector street. The signal operated in a five phase mode in the before condition with protected-only phasing on the east and west approaches. The combination phasing operated with leading left turns in the eastbound direction and lagging left turns in the westbound direction. The signal was also operated in the protected-only mode in the after condition.

The results of the before-and-after study are presented in Table 2. Delay per intersection left-turn approach vehicle

TABLE 1 Intersection Delay for Leading Versus Lagging Left-Turn Operation, Phoenix Area

Intersection		Delay per Approach Vehicle (sec/veh) Left Turn	Thru/Right	Total
1. 51st Ave./Glendale	Before	25.70	22.55	22.95
	After	57.79	34.34	37.66
	Difference	32.09	11.79	14.71
	Change	125%	52%	64%
2. 51st Ave./Northern	Before	23.51	44.57	41.57
	After	56.24	37.32	39.80
	Difference	32.73	-7.25	-1.77
	Change	139%	-16%	-4%
3. 51st Ave./Olive	Before	27.50	21.58	22.41
	After	45.30	27.65	30.19
	Difference	17.80	6.07	7.78
	Change	65%	28%	35%
4. 51st Ave./Peoria	Before	42.03	20.07	22.88
	After	65.64	33.83	38.00
	Difference	23.61	13.76	15.12
	Change	56%	69%	66%
5. 48th St./Southern	Before	54.95	21.56	27.23
	After	54.92	49.28	50.30
	Difference	-0.03	27.72	23.07
	Change	-0%	129%	85%
6. 48th St./Broadway	Before	63.39	39.27	44.51
	After	60.14	43.97	47.91
	Difference	-3.25	4.70	3.40
	Change	-5%	12%	8%
Analysis				
Sample Size		6	6	6
Mean of Difference		17.16	9.47	10.38
Overall Change		63.30%	45.54%	42.17%
Sample Standard Deviation		15.62	11.58	9.00
Test Statistic (t)		2.691	2.002	2.825
Significant @ 95%?		yes (p=.04)	no (p=.10)	yes (p=.04)

Before Condition: Leading Operation
After Condition: Lagging Operation

TABLE 2 Intersection Delay for Leading Versus Combination Operation, Phoenix Area

Intersection		Delay per Approach Vehicle (sec/veh) Left Turn	Thru/Right	Total
Southern/Stewart	Before	37.86	10.76	14.34
	After	33.25	9.63	13.02
	Difference	-4.61	-1.13	-1.32
	Change	-12%	-11%	-9%

Before Condition: Leading Operation
After Condition: Combination (leading EB/lagging WB)

decreased by 4.61 sec, or approximately 12 percent, in the after condition. The decrease was 1.13 sec per vehicle for the through or right-turn movements. This represents a change of approximately 11 percent. Total intersection delay decreased by 1.32 sec per approach vehicle in the after period. Total intersection delay decreased by approximately 9 percent with the conversion to a combination-leading-and-lagging operation.

PIMA COUNTY SIGNAL OPERATION ANALYSIS

The conversion from leading to lagging left-turn signals by Pima County in 1987 represented a unique opportunity to examine the effect of the operational change. With the cooperation of the Pima County Department of Transportation and the Arizona Department of Transportation, a before-and-after data collection effort was undertaken at selected intersections.

Selection of Intersections

The conversion program in Pima County involved a total of 37 signalized intersections in the Tucson area. At some of these intersections, various modifications to the signal operations were made in addition to the conversion of the left-turn phasing. At a limited number of intersections, the only planned change was to switch from the leading to the lagging left-turn operation; thus these intersections were selected for the before-and-after data collection. The intersections studied are given in Table 3.

Ultimately, the intersection of First Avenue and Ina Road underwent other changes in signal phasing as well as modifications in lane use that significantly changed the operation of the intersection. For this reason, the intersection was eliminated from the comparative analysis, although field data were collected at the site. In addition, the initiation of construction in the area of Ina Road and Thornydale Road significantly changed the traffic at that location before there had been an opportunity to collect the after data.

Signal Phasing

Pima County uses actuated control for traffic signals; thus all of the intersections in the study utilized full actuated control. In addition, each of the intersections operated on an isolated basis with no interconnection among adjacent signals. Table 3 identifies the operation of the left-turn signal phasings at each of the study locations.

It should be noted that phase overlaps were used for the leading left-turn conditions; however, the overlaps were not used with the lagging left-turn operations. At a limited number of intersections that utilized the protected-only left turns, a phase overlap condition would occur with the lagging left-turn operation. For example, one intersection had very low westbound approach volumes. For some cycles, the eastbound through and left-turn movements would occur at the same time.

With respect to the actual signal timing, the study utilized the signal settings employed by Pima County for the before and after conditions. There was no attempt by the research team to evaluate the signal timing settings used at the intersections.

Data Collection

For the field data collection, two time-lapse super-8-mm movie cameras were used to film the operation of each of the intersections. The filming of each intersection occurred during

TABLE 3 Delay Study Intersections, Pima County

Intersection	Type of Control (a,b)
Ajo Way/Alvernon Way	4 Phase (c)
Alvernon Way/Irvington Rd.	4 Phase (Protected/Permissive)
Campbell Ave./Skyline Rd.	3 Phase (Protected)
First Ave./Ina Rd.	4 Phase (d)
First Ave./Orange Grove Rd.	3 Phase (Protected/Permissive)
First Ave./River Rd.	3 Phase (Protected/Permissive)
Ina Rd./Thornydale Rd.	4 Phase (Protected/Permissive)
Kolb Rd./Valencia Rd.	4 Phase (Protected)
Palo Verde Rd./Valencia Rd.	3 Phase (Protected)

(a) The number of phases reflects the basic operation of the intersection. Phase overlaps were used in situations with opposing leading protected left turns.

(b) In the "after" condition, the "protected/permissive" left turn operation obviously becomes "permitted/protected."

(c) At the intersection of Ajo Way and Alvernon Way, a combination of types of control was used. For example, some approaches had protected left turn operations.

(d) At the intersection of First Avenue and Ina Road, a 4-phase signal operation was used in the before condition with protected/permissive left turns on the northbound and westbound approaches. For the after condition, the northbound and southbound approaches on First were treated as separate phases. In addition, the lane use on the northbound approach was changed.

the period from 3:00 to 6:00 p.m. on weekday afternoons in the before and after periods.

Analysis

Using the film record of the intersections during the before and after periods, data that reflected operational parameters were extracted. These operational data for each intersection were then used for the comparative analysis of the leading and lagging left-turn phasing. The discussion that follows presents the analysis of and results for each of the operational parameters.

Intersection Volume

In the design of the data collection effort, it was recognized that significant changes in volume can have a potential impact on the operational measures of intersection performance. For this reason, a number of precautions were taken in an attempt to minimize the possibility of major changes in volume between the before and after study periods.

Table 4 presents the average approach volumes for each intersection. At most of the study intersections, only minor differences in traffic volumes were observed. Given the relatively short period between the collection of the before and the after data, only small differences would be expected. Two exceptions were noted. There was no explanation for the cause of the significant increase in traffic volume at the intersection of Campbell Avenue and Skyline Road. The before data set was collected in April 1987, and the after data set was taken the following October. Although only 6 months passed between the data collection periods, there was a 21 percent increase in the approach volumes at that intersection. This increase generally occurred on all approaches and throughout the study period. In essence, there was a major increase in the use of the intersection.

In the second exception, there was a 16 percent decrease in the approach volume at the intersection of Kolb and Valencia roads. There had been a major change in employment in the vicinity of this intersection; thus the after condition was influenced by the reduction in employment.

Arrival of Vehicles

The arrival pattern of vehicles for a given intersection was examined by determining the percentage of the approach vehicles that had to stop because of the operation of the traffic signal. Basically, review of the film revealed the approach vehicles that were required to stop as well as those that were able to pass through the intersection without stopping. The percentage of vehicles stopped was then calculated by comparing the number of vehicles that stopped with the total approach volume. Table 5 summarizes this information for each intersection.

At most of the intersections the percentage of stopped vehicles was in the general range of 50 to 55. The main exception was the intersection of Palo Verde and Valencia roads, where the percentage for the before and after conditions was significantly lower than that for other intersections. This lower value can be explained by the presence of a free-flow right-turn lane on one of the approaches.

Vehicle Delay

Table 6 summarizes the results of the delay analysis and indicates the average stopped delay for the stopped vehicles as well as for the approach vehicles. These values reflect the overall delay for an intersection. At all the intersections where delay was actually measured, there were increases in the average delay per vehicle. Even for the intersections where there were decreases in the approach volume, the average vehicle delay increased.

Cycle Length

The average signal cycle lengths for the before and after periods for each intersection are given in Table 7. A general review of the table reveals that the differences in the cycle lengths vary from intersection to intersection, with increases at some of the sites and decreases at others. At intersections where there was a decrease in the cycle length, the permitted/protected left turn was utilized. The increases in cycle length were at intersections where protected-only left turns were

TABLE 4 Intersection Total Approach Volumes, Pima County

	Average Approach Volume (vph)*		
Intersection	Before	After	Difference
Ajo Way/Alvernon Way	3644	3523	-3%
Alvernon Way/Irvington Rd.	2788	2882	3%
Campbell Ave./Skyline Rd.	2527	3070	21%
First Ave./Orange Grove Rd.	2519	2472	-2%
First Ave./River Rd.	3379	3107	-8%
Ina Rd./Thornydale Rd.	3495	**	**
Kolb Rd./Valencia Rd.	7052	5950	-16%
Palo Verde Rd./Valencia Rd.	2560	2472	-3%

* The average approach volumes are for the entire intersection. The value in the table reflects the sum of all approaches.
** After values not available for Ina Rd./Thornydale Rd.

TABLE 5 Percentage of Approach Vehicles Stopped, Pima County

Intersection	Percent Stopped	
	Before	After
Ajo Way/Alvernon Way	54.1	53.0
Alvernon Way/Irvington Rd.	54.5	53.8
Campbell Ave./Skyline Rd.	50.7	55.6
First Ave./Orange Grove Rd.	55.6	49.5
First Ave./River Rd.	54.4	55.7
Ina Rd./Thornydale Rd.	60.6	*
Kolb Rd./Valencia Rd.	60.1**	70.4**
Palo Verde Rd./Valencia Rd.	31.3	33.3

* After value not available for Ina Rd./Thornydale Rd.
** At the Kolb Rd./Valencia Rd. intersection, the values are for the eastbound and westbound approaches only. For the before condition, the percent vehicles stopped for all approaches was 49.2 percent. The after condition value for all approaches was not available.

TABLE 6 Vehicle Delay Comparison, Pima County

Intersection	Delay Per Stopped Vehicle (Sec)	Delay Per Approach Vehicle (Sec)
Ajo Way/Alvernon Way		
Before	32.68	17.75
After	39.68	21.04
Difference	7.00 (21%)	3.29 (19%)
Alvernon Way/Irvington Rd.		
Before	22.82	12.44
After	32.32	17.39
Difference	9.50 (+42%)	4.95 (+40%)
Campbell Ave./Skyline Rd.		
Before	27.45	13.93
After	31.43	17.47
Difference	3.98 (+14%)	3.54 (+25%)
First Ave./Orange Grove Rd.		
Before	22.88	12.72
After	27.11	13.43
Difference	4.23 (+18%)	0.71 (+6%)
First Ave./River Rd.		
Before	32.15	17.48
After	33.55	18.68
Difference	1.40 (+4%)	1.20 (+6%)
Ina Rd./Thornydale Rd.		
Before	33.03	20.01
After	*	*
Difference		
Kolb Rd./Valencia Rd.		
Before	26.04	12.69
After	*	19.27
Difference		6.58 (+52%)
Palo Verde Rd./Valencia Rd.		
Before	19.25	6.03
After	23.58	7.85
Difference	4.33 (+22%)	1.82 (+30%)
Average Change	+20%	+30%

* After value not available

TABLE 7 Average Cycle Length, Pima County

Intersection	Average Cycle Length (Sec)		
	Before	After	Difference
Ajo Way/Alvernon Way	95.3	114.3	19.0
Alvernon Way/Irvington Rd.	72.6	70.4	-2.2
Campbell Ave./Skyline Rd.	79.9	90.3	10.4
First Ave./Orange Grove Rd.	77.3	71.9	-5.4
First Ave./River Rd.	95.6	90.7	-4.9
Ina Rd./Thornydale Rd.	85.8	*	
Kolb Rd./Valencia Rd.	65.7	76.7	11.0
Palo Verde Rd./Valencia Rd.	62.1	62.6	0.5

* After value not available for Ina Rd./Thornydale Rd.

utilized. Changes in cycle length, therefore, were a function of whether left turns were permitted along with the through movement or not.

The exception to an increase in cycle lengths with protected-only lagging left turns occurred at the intersection of Palo Verde and Valencia roads. At this intersection, the average cycle lengths remained virtually the same even with the protected left-turn operations. Because of the low approach volumes for some movements, this is one of the intersections that resulted in a phase overlap type of operation. Because of this condition, the average cycle length remained the same.

Discussion of Results

In considering the results of the analysis of the Pima County intersections, it must be recognized that

- All of the study locations were operating with actuated control;
- The signals were basically isolated from other intersections, and there was no coordination with adjacent intersections at the time of the data collection;
- The intersections were not operating at what could be considered saturated conditions; and
- Vehicle queues generally cleared during each cycle.

There was some variation in the measured approach volumes at the study intersections; however, major changes occurred at only two intersections. Because the intersections were not operating at saturated conditions, increase in volumes would not necessarily result in significant increases in delay.

Generally, there was little change in the percentage of vehicles stopped. This would suggest that the arrival pattern was random in terms of the signal cycle. For this reason, the effect of platooning should not be a factor with respect to delay calculations and measurements.

It is significant to note that the reduction in cycle length was associated with intersections where permitted left turns were allowed. On the other hand, intersections with protected left turns only had increases in cycle length with the lagging left-turn operation. This result is reasonable because of the fact that the opportunity for phase overlap was lost when the lagging left turn was used.

An interesting result of the analysis is that vehicle delay increased at all study intersections. There was an average increase of 20 percent in the delay per stopped vehicle and an average increase of 30 percent in the delay per approach vehicle. Even when there was a decrease in approach volumes, there were increases in delay. Delay might be expected to increase with longer cycle lengths; however, it also increased at intersections with reductions in average cycle lengths.

PHOENIX AREA TRAVEL TIME STUDY

As part of this research, alternative phasing sequences were tested using travel time data along five routes in Glendale, four routes in Tempe, and one route in Mesa.

In order to obtain a true comparison between leading and lagging left turns, it was necessary to use signal timing patterns developed by a common optimization program. Because of the ease of operation and the numerous runs that would be required as part of the combination portion of the study, FORCAST was utilized to optimize the signals.

Once the timing plans were implemented on the street, travel time runs were performed using the "floating car" method with the TIMELAPSE Travelog data collection computer.

Six travel time runs were performed for each route in each direction for three time periods: a.m. peak, p.m. peak, and off-peak. One driver collected all the data in Glendale and another driver collected the travel time data for Tempe. The same driver was used for all runs in each city in order to eliminate the variability of different drivers.

The six runs were averaged for each route to determine the average stops, delay time, and travel time for each route. Each of the estimates for the routes was multiplied by its respective volume to produce a weighted point estimate based on the route volume. A paired Student's *t*-test was then performed between each sample. The following comparisons were made:

- Existing leading minus FORCAST optimized leading,
- Existing leading minus FORCAST optimized lagging,
- Existing leading minus FORCAST optimized combination,
- FORCAST leading minus FORCAST lagging,
- FORCAST leading minus FORCAST combination, and
- FORCAST lagging minus FORCAST combination.

Because FORCAST develops timing plans that weight the benefit of reduced stops with reduced delay and travel time, a representative cost for each timing plan was developed using

the information in *A Manual on User Benefit Analysis of Highway and Bus Transit Improvements* (*3*). These values have been updated to 1988 dollars by using the transportation portion of the consumer price index.

ANALYSIS

Glendale Travel Time Study

In the Glendale study area, all the major arterial–major arterial intersections were operating in a protected-permissive leading left-turn mode in the before condition. All the signals within the study area were optimized using the FORCAST signal timing program, but only the signals along 51st Avenue had the phasing patterns changed during the course of the study. The five routes chosen for the Glendale study were 51st Avenue, 59th Avenue, Peoria Avenue, Olive Avenue, and Northern Avenue.

The comparison was made among (*a*) existing leading, (*b*) FORCAST-optimized leading, (*c*) FORCAST-optimized lagging, and (*d*) FORCAST-optimized combination, with the results in shown in Table 8. Figure 1 shows the equivalent motorists' cost based on stopped-time delay, travel time, and stops.

As the data in Table 8 suggest, there is a significant difference in travel time and delay between both the FORCAST leading–FORCAST lagging and the FORCAST leading–FORCAST combination plans. If the cost parameters are viewed separately, it appears that the existing leading timing plan works best for the a.m. peak, the combination plan works best for the midday and p.m. peak, and the lagging plan works best for the off-peak period. In the a.m. peak, the lagging plan also works better than the FORCAST leading or the combination plan. It appears, at least from this information, that lagging left turns work best in situations such as an off-peak period in which left-turn volumes are relatively light.

Tempe Travel Time Study

In the Tempe area, all major arterial–major arterial intersections were operating in protected-permissive leading left-turn operation with the exception of the north and south approaches at 48th Street and Broadway. Because of the dual left turns, these approaches operate in a protected-only leading left-turn mode. FORCAST was used to create timing plans for all signals within the study area; however, alternative phasings were implemented only at 48th Street and Broadway and 48th Street and Southern.

As shown in Table 9, only one result is significant in the Tempe travel time data. FORCAST leading had significantly fewer stops than FORCAST lagging.

From Figure 2, it may be noted that lagging has a higher cost than FORCAST leading or combination in the midday and p.m. peak, but the FORCAST combination has a higher cost in the a.m. peak. The cost difference between leading and lagging is least in the a.m. peak and greatest in the p.m. peak. At the two intersections in Tempe where lagging left turns were implemented, there is a great directional split between left turns in the p.m. peak. Forcing these two movements together has greatly increased the motorists' cost in the p.m. peak.

Mesa Travel Time Study

The city of Mesa changed the phasing at Southern and Stewart avenues from leading east-west to leading east and lagging west. Lee Engineering collected travel time data along Southern Avenue in the a.m., midday, and p.m. peak time periods to determine the effect of this changeover. The results of this change are shown in Table 10.

Although they are not significant, substantial reductions are shown in delay, stops, and travel time because of the change from an all-leading phasing pattern to a combination leading-lagging phasing pattern.

TABLE 8 Travel Time Study Comparisons, City of Glendale

Comparison	Least Delay	Level of Significance (p)	Least Travel Time	Level of Significance (p)	Least Stops	Level of Significance (p)
Existing Leading-FORCAST leading	Existing Leading	.07	Existing Leading	.16	FORCAST Leading	.27
Existing Leading-FORCAST lagging	Existing Leading	.08	FORCAST Lagging	.34	Existing Leading	.73
Existing Leading-FORCAST Combination	FORCAST Combination	.86	FORCAST Combination	.27	FORCAST Combination	.26
FORCAST Leading-FORCAST lagging	FORCAST Lagging	.03	FORCAST Lagging	.01	FORCAST Leading	.43
FORCAST Leading-FORCAST Combination	FORCAST Combination	.02	FORCAST Combination	.01	FORCAST Combination	.87
FORCAST Lagging-FORCAST Combination	FORCAST Combination	.47	FORCAST Combination	.58	FORCAST Combination	.29

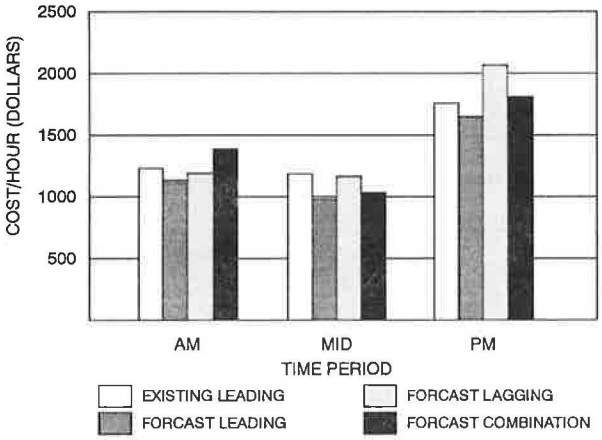


FIGURE 1 Travel time study cost per hour for city of Glendale.

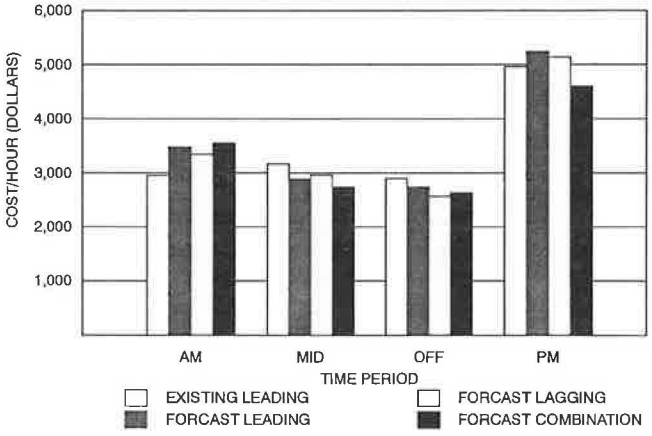


FIGURE 2 Travel time study cost per hour for city of Tempe.

Discussion of Results

It is difficult to determine if either leading or lagging left turns result in better operation for a given situation. Although the difference is not statistically significant, lagging left turns appeared to operate better for three time periods in Glendale (on the basis of FORCAST plans).

The combination timing plan worked better than leading or lagging in Glendale for only midday and the p.m. peak. In Tempe the combination was never the lowest-cost plan. This was surprising, for it was believed that the opportunity for leading or lagging at a particular intersection would help improve progression. It should be stressed again that the FORCAST timing plan must overcome two obstacles in order to choose lagging left turns for intersection phasing. Because it does not recognize left turns made in the permissive period, it does not determine the true best combination plan.

The combination timing plan fared best in Mesa, where stops, delay, and travel time were all reduced substantially. This type of combination phasing is different from those tested in either Glendale or Tempe. The Mesa combination plan was leading eastbound and lagging westbound. In Tempe the phasing tested was leading north-south and lagging east-west. It would appear that there is no substantial reduction in motorists' cost with the Tempe type of phasing, but there is with the Mesa phasing. It is important to realize that to implement the Mesa phasing, it is necessary to have either protected-only operation or programmed-visibility traffic signal heads as are currently being used in Texas.

In conclusion, the following points should be mentioned:

- One of the greatest benefits of lagging left turns is decreased need for a protected left-turn phase. In order for a timing program to implement the best phasing, it is necessary

TABLE 9 Travel Time Study Comparisons, City of Tempe

Comparison	Least Delay	Level of Significance (p)	Least Travel Time	Level of Significance (p)	Least Stops	Level of Significance (p)
Existing Leading-FORCAST leading	Existing Leading	.59	FORCAST Leading	.86	FORCAST Leading	.08
Existing Leading-FORCAST lagging	Existing Leading	.16	Existing Leading	.41	FORCAST lagging	.99
Existing Leading-FORCAST Combination	Existing Leading	.69	FORCAST Combination	.43	FORCAST Combination	.35
FORCAST Leading-FORCAST lagging	FORCAST Leading	.47	FORCAST Leading	.37	FORCAST Leading	.05
FORCAST Leading-FORCAST Combination	FORCAST Combination	.78	FORCAST Combination	.56	FORCAST Leading	.13
FORCAST Lagging-FORCAST Combination	FORCAST Combination	.26	FORCAST Combination	.12	FORCAST Combination	.23

TABLE 10 Travel Time Studies for All-Leading Minus Combination Leading-Lagging Signal Phasing, City of Mesa

Weighted Delay (Vehicle-hours)		Weighted Travel Time (Vehicle-hours)			Weighted Stops (Thousand Vehicle Stops)					
Route	Time	Leading	Lagging	Difference	Leading	Lagging	Difference	Leading	Lagging	Difference
Southern Ave. EB	AM	5	0	5	27	22	6	0.8	0.0	0.8
Southern Ave. EB	MID	52	1	51	129	71	58	5.2	0.0	5.2
Southern Ave. EB	PM	41	8	33	132	90	42	5.9	2.9	2.9
Southern Ave. EB	AM	10	3	8	68	60	8	2.1	2.1	0.0
Southern Ave. EB	MID	5	5	0	76	74	2	2.4	2.4	0.0
Southern Ave. EB	PM	0	5	-5	74	66	8	0.0	2.5	-2.5
Total		113	21		506	382		16.4	10.0	
		Sample Size		6	Sample Size		6	Sample Size		6
		Mean Difference		15.197	Mean Difference		20.579	Mean Difference		1.069
		Std Deviation		21.871	Std Deviation		23.436	Std Deviation		2.675
		Test Stat.		1.702	Test Stat.		2.151	Test Stat.		0.979
		Significant Level of Significance(p)		N .15	Significant Level of Significance (p)		N .09	Significant Level of Significance (p)		N .37

for that program to evaluate the left-turn movement in conjunction with gaps in the opposing traffic stream. Since FORCAST does not do this, it is not a good program for optimizing combination phasing.

- Combination timing seems to work best when leading and lagging are implemented for opposing directions, for example, leading eastbound and lagging westbound.
- In locations like Tempe, where there is a high directionality with opposing left-turn volumes, substantial delay is associated with lagging operation because of the loss of phase overlap.

RESULTS AND CONCLUSIONS

On the basis of the field studies, it was found that intersection delay is significantly greater with lagging left-turn operation. Many factors potentially affect delay, such as loss of phase overlap. In addition, no significant differences were found in progression between the leading, lagging, and mixed operations.

More specifically, the following results were found:

1. Significantly greater delay per approach vehicle occurs with lagging operation than with leading operation for the intersections and time periods tested. It is important to note that the time period tested was generally the p.m. peak hour. During this period it would not be as likely to have sufficiently low left-turn and through volumes to eliminate many protected left-turn phases in the lagging condition.
2. There were no statistically significant differences in stops, delay, or travel time with the different operating conditions. The requirement that the Glendale and Tempe "mixed" operation be limited to either both leading or both lagging on the same street in order to avoid the "trap" restricted potential progression benefit.

The most promise for benefit from lagging or mixed operation was found in the Mesa study, in which leading left-turn operation was utilized for eastbound traffic and lagging for westbound traffic in the after condition. This mixed operation was possible without the trap condition because of the use of protected-only left turns.

The field studies provided valuable insight into the understanding of the many variables that influence left-turn operations. A number of variables have an impact on the effectiveness of left-turn alternatives at a specific site. These variables fall into the general categories of signal control, network considerations, traffic characteristics, and driver perception.

ACKNOWLEDGMENTS

This paper was based on research undertaken as part of project HPR-PL-1 (35), Item 321, "Comparative Analysis of Leading and Lagging Left Turns," and conducted by Lee Engineering, Phoenix, Arizona, in conjunction with the Arizona Department of Transportation and FHWA, U.S. Department of Transportation.

REFERENCES

1. *Traffic Engineering Handbook.* Institute of Traffic Engineers, Washington, D.C., 1965.
2. Machemehl, R. B., and A. M. Mechler. Comparative Analysis of Left-Turn Phase Sequencing. In *Transportation Research Record 956,* TRB, National Research Council, Washington, D.C., 1984.
3. *A Manual on User Benefit Analysis of Highway and Bus Improvements.* American Association of State Highway and Transportation Officials, Washington, D.C., 1977.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Selection Criteria for Left-Turn Phasing and Indication Sequence

SETH A. ASANTE, SIAMAK A. ARDEKANI, AND JAMES C. WILLIAMS

The development of guidelines and recommendations for the selection of left-turn phasing and indication sequences at signalized intersections is documented in this paper. The guidelines developed are based on field studies and use easy-to-obtain data for the selection process. A simple three-level decision process regarding the most suitable left-turn phasing treatment to be used is established. The process favors the least restrictive permissive left-turn phase unless traffic and geometric conditions warrant the more restrictive protected/permissive or protected-only phasings. These guidelines are based on threshold values designed to determine what constitutes an excessive value for any particular variable, beyond which more restrictive left-turn phasing treatments may be justified. The guidelines developed also reflect a selection process that recognizes the trade-off between operational efficiency and safety. The study shows that selection of a particular phase is a multiobjective process involving a number of factors, and in many cases more than one condition must be met to justify the selection of a particular phase that will ensure an optimal solution.

Selection of the appropriate left-turn signal treatment at a signalized intersection involves many options of phase type and phasing sequence. Meanwhile, no comprehensive guidelines have been developed to assist traffic engineers in this task. Existing guidelines are inadequate, and the engineer has to rely on experience or try different treatments until a suitable one is found. Comprehensive guidelines are needed for the selection of the appropriate combination of signal phase pattern and phasing sequence at a signalized intersection. Substantial gains in efficiency and safety of left-turn operations, as well as reductions in fuel consumption and emissions, can be achieved through the implementation of appropriate left-turn signal treatments.

OBJECTIVE

The objective of this study is to develop guidelines for the selection of an appropriate left-turn phase pattern and phasing sequence for a signalized intersection. The phasing patterns considered include protected only, protected/permissive, permissive only, and Dallas phasing. Once a phasing pattern is selected, a decision must be made on the appropriate phasing sequence to be used. Depending on the phasing pattern selected, leading, lagging, or a leading/lagging sequence may be applicable. Phase overlaps are also possible for protected-only and protected/permissive phases. Dallas phasing, which is a modified leading/lagging, protected/permissive sequence, is shown in Figure 1. During the portion of the cycle when one of the left turns is protected and its adjacent through movements plus right turns are displayed a circular green signal, the opposing left turn is permitted, that is, is shown the circular green. Because the throughs and rights adjacent to the permitted left are shown a red signal (because the opposing left is protected), Dallas phasing leads to a unique display: circular green in a five-section head for the lefts, indicating permitted (not protected) turning, and circular red for the throughs plus rights. The five-section head is required for the lefts since the Dallas phasing provides for both protected and permissive left turns. Motorist surveys have indicated that drivers understand the Dallas phasing as well as or better than they understand other types of left-turn phasing (*1*). Furthermore, Dallas phasing provides the advantage of a true protected/permissive, leading/lagging operation without the yellow trap. A detailed description is provided by de Camp and Denney (*2*).

In the development of these guidelines, the following goals were considered:

- Maintain continuity and build on previous research studies in this area,
- Rely on actual field data,
- Provide easy-to-use quantitative measures, and
- Identify, on the basis of statistical analyses, the most suitable left-turn phasing and signal sequence change for a given set of intersection conditions.

BACKGROUND

A detailed review of previous research on guidelines for the selection of left-turn phasing and indication sequences was undertaken. The focus of previous research in this area has been on the development of guidelines for left-turn phasing, that is, whether left-turn protection is needed rather than a specific phasing type and indication sequence. Most of these studies used accidents and delays as decision criteria, but in many cases, either a subset of the factors involved was studied or sample sizes were very small, or both, limiting the scope of the conclusions. Left-turn studies undertaken in various states include those in Kentucky (*3*, *4*), Texas (*5*, *6*), Arizona (*7*, *8*), Florida (*9*), and Virginia (*10*). Through these studies a number of guidelines were formulated for selecting among three types of left-turn phasing, namely, permissive only (PMO), protected/permissive (P/P), and protected only (PTO).

Civil Engineering Department, University of Texas at Arlington, P.O. Box 19308, Arlington, Tex. 76019.

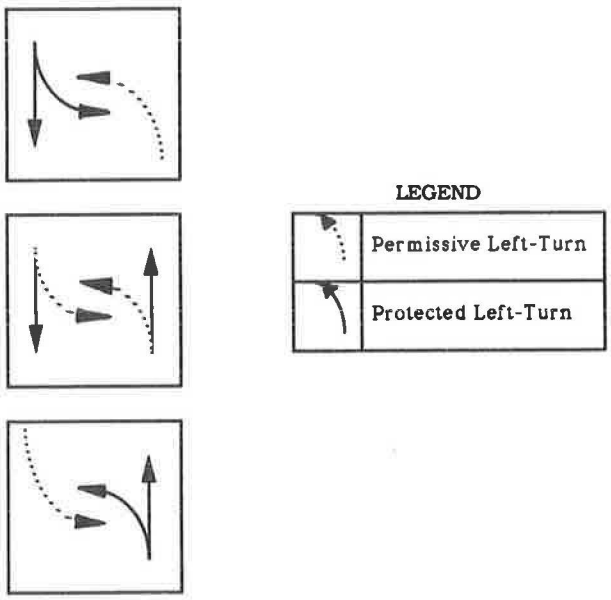


FIGURE 1 Dallas phasing.

STUDY APPROACH

Data were collected from over 100 sites in nine counties across Texas, incorporating a variety of population characteristics. Other site selection parameters included intersection geometry, approach speeds, and signal phasing types and sequences. Such diversification also led to a reduction in coverage error and incorporation of local left-turn signalization policies. The nine counties studied were Bexar, Cameron, Dallas, Ector, Harris, Lubbock, Nueces, Tarrant, and Travis.

In each county, 216 possible intersection approach combinations were considered. These included combinations of six opposing speed limits, 50 to 90 km/hr in 8-km/hr increments (30 to 55 mph in 5-mph increments); three opposing lanes (one, two, and three); two left-turn lanes (one and two); and six phase patterns (leading protected only, lagging protected only, leading protected/permissive, lagging protected/permissive, permissive only, and Dallas phasing). In fact, these 216 approaches represent the upper bound and not the actual number of sites studied, as some combinations did not exist or were not found. The selections were made from lists of intersections supplied by the city and county transportation offices in the counties under study. The intersections selected have little or no pedestrian traffic and exclusive left-turn lanes on the study approaches. The number of intersections and approaches studied within each county is shown in Table 1.

DATA COLLECTION AND REDUCTION

Traffic, geometric, and delay data were collected at all the intersections selected for the study. The objective of this field delay study was to compare the results with the most commonly used delay models (*11*) and make adjustments in the model parameters if necessary. All approaches were studied on weekdays during one of the following peak periods: a.m. (7:00 to 8:30 a.m.), noon (12:00 to 1:00 p.m.), and p.m. (4:30 to 6:00 p.m.). One hour of continuous video recording of the approaches of interest was made at each site. Depending on the number of approaches under study at each intersection, one or two video cameras were used. Available sight distance was also determined at all approaches that had potentially restricted sight distance.

The data reduction from the videotapes consisted of volume counts, vehicle mix, signal timing, and conflicts involving left-turning vehicles. Of particular interest to this study was the left-turning traffic mix, which is believed to affect left-turn operations. The average cycle length was used for those intersections with actuated signals. This approach converts actuated signal settings to pretimed signal settings; it also reduces variability in data and the need to collect cycle-by-cycle signal information. The calculated stopped delays, obtained with the Highway Capacity Software (*12*), signalized intersection program, were compared with the field-measured (observed) stopped delay to assess the correspondence between the two, and any subsequent adjustments necessary were made. A plot of the observed and calculated left-turn stopped delays for the PTO phase is shown in Figure 2, along with the 95 percent confidence bands for the mean for each observation

TABLE 1 Number of Study Intersections by County

County	Number of Intersections	Approaches Studied	Study Date
Bexar	8	16	March 1991
Cameron	7	14	January 1991
Dallas	20	29	June/July 1991
Ector	5	10	April 1991
Harris	12	24	January 1991
Lubbock	5	10	April 1991
Nueces	6	12	February 1991
Tarrant	33	55	June/Oct. 1991
Travis	12	24	November 1991
Total	108	194	

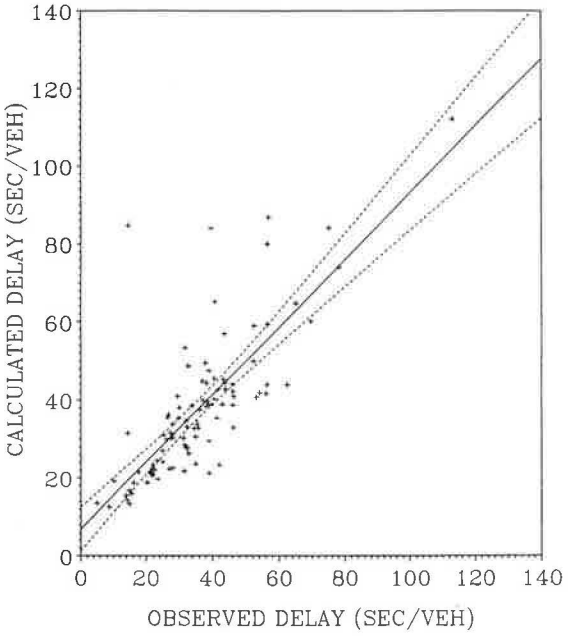


FIGURE 2 Observed and calculated left-turn stopped delays for PTO phase.

(*dashed lines*). The parameter estimates are presented in Table 2. From the hypothesis tests (Table 2), the slopes (β_1) do not significantly differ from 1. Likewise, the intercepts (β_0) do not significantly differ from zero. The observed and calculated delays were in good statistical agreement, as indicated by the parameter estimates and the hypothesis tests. The intercept for the PTO phasings was, however, slightly above zero. The *Highway Capacity Manual* (*11*) model was therefore used to estimate the stopped delays.

ACCIDENT STUDIES

To assess the safety aspects of the various phasing treatments, 42 intersections were selected for accident studies. Accident records for those intersections were obtained from the Texas Department of Transportation (TxDOT) and the transportation departments of the cities of Arlington, Dallas, and Fort Worth. The TxDOT records represent three successive years from 1988 through 1990, and the remaining records are for 1989 through 1991. Signal timings, phasing, and geometric history of the intersections involved were also obtained from the city transportation departments to ensure that no major geometric or signalization changes had taken place that might have influenced accidents. Only accidents involving left turns were extracted for analysis. Accident totals rather than rates were used in the study because of their availability and ease of application. Accident rates tend to be biased, with low-volume approaches having higher rates and vice versa. Another problem with using rates is measurement of exposure, since both left-turn and opposing-traffic volumes can influence the number of accidents. In addition, traffic volume data were for only one peak hour.

The average number of left-turn accidents for the various phases and sequences is shown in Table 3. Aside from phasing patterns, a number of other factors, such as intersection geometry, traffic volume, and weather conditions, affect left-turn accidents and could account for the large standard deviations observed. However, on average, P/P approaches have significantly higher left-turn accident totals ($\alpha = 0.10$) than PTO approaches. Higher left-turn accident rates at P/P approaches have also been reported by Agent (*4*) and Upchurch (*7*). The low left-turn accident totals for PMO approaches stem from the fact that they are generally low-volume intersections and were not problematic; otherwise, they would have likely been corrected by providing some form of left-turn protection.

The safety of leading and lagging sequences, as measured through the number of accidents, is not significantly different (see Table 3). However, leading sequences are likely to have higher left-turn accident totals, as indicated by their large variance (although an *F*-test showed the variances not to be significantly different in this case). Hummer et al. (*13*) have also reported higher left-turn accidents for approaches with leading as compared with lagging sequences. However, Lee et al. (*8*) found no significant difference between leading and lagging operations.

The focus of the accident analysis was to establish what constitutes an excessive number of left-turn accidents for a P/P phasing treatment. The 85th-percentile accident numbers were selected as a criterion. The 85th-percentile, 3-year, left-turn accident total was eight for P/P approaches and six for PTO approaches. PTO phases also had smaller variance com-

TABLE 2 Estimates and Tests of Hypotheses for Intercepts and Slopes of the Phase Patterns

Phase Pattern	Parameter	Estimate	Std. Error	R^2	Hypothesis	P-value
Protected Only	β_o	6.60	2.860	0.59	Ho : $\beta_o = 0$	0.02*
	β_1	0.86	0.073		Ho : $\beta_1 = 1$	0.07
Protected/Permissive	β_o	-2.05	1.570	0.86	Ho : $\beta_o = 0$	0.20
	β_1	0.96	0.053		Ho : $\beta_1 = 1$	0.47
Permissive Only	β_o	-1.65	2.100	0.76	Ho : $\beta_o = 0$	0.44
	β_1	0.93	0.100		Ho : $\beta_1 = 1$	0.51
All Phases (pooled)	β_o	0.34	1.427	0.75	Ho : $\beta_o = 0$	0.81
	β_1	0.98	0.042		Ho : $\beta_1 = 1$	0.58

* Significant at $\alpha = 0.05$.

TABLE 3 Three-Year Left-Turn Accident Totals for Left-Turn Phasing Types and Sequences

Phase Pattern	Mean Value	Standard Deviation	85 Percentile Value	Number of Approaches
Protected Only	2.57	3.16	5.73	77
Protected/Permissive	3.69	3.96	7.65	36
Dallas Phasing	2.92	4.60	7.52	18
Permissive Only	1.27	1.66	2.93	26
Leading Sequence	2.9	3.50	6.40	102[a]
Lagging Sequence	2.8	2.90	5.70	11[a]

[a] These approaches are the same as those listed in the first 2 rows of table.

pared with P/P treatments (Table 3). The 85th percentile was selected because it represents the mean plus one standard deviation, a commonly used level of confidence.

CONFLICTS

Conflict analysis can also be a powerful tool in determining the relative safety of intersections. Many latent risk factors at intersections are not reflected in accident records and can only be identified through conflict studies. Many left-turn conflicts of the near-miss type do not result in accidents and hence are not recorded. The left-turn conflict rate (C_{lt}) was determined from videotapes. The elapsed time from the start of each study period to the occurrence of the first conflict was used as a surrogate variable to determine the conflict rate. Where no conflict was observed during the 1-hr observation period, a time value of 60 min was used. The left-turn conflict rate is then calculated as

$$C_{lt} = \frac{\left(\frac{60}{T} \times 10^6\right)}{\left(\frac{V_{lt}}{N_{lt}} \times \frac{V_{op}}{N_{op}}\right)} \quad (1)$$

where

C_{lt} = number of conflicts per million [vehicles per hour (vph)/lane]^2,
N_{lt} = number of left-turn lanes,
N_{op} = number of opposing lanes,
V_{lt} = left-turn volume (vph),
V_{op} = opposing volume (vph), and
T = time to first conflict (min).

The four left-turn conflict types used in this study are

Type 1. Left-turn vehicle causing the opposing vehicle to brake or weave to avoid collision,

Type 2. The second through vehicle in the opposing path also having to take an evasive action,

Type 3. Vehicles entering the intersection on the green or yellow and turning left on the red, and

Type 4. Rear-end conflict in the left-turn lane.

The conflict study results are shown in Table 4. It can be seen that the lagging sequences have a significantly lower conflict rate ($\alpha = 0.04$). This was not clearly evident in the accident data analysis as reported in Table 3. The conflict studies do confirm the inference made from the accident studies that lagging sequences are safer than leading sequences. As shown in Figure 3, conflicts at P/P approaches with more

TABLE 4 Conflict Rates in Conflicts per Million (vehicles per hour per lane)^2 for Left-Turn Phasing Types and Sequences

Phase Pattern	Mean Value	Standard Deviation	85 Percentile Value	Number of Approaches
Protected Only	116	146	262	62
Protected/Permissive	176	272	448	47
Dallas Phasing	161	152	313	10
Permissive Only	914	1130	2044	36
Leading Sequence	156	230	386	86[a]
Lagging Sequence	90	101	191	23[a]

[a] These approaches are the same as those listed in the first 2 rows of table.

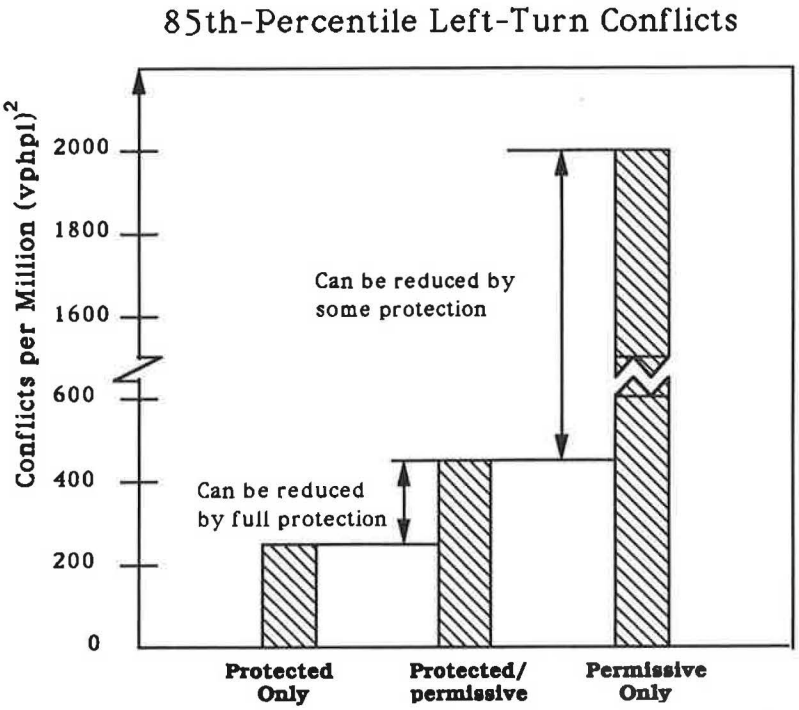


FIGURE 3 Summary of conflict studies.

than 260 conflicts per million (vph/lane)² could be reduced by providing full protection. Likewise, conflicts at PMO approaches with more than 450 conflicts per million (vph/lane)² could also be reduced by providing some form of protection. For more than 190 conflicts per million (vph/lane)², the lagging sequence can be used to improve the safety of the approach.

STOPPED DELAY

Delay is the direct result of interaction between the set of design variables and the decision variables. The objective is to find a set of design variables that minimizes delay without reducing safety. P/P approaches are generally associated with lower left-turn delays compared with PTO approaches. The average left-turn stopped delay for the P/P phasing was 20.3 sec/vehicle compared with 37.7 sec/vehicle for PTO phasing. The statistical test showed the difference to be significant at the 0.05 level. This observation is somewhat intuitive, as P/P phasing allows the left-turn traffic to filter through gaps in the opposing traffic. Lower left-turn stopped delays were also observed for leading (29.3 sec/vehicle) as compared with lagging (39.4 sec/vehicle) sequences. The Dallas phasing treatment also results in significantly ($\alpha = 0.05$) less left-turn stopped delay for the leading than the lagging sequence. The average left-turn stopped delay for the leading and lagging sequences of Dallas phasing were 29.3 and 36.0 sec/vehicle, respectively. However, the operational efficiency of the Dallas phase is heavily dependent on the magnitude of the opposing traffic [Collins (*14*)].

DATA ANALYSIS

To develop a selection criterion based on a systematic procedure that can be easily followed in practice, the data were classified into three main groups: design variables, decision variables, and measures of effectiveness.

- *Decision variables* are those variables that affect the intersection performance. Those selected for the study were left-turn volume (V_{lt}), opposing volume (V_{op}), volume cross-product (V_{xp}), vehicle mix (M_{ix}), ratio of green to cycle length (g/C) for left turns, number of left-turn lanes (N_{lt}), number of opposing lanes (N_{op}), volume-to-capacity ratio (v/c) of the approach, speed of opposing traffic (S_{op}), and sight distance (D_{iff}).
- *Design variables* are those aspects of the left-turn signal treatment for which guidelines are to be developed, such as the type of left-turn phasing, the sequence of the indications, and the type of auxiliary left-turn signs, if needed. A set of these design variables must be selected so that they optimize the intersection operation as a whole, taking into consideration delay and safety.
- *Measures of effectiveness* (MOEs) are those variables through which the performance of the signal phasing and indication sequences is assessed. The MOEs include left-turn and through stopped delay, intersection stopped delay, left-turn accidents, and left-turn conflicts.

A correlation analysis to minimize multicollinearity effects in modeling was conducted by examining the pairwise cor-

relation among the individual decision variables. The analysis showed no definite trend among any of the decision variables. Similar conclusions were also previously reported by Agent (*3*) and Upchurch (*15*).

STATISTICAL ANALYSIS OF DECISION VARIABLES

The statistical analysis addressed a number of decisions regarding the suitability of the left-turn phasing treatment to be used:

- Is a PMO left-turn phase adequate or should left-turn protection (green arrow) be provided?
- If left-turn protection is called for, is a more restrictive PTO phase justified or would the P/P pattern suffice?
- If the P/P phase is prescribed, would a leading operation be sufficient or should a lagging, leading/lagging, or Dallas phasing sequence be provided?

Permissive Versus Some Protection

A probabilistic approach using logistic regression was adopted to address this issue. This approach uses characterizing variables that distinguish one phase type from another. Given a set of decision variables, probability values are associated with the suitability of each phase type. The logistic model (*16*) is

$$U(P) = \log_e\left(\frac{P}{1 - P}\right) = \beta_0 + \beta_1 X_1 + \ldots \quad (2)$$

$$P(\phi) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots}} \quad (3)$$

where

$U(P)$ = utility function associated with a set of decision variables (X_i),
β_0, β_1 = model parameters,
ϕ = phase type, and
$P(\phi)$ = probability of selecting phase type ϕ.

Three decision variables were significant in differentiating permissive phasings from those with some protection, namely, left-turn volume, speed limit of the opposing approach, and number of opposing lanes. The maximum likelihood estimates of the parameters in Equation 2 and their standard errors shown in parentheses are

β_0	β_1	β_2	β_3
−5.100	0.705	0.024	0.085
(1.79)	(0.34)	(0.01)	(0.05)

The coefficients β_1, β_2, and β_3 are associated with the decision variables N_{op}, V_{lt}, and S_{op}, respectively. All the parameter estimates were statistically significant at the 0.10 level. Since classification is very sensitive to the relative sizes of the two components being classified and always favors classification into the larger group [Hosmer and Lemeshow (*17*)], a cutoff point of 0.7 was selected for classification to account for the unbalanced data. The positive coefficients for the decision variables indicate preference for some protection for higher values. Using the cutoff point of 0.7, the corresponding utility value [$U(P) = 0.85$] is obtained, which yields the following indifference lines:

$$V_{lt} = 220 - 3.54\,(S_{op}) \quad \text{for } N_{op} = 1$$

$$V_{lt} = 190 - 3.54\,(S_{op}) \quad \text{for } N_{op} = 2$$

$$V_{lt} = 160 - 3.54\,(S_{op}) \quad \text{for } N_{op} = 3$$

Figure 4 shows the plots of left-turn volume versus speed limit on the opposing approach for one, two, and three opposing lanes, respectively. The lower portion (*shaded*) indicates preference for PMO operation, and the upper portion (*unshaded*) signifies the need for some protection.

Protected Only Versus Protected/Permissive

The PTO and P/P phase types display similar characteristics in terms of the decision variables, making the logistic approach ineffective. The analysis approach consists of setting threshold values that could be used as a distinction criterion between the two forms of protective phasings. Threshold values were determined by establishing the 85th-percentile values of each decision variable for the P/P phasings. The number of approaches under each of these two phasing types that meet the threshold values and the percentage of these approaches that have PTO phasing are presented in Table 5. In determining the conditions under which PTO phasing is recommended, all pairwise combinations of the decision variables for which 80 percent or more of the approaches studied had PTO phasings were identified (Table 6).

A sensitivity analysis identified the 80th percentile as the point of diminishing returns for the selection of conditions under which PTO phasing is recommended. The analysis indicated that if a percentile lower than the 80th is used, very few additional conditions for PTO phasings will be added. For example, a 70th- or 75th-percentile cutoff will result in only one added condition for PTO phasing, that is, more than two left-turn lanes. On the other hand, considering a higher value than the 80th percentile will exclude a large number of conditions for which PTO phasing should be recommended. For example, considering the 85th percentile will exclude four of the eight conditions identified under the 80th-percentile criterion, including sight distance, for which PTO phasing should definitely be considered. With the exception of sight distance, two or more conditions are required to justify the use of PTO phasing. The recommended guidelines are presented in the section on summary of guidelines.

Leading or Lagging or Leading/Lagging Phase Pattern

Leading sequences are, from an efficiency standpoint, more desirable since they are associated with lower delays and increased intersection capacity. Lagging sequences, on the other hand, appear to be safer compared with leading sequences. Leading/lagging operation may be implemented for reasons

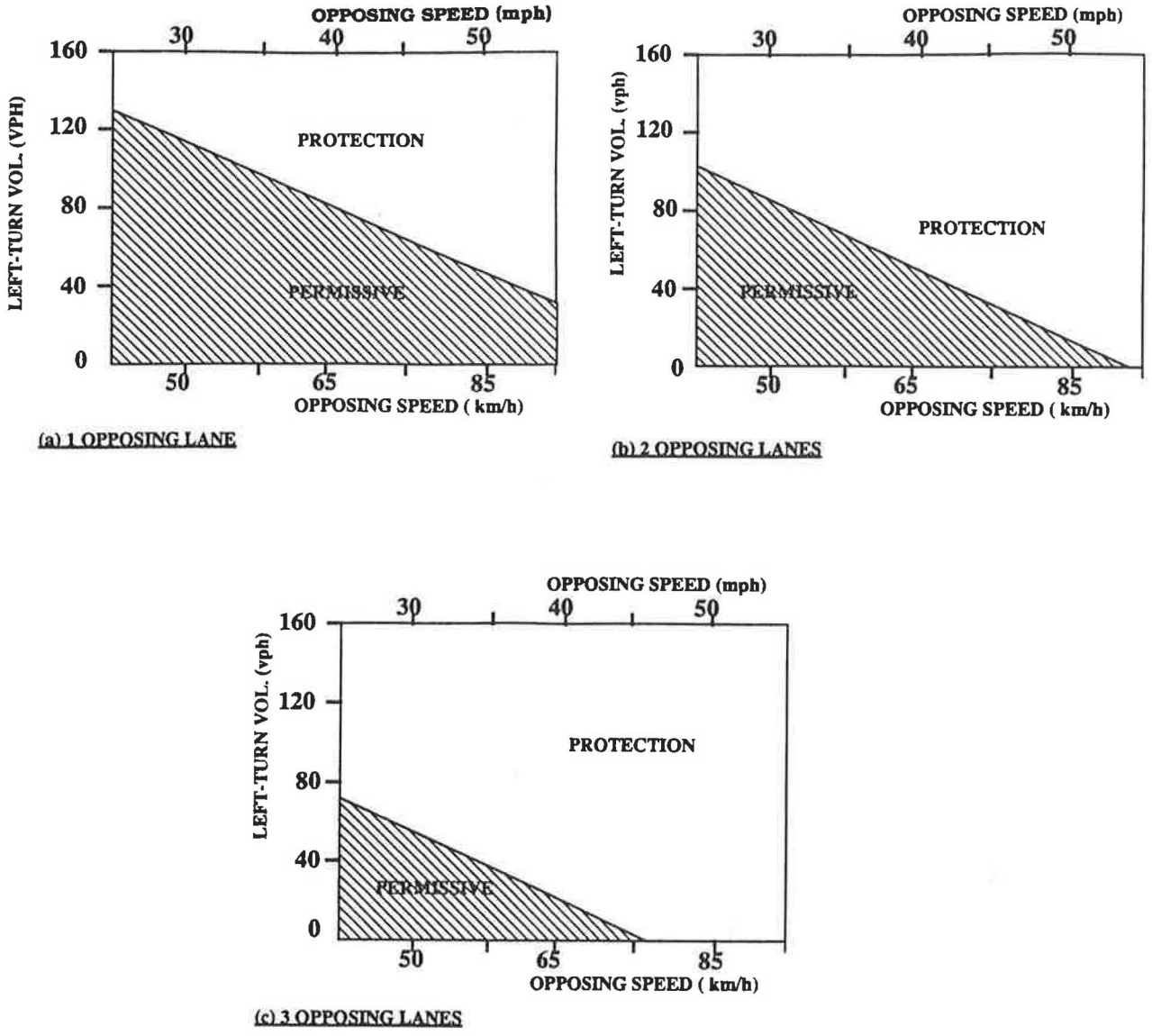


FIGURE 4 Selection guide for the choice between PMO versus some left-turn protection when left-turning vehicles face one, two, and three opposing lanes.

TABLE 5 Number and Percent of Approaches Satisfying the Threshold Values

Decision Variable	Threshold Value	Number of P/P+	Number of PTO+	Percent PTO
Left-turn volume (V_{lt})	> 320 vph	11	19	65
Opposing volume (V_{op})	> 1100 vph	11	27	70
Left-turn lanes (N_{lt})	≥ 2	12	30	70
Opposing lanes (N_{op})	≥ 3	25	33	60
Sight distance (D_{lft}) *	< 0	5	19	80
Left-turn Mix (M_{lx})	> 2.5%	6	10	65
Opposing speed limit (S_{op})	≥ 75 km/h	12	36	75
Volume cross product (V_{xp})	> 250,000 vph²	7	9	55

+(PTO = Protected only, P/P = Protected/Permissive)

*The difference between available and required sight distance based on opposing speed limit.

1 km = 0.6 mi

TABLE 6 Number and Percent of Approaches That Meet Pairwise Threshold Combinations and Have PTO Phasing

Pairwise Combination	Number of P/P	Number of PTO	Percent of PTO
V_{lt} > 320 vph & V_{op} > 1100 vph	2	7	80
V_{lt} > 320 vph & S_{op} ≥ 75 km/h	2	10	80
V_{lt} > 320 vph & M_{lt} > 2.5 %	0	3	100
V_{op} > 1100 vph & S_{op} ≥ 75 km/h	2	18	90
V_{op} > 1100 vph & M_{lt} > 2.5 %	0	2	100
N_{lt} ≥ 2 & S_{op} ≥ 75 km/h	4	16	80
N_{op} = 3 & S_{op} ≥ 75 km/h	1	10	90

1 km = 0.6 mi

other than the local reduction in delay. Arterial progression is often greatly improved by the use of leading/lagging operation, since the through green times are not constrained to occur at the same time. In addition, an intersection may not be wide enough to accommodate dual left turns, particularly if at least one of the approaches has two left-turn lanes. In this case, a leading/lagging phase pattern would be necessary in order to provide turning in both directions.

SUMMARY OF GUIDELINES

An engineering study of the following site conditions and characteristics is required to determine the appropriate left-turn signal treatment: traffic volumes, traffic mix, intersection geometry, sight distance, delays, speed of traffic, accidents, conflicts, and traffic progression scheme. The guidelines are based on threshold values designed to determine what constitutes an excessive value of any particular variable, beyond which a specific left-turn phasing treatment can be justified. The guidelines were developed for intersections with little or no pedestrian traffic and with exclusive lanes for protected left turns. The decisions to be made are classified into three levels, discussed in the following sections.

Level 1: PMO Versus Some Protection

Level 1 is intended for application where the decision entails determining whether a PMO phase is appropriate or some protection (green arrow) is necessary. It is recommended that PMO phasing be replaced by phasing with some left-turn protection when any one of the following conditions exists:

- The plotted point representing the peak-period volume in vehicles per hour (based on the peak 15 min) and the corresponding opposing-traffic speed limit fall above the curve (unshaded portion) in Figure 4 for the existing number of opposing lanes,
- The sight distance for the left-turning vehicle is restricted on the basis of the posted speed limit for the opposing traffic (in such cases, full protection is recommended),
- More than eight left-turn-related accidents have occurred within the last 3 years at any one approach with PMO phasing, and
- More than 450 left-turn-related conflicts per million (vph/lane)^2 are observed at an approach with PMO phasing.

Level 2: P/P Versus PTO Phasings

Once the decision has been made to provide some left-turn protection, it must be determined whether P/P phasing would suffice or whether a more restrictive PTO phasing should be prescribed. If possible, the more efficient P/P phasing should be used unless PTO phasing is absolutely necessary. PTO phasing is recommended under any of the following conditions:

- Approaches with restricted sight distance, as determined on the basis of posted speed limit on the approach opposing the left-turn traffic, or
- Approaches with four or more opposing lanes that must be crossed by the left-turning traffic;

or on any two of the following conditions:

- Peak-hour volume measured at 15-min intervals for the left-turning traffic greater than 320 vph,
- Peak-hour volume measured at 15-min intervals for the opposing traffic greater than 1,100 vph,
- Opposing speed limit greater than or equal to 75 km/hr (45 mph), or
- Two or more left-turn lanes;

or when one of the following conditions or combinations of conditions exist:

- Three opposing lanes and opposing speed limit greater than or equal to 75 km/hr (45 mph), or
- Left-turn volume greater than 320 vph and percent of heavy vehicles in the left-turning traffic exceeding 2.5 percent, or
- Opposing volume greater than 1,100 vph and percent of heavy vehicles in the left-turning traffic exceeding 2.5 percent, or
- Seven or more left-turn-related accidents within a 3-year period for a P/P approach, or

- More than 260 left-turn-related conflicts per million (vph/lane)^2 observed for a P/P approach, or
- Average stopped delay to left-turning traffic acceptable (i.e., within the desired level of service) for PTO phasing and traffic engineer judges that the use of P/P phasing will result in a greater number of left-turn accidents.

Level 3: Sequence of Phasing—Leading, Lagging, or Leading/Lagging?

Level 3 is for the selection of the appropriate phasing sequence once the phase type to be used has been determined.

- A leading sequence is recommended when P/P or PTO phasing has been determined to be suitable under a Level 1 or 2 decision as outlined above, provided that it will not disrupt any progression scheme on either street;
- In regions where Dallas phasing is a viable option, it is recommended when P/P phasing has been determined to suffice but the resulting level of service is not acceptable;
- A lagging sequence is recommended when
  - –It is intended to improve the safety of an already installed leading sequence under which more than 190 left-turn conflicts per million (vehicles per hour per lane)^2 are observed, or
  - –The lagging left-turn sequence is necessary as part of an overall network progression scheme;
- A leading/lagging sequence is recommended for intersections when
  - –There is inadequate space within the intersection to safely accommodate a dual left-turn operation, or
  - –It is necessary for the progression scheme.

DISCUSSION OF RESULTS

The guidelines developed provide a simple three-level procedure to aid in the selection of the appropriate left-turn signal treatment. The data requirements for each decision level are different, reflecting the different objectives to be dealt with at each level. These guidelines also reflect a selection process that recognizes the trade-off between operational efficiency and safety. In some cases, more than one condition is required to justify the selection of a particular phase so as to ensure an optimum solution.

It must be noted that intersections without exclusive left-turn lanes have not been included in this study, and therefore the proposed guidelines are not applicable to this geometric condition. In general, P/P operation is not appropriate for shared-lane operations. Split phasing is often used when protection is deemed necessary. An aspect of phase type selection overlooked in this study is the inclusion of pedestrian volume as a decision variable. When left-turn protection is required, traffic engineers often favor a PTO phase at intersections with heavy pedestrian volume because permissive left turns are often confusing and unanticipated by pedestrians, particularly when they cross the intersection using signal indications based on vehicular traffic. Permissive phasings are also undesirable for bicyclists, who require a considerably larger gap in the opposing traffic for permissive turns. Threshold values for pedestrian and bicycle volumes and accidents would have to be statistically established beyond which a PTO phasing could be recommended.

ACKNOWLEDGMENTS

This research has been sponsored by the Texas Department of Transportation (TxDOT) in cooperation with the Federal Highway Administration, U.S. Department of Transportation. The authors would particularly like to thank Joni Brookes of the Colorado Department of Transportation, formerly the project technical coordinator with TxDOT, for her assistance and suggestions. They would also like to thank Ray Derr of TxDOT, Robert Musselman, and Davey Warren of FHWA for their comments.

REFERENCES

1. Asante, S. A., S. A. Ardekani, and J. C. Williams. *Selection Criteria for Left-Turn Phasing, Indication Sequence, and Auxiliary Signs*. Research Report 1256-1F. Transportation Research, University of Texas at Arlington, Feb. 1993.
2. de Camp, G., and R. W. Denney, Jr. Improved Protected-Permitted Left-Turn Signal Displays—The Texas Approach. *ITE Journal*, Vol. 62, No. 10, Oct. 1992, pp. 21–24.
3. Agent, K. R., and R. C. Deen. Warrants for Left-Turn Signal Phasing. In *Transportation Research Record 737*, TRB, National Research Council, Washington, D.C., 1979, pp. 1–10.
4. Agent, K. R. *Guidelines for the Use of Protected/Permissive Left-Turn Phasing*. Research Report UKTRP-85-19. Kentucky Transportation Research Program, University of Kentucky, Knoxville, Aug. 1985.
5. Machemehl, R. B., and H. J. Lin. Guidelines for Left-Turn Lanes and Signal Phases. In *Compendium of Technical Papers*, Institute of Transportation Engineers, Washington, D.C., Aug. 1982, pp. 1–8.
6. Machemehl, R. B., and A. M. Mechler. *Comparative Analysis of Left-Turn Phase Sequencing*. Research Report 258–2. Center for Transportation Research, University of Texas at Austin, Nov. 1983.
7. Upchurch, J. E. *Left-Turn Signal Warrants for Arizona*. Report FHWA/A2-85/192. Arizona Department of Transportation; FHWA, U.S. Department of Transportation, May 1985.
8. Lee, J. C., R. H. Wortman, D. J. P. Hook, and M. J. Poope. *Comparative Analysis of Leading and Lagging Left-Turns*. Report FWHA-AZ91-321. Arizona Department of Transportation; FHWA, U.S. Department of Transportation, 1991.
9. Florida Section, Institute of Transportation Engineers. Left-Turn Phase Design in Florida. *ITE Journal*, Sept. 1982, pp. 28–35.
10. Cottrell, B. H. Guidelines for Protected/Permissive Left-Turn Signal Phasing. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 54–61.
11. Signalized Intersections. In *Special Report 209: Highway Capacity Manual*, TRB, National Research Council, Washington, D.C., 1985.
12. *Highway Capacity Software: User's Manual*. Release 1.50, Federal Highway Administration, U.S. Department of Transportation, 1990.
13. Hummer, J. E., R. E. Montgomery, and C. S. Kumares. Guidelines for the Use of Leading and Lagging Left-Turn Signal Phasing. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991, pp. 11–20.
14. Collins, R. *A Comparative Analysis of Left-Turn Delay Associated with Two Different Lead/Lag Phasing Arrangement*. Master's thesis. University of Texas at Austin, 1988.
15. Upchurch, J. E. Guidelines for Selecting Type of Left-Turn Phas-

ing. In *Transportation Research Record 1069*, TRB, National Research Records, Washington, D.C., 1986, pp. 30–38.
16. Wrigley, N. *Categorical Data Analysis for Geographers and Environmental Scientists*. Longmans, London and New York, 1983.
17. Hosmer, D., and R. Lemeshow. *Applied Logistics Regression*. Wiley Series. Wiley, New York, 1989.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Analysis of Flashing Signal Operation

KENT C. KACIR, ROBERT J. BENZ, AND H. GENE HAWKINS, JR.

Flashing traffic signal operation can offer reduced delay over alternative modes of signal operation, such as pretimed and actuated. The research described in this paper was conducted as part of a study on flashing traffic signals. Significant research activities included a review of previous literature, a survey of current practice, an operational analysis of alternative modes of signal operation, and the analysis of traffic volumes during late-night low-volume periods. The literature review found few comprehensive guidelines, although there is evidence that substantial interest exists. The survey of current practice indicated that traffic engineers primarily rely on engineering judgment instead of standards or guidelines. The operational analysis determined that for low volumes, the flashing yellow/red operation will reduce total delay by 50 percent versus pretimed and actuated operation. In general, the red/red flash operation will produce the most amount of delay. Data collection efforts revealed that typically 2.5 percent of the average daily traffic (ADT) occurs during the period between midnight and 6:00 a.m. and that the hourly volume during this period ranges from 0.2 to 0.8 percent of the ADT.

Traffic signals provide a safe and effective means of controlling vehicular and pedestrian traffic at intersections. However, because they assign the right-of-way to the various traffic movements, traffic signals exert a profound influence on traffic flow. *The Manual on Uniform Traffic Control Devices* (MUTCD) (*1*) states that signals should not be installed unless one or more of the signal warrants are met. Two of the warrants are volume based. For an intersection to meet one of these warrants, traffic must be greater than a specified level for at least 8 hr of the day. Even when an intersection meets one of the warrants, there may be periods of time during the day when traffic volumes are below the warrant levels. During these low-volume periods, flashing signal operation is an alternative to normal (green-yellow-red) signal operation. The primary justification for flashing operation is that vehicular delay can be reduced by eliminating or reducing the number of stops.

STUDY ACTIVITIES

The findings presented in this paper were developed from several research activities, which included a review of literature on flashing operation, a survey of traffic engineers on the use of flashing traffic signals, an evaluation of low-volume traffic signal operations, and the analysis of late-night traffic volumes. A significant portion of this paper presents research results based on an operational evaluation. Traffic safety is an important factor to consider before a signal is placed into flashing operation. Because of the amount of accident data and statistical results required to adequately cover the safety issues, they cannot be presented here with the operational analysis. However, the literature review in this paper does address past research efforts in identifying the safety aspects of flashing operation.

Texas Transportation Institute, Texas A&M University System, College Station, Texas 77843-3135.

Literature Review

Probably the most important source of guidelines or standards on any traffic control device is *The Manual on Uniform Traffic Control Devices* (MUTCD). In the 1988 edition (*1*), the use of flashing signals is mentioned, but few guidelines are provided for implementing flashing operation. It is noted that earlier editions of the MUTCD provided more guidelines than does the 1988 edition. For example, in the 1935 edition (*2*) it is stated that "when the total vehicular volume entering an intersection having fixed-time signals falls below 500 vph for a period of two or more consecutive hours, the fixed-time signal shall be operated as flashing." With regard to actuated signal operation, the 1935 MUTCD states that "because actuated control adjusts itself to varying traffic volume and involves relatively little vehicular delay during light traffic, it is not necessary to change to flashing operation at any time."

Texas is one of the states that publishes its own edition of the MUTCD. The 1980 edition of the Texas MUTCD (*3*) contains the following guideline for implementing flashing operation:

> When for a period of four or more consecutive hours of the late evening and/or early morning periods, any traffic volume drops to 50 percent or less of the stated volume warrants, pretimed traffic control signals should be placed on flashing operation rather than continue normal operation.

Guidelines Based on Traffic Volume

The most significant research study of flashing traffic signals was conducted as part of an FHWA study of traffic signal operation (*4*). The FHWA study recommended using flashing yellow/red operation when the two-way traffic volumes on the major street are below 200 vph. Flashing yellow/red operation may also be used when the two-way major street volume is greater than 200 vph provided the ratio of major-street volume to minor-street volume is greater than 3.

Guidelines Based on Accidents

The safety aspect of flashing operation has been addressed in several different studies (*4–6*). The FHWA report (*4*) found

that, in general, flashing yellow/red operation increased the accident rate. The exception was at intersections with a high ratio of major-street to minor-street volume. Accident rates at these intersections were lower with flashing operation than with normal operation. Accidents, particularly right-angle accidents, were higher with flashing yellow/red than with normal operation.

The FHWA study (*4*) analyzed accident files from around the nation and found that flashing yellow/red operation, in general, significantly increased the hazard of driving at night. The major exceptions were intersections at which the volume ratio is equal to or greater than 3 or at which the major-street two-way volume is less than 200 vph during flashing operation. In addition, the study found that the most hazardous driving time was the first hour after drinking establishments closed.

The FHWA study recommended against using flashing yellow/red operation if the following conditions are met or exceeded at an intersection:

- Three right-angle accidents in one year during flashing operation (short-term rate).
- Two right-angle accidents per million vehicles during flashing operation if the rate is based on an average of three to six observed right-angle accidents per year (long-term rate), or
- 1.6 right-angle accidents per million vehicles during flashing operation if the rate is based on an average of six or more right-angle accidents per year (long-term rate).

Studies performed by local agencies in Portland, Oregon (*5*), and Oakland County, Michigan (*6*), agreed with the FHWA study finding indicating that right-angle accidents occur at significantly higher rates when flashing yellow/red signal operation is used. The Portland study recommended the use of flashing operation for low-volume conditions when the volume ratio is less than or equal to 2. The Michigan study recommended removing flashing operation if the volume ratio is 4 or less.

A follow-up to the Michigan study (*7*) analyzed traffic accident data at 59 intersections that were changed from flashing signal operation to 24-hr normal operation. The results of this study indicated that changing signal operation from flashing to normal operation was effective in reducing the frequency of total right-angle and personal injury right-angle accidents during nighttime hours. However, there was no noticeable change in the frequency of total rear-end or personal injury rear-end collisions. The follow-up study reconfirmed the findings of the original study and no new recommendations or findings were added.

Survey of Flashing Practice

Because of the limited amount of previous research on flashing operation, most traffic engineers make decisions related to flashing operation on the basis of engineering judgment or field experience. Therefore, a survey was developed to identify flashing signal practice in Texas. The survey provided the opportunities to gather information about many different aspects of flashing operation and also to assist the research team in their data collection efforts. Recognizing these opportunities, the research team established two objectives for the survey of flashing signal operation: (*a*) to identify where and how flashing operation is currently utilized on a regular basis in Texas and (*b*) to determine the guidelines or warrants that agencies use to implement flashing signal operation.

Survey Methodology

The survey was developed so that it could be sent to many agencies and quickly answered by the traffic engineering personnel at the various agencies. Some of the more significant questions and their responses are shown in Table 1. Surveys

TABLE 1 Survey Questions and Responses

Question	Responses
Use flashing operation on a regular basis.	70%-Yes 30%-No
Conditions when flashing operation is used.	96%-Emergency (due to signal failure) 68%-Signal installation and/or removal 66%-Early morning hours 55%-Railroad preemption 21%-Low-volume periods other than early morning 17%-School areas
Factors addressed in flashing guidelines.	49%-Traffic volume 47%-No guidelines 40%-Time of day 23%-Accidents 19%-Other
Usefulness of guidelines for flashing operation.	59%-Useful 32%-Might be useful 9%-Not useful
Use flashing operation with actuated controllers.	47%-Use 11%-Sometimes use 38%-Do not use
Analysis of flashing operation.	8%-Have performed an analysis 92%-Have not performed an analysis
Basis for selecting flashing indications.	Volume ratio Consistency

were received from 24 Texas Department of Transportation (TxDOT) districts and 23 city transportation departments in Texas.

Survey Results

The survey results indicated that 70 percent of the respondents use flashing operation on a regular (or normal) basis. Normal operation includes all types of flashing other than emergency or railroad preemption flashing. Among the choices (multiple-choice format) for regular flashing operation, early morning flashing and signal installation and removal flashing were the two most common uses; both were used by approximately two-thirds of the respondents. Flashing during early morning hours was more common among the local agencies (74 percent) than the TxDOT districts (58 percent), and flashing for signal installation and removal was more common with the districts (75 percent) than with the local agencies (61 percent).

Flashing in an emergency situation related to signal failure is used by virtually all the respondents (96 percent). Railroad preemption flashing is also commonly used (55 percent). There was a fairly significant difference between the districts and the local agencies in the use of flashing with railroad preemption (71 versus 39 percent). This difference seems logical when one considers that railroad preemption as used by the local agencies probably displays a green indication to the nonconflicting movement.

Almost half (47 percent) of all the respondents indicated that they have no formal guidelines for implementing flashing operation. Among those agencies that do have guidelines, the specific factors that are considered in a decision to implement flashing operation are traffic volume (49 percent), time of day (40 percent), and accidents (23 percent). Other factors were also identified but are used to a lesser extent: day of the week, relationship to other intersections, geometrics, posted speeds, weather effects, and type of signal operation.

The survey also addressed operating an actuated signal in the flashing mode. This question was added to the survey because of conflicting opinions about whether flashing operation is appropriate at intersections with actuated signals. Intuitively, it would seem that flashing operation would not be needed at such an intersection because of the ability of an actuated signal to respond to traffic demand. However, the responses to this question indicate that there is greater use of flashing with actuated controllers than originally thought. One possible explanation for the large number of responses for use of flashing with actuated control is that the flashing might be limited to certain applications resulting from emergencies, conflicts, maintenance, preemption, and installation and repair. However, conversations with some of the survey respondents indicated that flashing an actuated signal during low-volume periods is not unusual.

The basis on which traffic engineers select the mode of flashing operation (yellow/red or red/red) varied among respondents. This can be attributed largely to the lack of any formal guidelines. One of the biggest concerns of the survey participants was the potential for confusing drivers by displaying two different modes of flashing operation at a single intersection during different times of the day. Typically this would involve using a yellow/red flash for normal operation and a red/red flash for conflict operation. Some of the common responses to the question on selecting the mode of flashing were as follows:

- If the intersection traffic volume is close to being equal on all approaches, flash red to all approaches. If the intersection traffic volume is not equal but generally greater than 2:1, flash yellow on the major street and red on the minor street.
- The normal flash is yellow/red, and the conflict or emergency flash is red/red.
- The red/red flash is used for all occasions (normal and emergency).
- Flashing operation is not used at all, to avoid driver confusion. High accident rates have been linked with motorist confusion associated with the yellow/red flash.
- The red/red flash is used for all diamond interchanges.
- The red/red flash is used at intersections with sight distance restrictions or accident history.
- Other responses to the question of guidelines for use of flashing signals included consistency with other signals, speed, geometrics, accident history, arrangement of STOP signs before signal installation, and police input.

Operational Analysis

Traffic simulation models were used to compare flashing operation with normal operation. The total intersection delay for yellow/red flash, red/red flash, pretimed, and actuated signal operation were compared for isolated intersections and three-intersection signal systems.

Signal operation at isolated intersections was simulated using the TEXAS and TRAF-NETSIM models. Although the TEXAS model can simulate all four types of signal operation, NETSIM does not have the capability to model a four-way stop-controlled intersection because it cannot model red/red flashing operation. The latest version of the TEXAS model, which has been improved to provide more accurate representation of delay at four-way stop intersections, was used in the simulation (*8*). Earlier versions overestimated delay at four-way stop-controlled intersections (*9*). Operation of a signal system was simulated using NETSIM, because the TEXAS model cannot simulate more than one intersection.

Because both models are stochastic, replicate runs using different random number seeds were made. As a minimum, five runs were made for each scenario used in the analysis. Afterwards, the average total delay per vehicle was computed and used in the analysis.

Simulation Assumptions

In order to simplify the analysis, the study team identified a basic intersection geometric configuration associated with common or fundamental signal operation, or both. In most cases, assumptions were established that would simplify the operation and thus provide a clear understanding of the fundamental relationship between the alternative modes of signal operation. For the selected traffic volumes modeled, an optimal signal timing was developed. Therefore, the delay data

are reflective of the signal operation (i.e., normal versus flashing operation) and not a misallocation of effective green time. The general assumptions are given in Table 2.

Results of Operational Analysis

The various types of signal operation at an isolated intersection were analyzed using both the TEXAS and NETSIM models, and the analysis of a signal system was performed with the NETSIM model. Total intersection delay was used as the basis of comparison and the impact of volume was normalized by using the ratio of major- to minor-street approach volume. The presentation of the results was simplified by dividing the analysis into three groups of major-street approach volumes: less than 125 vph, 250 to 500 vph, and 750 to 900 vph.

Isolated Intersection Analysis The analysis results from the two models were basically similar. However, because NETSIM cannot analyze red/red flashing operation, only the results of the TEXAS model are presented here.

The results show a clear relationship between the volume ratio and total delay. Typically, as the traffic volume on the major arterial increases, the total delay increases. This trend is consistent with fundamental practice and understanding of traffic signal operation. The minor-street volume does not influence the total delay if the volume ratio is greater than 3.

Figures 1 through 3 show the results from the TEXAS model for major-street volumes of less than 125 vph, 250 to 500 vph, and 750 to 900 vph, respectively. It can be seen that, in general, flashing yellow/red signal operation produces the lowest amount of delay, followed by actuated, pretimed, and then flashing red/red signal operation. The only exception to this trend is found in Figure 3 for equal traffic volumes on the major and minor streets. At this point, the delay for flashing yellow/red increases asymptotically and parallel to the 1.0 volume ratio. At this volume ratio, normal signal operation produces the least total delay.

Red/Red Flashing Operation Red/red flashing produces the highest total intersection delay for all volume ratios greater than 3. For volume ratios less than 3, pretimed operation generally produces the highest total delay. For major arterial volumes less than 500 vph, the red/red flash curve is relatively flat. In other words, very little change in total delay results from a decrease in minor-street volume. For major-street volumes greater than 500 vph, the volume ratio must be less than 4 for the minor street to have an effect on the total intersection delay.

Yellow/Red Flashing Operation As stated previously, yellow/red flashing operation generally involves the least amount of total delay, producing less than 5 sec of delay per vehicle if the major-street volumes are less than 125 vph (Figure 1). When the major-street volumes are greater than 125 vph but less than 750 vph, the intersection delay is less than 6 sec (Figure 2). If the major-street volume is greater than 750 vph, total intersection delay is more than under pretimed or actuated operation and red/red flashing operation with balanced flow (Figure 3). Beyond a volume ratio of 3, the total delay drops below 5 sec per vehicle.

Pretimed and Actuated Operation For lower traffic volumes on the major arterial, actuated signal control produced approximately one-half the delay produced by pretimed signal operation. For higher traffic volume on the major arterial, actuated operation reduces delay by 3 or more seconds over pretimed operation.

TABLE 2 Assumptions for Operational Analysis

Geometric Assumptions
• 4-legged intersection with 5 lanes on major arterial and 4 lanes on minor street.
• Lane width of 3.66 meters.
• System of 3 intersections separated by 305 meters.
Operational Assumptions
• Turning movements: 10% left, 10% right, and 80% through vehicles.
• Traffic volume ranged from 25 to 900 vehicles per approach on both major and minor street.
• Travel speed of 48 kilometers per hour on both major and minor streets.
• Pretimed operation used a 3-phase leading-left phasing arrangement.
• The minimum cycle length was 40 seconds.
• Yellow clearance of 3 seconds and an all-red clearance of 1 second.
• The Cycle Capacity Probability Design Curve (10) used to determine phase clearance values.
• Offset values for the signal system determined by PASSER II-90 traffic operations optimization program.
• Actuated operation used the 3-phase operation developed for pretimed operation.

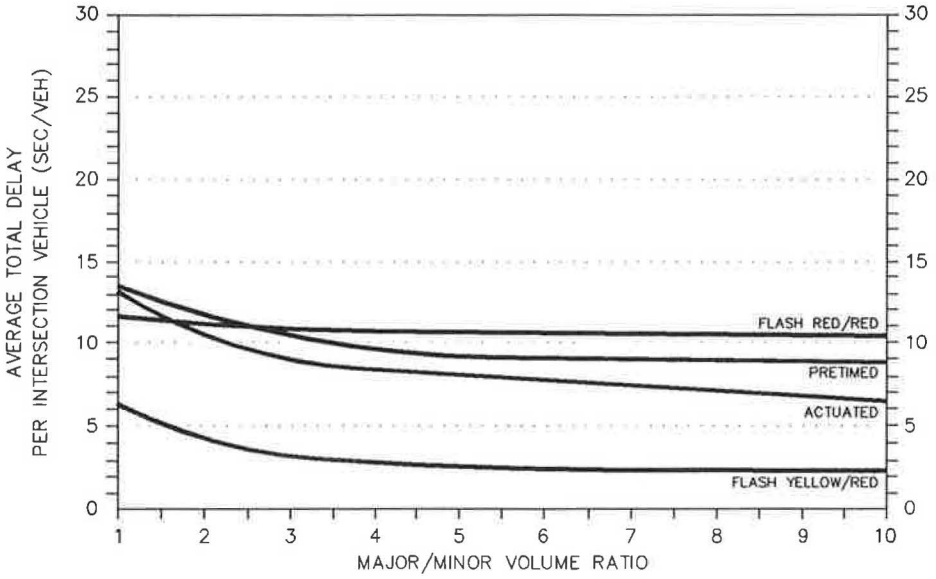


FIGURE 1 Isolated intersection operation with major-street volume less than 125 vph.

Summary of Isolated Intersection Analysis Results The following conclusions can be drawn from Figures 1 through 3:

- Yellow/red flashing operation produces the least amount of total intersection delay, followed by actuated, pretimed, and red/red flashing operation.
- Red/red flashing produces approximately 7 sec more than yellow/red flashing when the major-street volume is less than 125 vph and approximately 8 and 13 sec for major-street volume with less than 500 and 900 vph, respectively.
- Red/red flashing produces a relatively constant delay when the major-street volume is less than 750 vph. When the major-street volume is greater than 750 vph, the delay becomes constant at volume ratios greater than 4.
- If the major-street volume is less than 250 vph, total intersection delay can be reduced by one-half if pretimed operation is changed to actuated operation and reduced by approximately one-third if changed to yellow/red flashing.
- If the major-street volume is less than 750 vph, total intersection delay can be reduced by one-half if red/red flashing is changed to actuated signal operation and reduced again by one-half if actuated operation is changed to yellow/red flashing.
- If the major-street volume is greater than 750 vph, the same general trend is present as previously described, but only for volume ratios greater than 3.

Results of Signal System Analysis The signal system was analyzed using the NETSIM model. A clear relationship was shown between the volume ratio and total delay. Typically, pretimed, actuated, and yellow/red flashing operation pro-

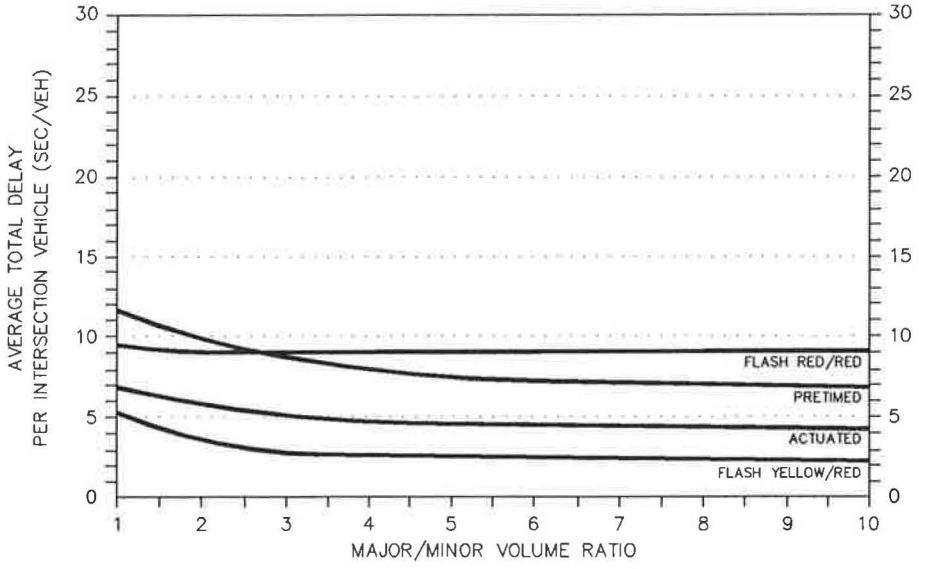


FIGURE 2 Isolated intersection operation with major-street volume between 250 and 500 vph.

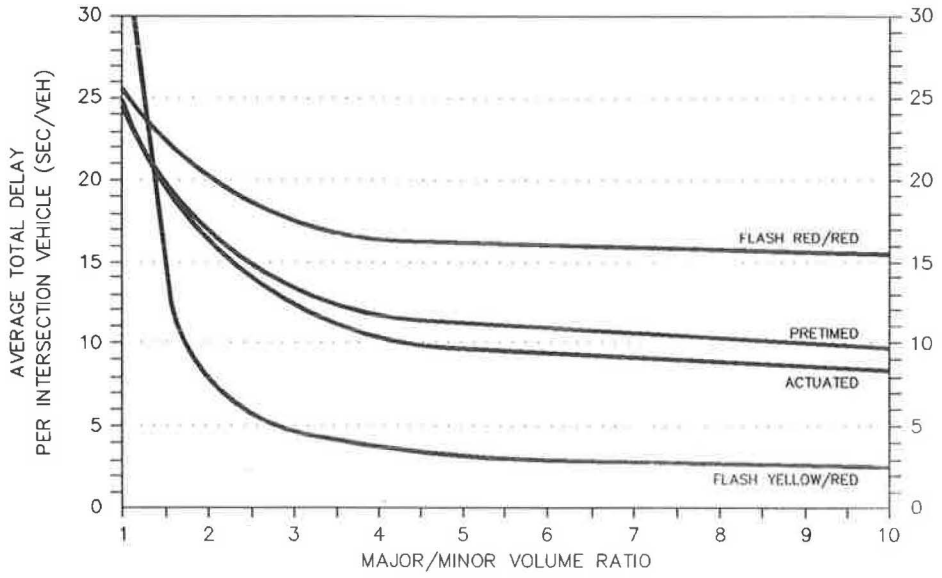


FIGURE 3 Isolated intersection operation with major-street volume between 750 and 900 vph.

duced 2 to 3 sec more delay than that produced in the isolated intersection simulation. Figures 4 through 6 show the results for major-street volumes of less than 125 vph, 250 to 500 vph, and 750 to 900 vph, respectively.

Yellow/Red Flashing Operation As found in the isolated intersection analysis, yellow/red flashing produced the least amount of total delay, less than 7 sec if the major-street volume is less than 125 vph. This is compared with 5 sec of delay produced for the isolated intersection for the same traffic volume. The added delay results from minor-street vehicles, which are required to stop for the flashing red signal. The major arterial continues progressive movement without any additional delay. As the major-street volume increases, there is added friction among the vehicles. Total delay increases with increasing major-street volume and increasing minor-street volumes for volume ratios less than 4 (Figures 4–6).

Red/Red Flashing Operation It was stated previously that the NETSIM model cannot simulate red/red flashing operation; therefore, no systemwide results are available for this type of operation. However, it seems unlikely that red/red flashing would be used in a signal system, so the lack of results for this case is not significant.

Pretimed and Actuated Operation Pretimed signal operation produced no more than 3 times the delay produced by

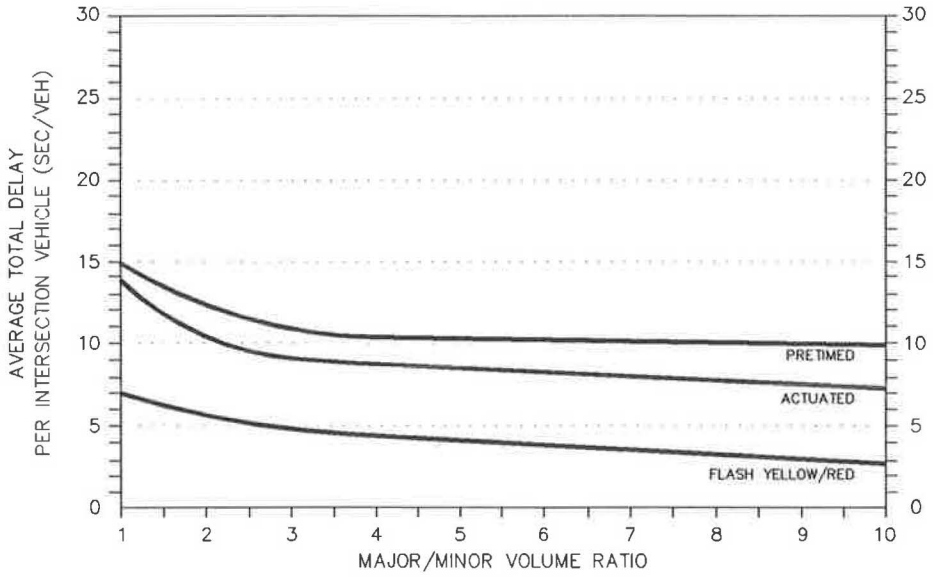


FIGURE 4 Signal system operation with the major-street volume less than 125 vph.

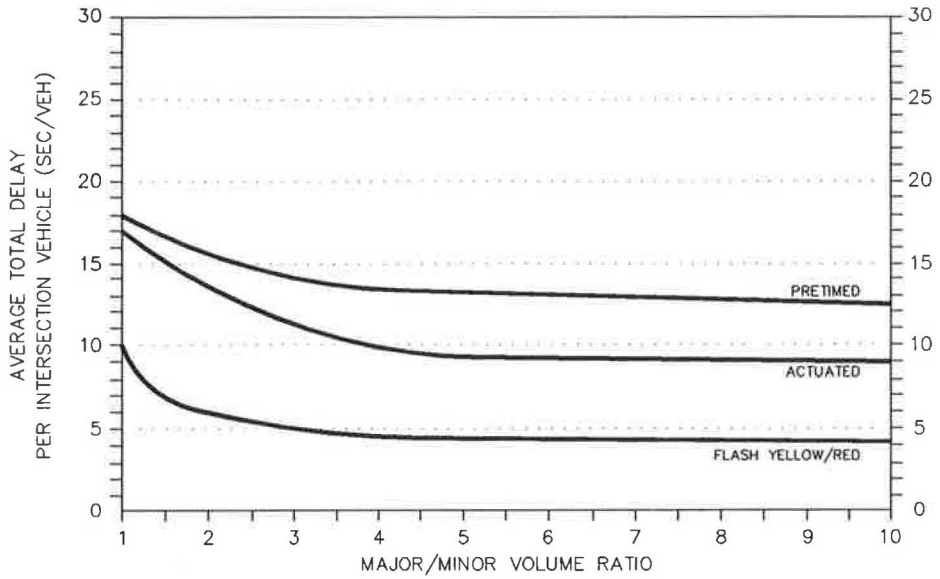


FIGURE 5 Signal system operation with the major-street volume between 250 and 500 vph.

yellow/red flashing (Figures 4–6). Less than 15 sec of total intersection delay was produced by pretimed operation with major arterial street volume less than 125 vph. As the volume ratio approached 10, total delay was reduced to 10 sec. For this same traffic volume (125 vph and less), actuated operation produced approximately 3 sec less delay than pretimed operation. As the major-street volume increased, actuated operation produced approximately two-thirds the total delay found with pretimed operation.

Summary of Signal System Analysis Results The following conclusions can be drawn from Figures 4 through 6:

- Yellow/red flashing operation produces the least amount of total intersection delay, followed by actuated and pretimed signal operation.
- Total intersection delay can be reduced by as much as one-third if pretimed operation is changed to actuated (coordinated) operation and reduced by approximately two-thirds if changed to yellow/red flashing operation.

Analysis of Late-Night, Low-Volume Periods

Flashing signal operation is normally implemented during low-volume periods. During the day, there are periods that are commonly referred to as "off-peak" times. However, the traffic volumes during which flashing operation should be considered are substantially lower than the daytime off-peak volumes. For Figures 1 through 6 to be useful to the practicing traffic engineer, the nighttime hourly volumes must be known.

Late-night traffic volumes were obtained as part of this study. The objective was to identify relationships between

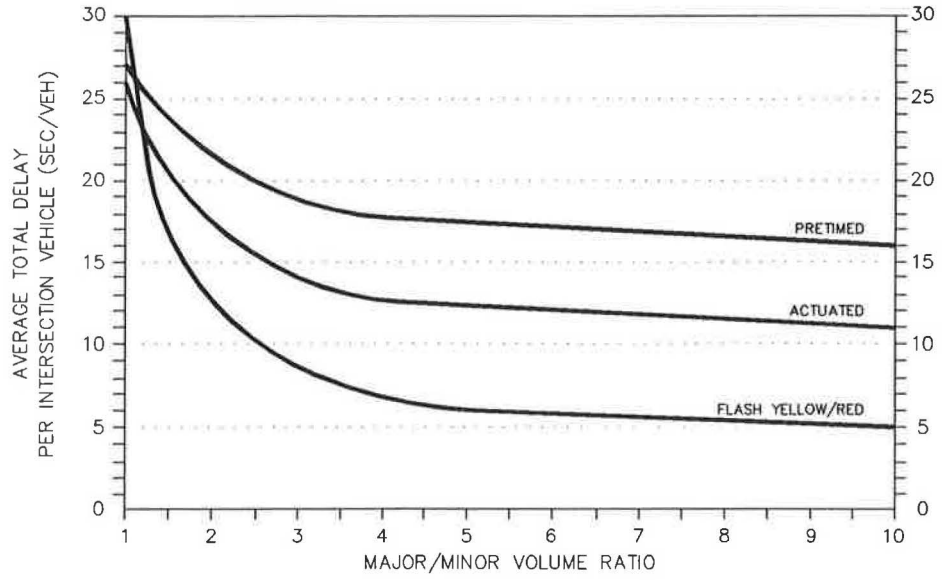


FIGURE 6 Signal system operation with the major-street volume between 750 and 900 vph.

these volumes and ADT. The study team analyzed 24-hr traffic volumes from 12 study sites in Austin, Bryan, and College Station, Texas. The ADTs for these sites varied from 6,000 to 30,000. An attempt to establish a relationship by ADT was made using the following six classes: less than 10,000; 10,001 to 15,000; 15,001 to 20,000; 20,001 to 25,000; 25,001 to 30,000; and more than 30,000.

The results identified the relationship between total ADT and late-night traffic volumes. It was found that a typical hour between midnight and 6:00 a.m. averaged 0.4 percent of the ADT, with a range of 0.1 to 1.7 percent. Total average values for the period between midnight and 6:00 a.m. ranged between 1.6 and 4.4 percent of the ADT. The overall average of traffic volumes was 2.6 percent of the ADT for the total volume between midnight and 6:00 a.m. Table 3 shows the average volume for each of the hours from midnight to 6:00 a.m. No relationship was found among the different ADT classifications.

CONCLUSIONS

The research described in this paper led to a number of conclusions related to flashing signal operation and how flashing operation compares with normal signal operation. These findings should provide some assistance to traffic engineers faced with deciding whether to implement or remove flashing signal operation.

The literature review found that many guidelines have been suggested for implementation, removal, or both of flashing operation of traffic signals. The majority of these guidelines are based on traffic volumes and accidents. The traffic volume guidelines are typically based on (*a*) the traffic entering an intersection, (*b*) the ratio of major-street to minor-street volume, and/or (*c*) a percentage of the existing traffic signal warrant volumes. Accident guidelines are usually stated as a condition to remove flashing signals. In some cases, the guidelines from different studies disagreed or conflicted with one another. For example, the Portland study (*5*) recommended placing signals into flashing operation if the volume ratio was equal to or less than 2, whereas the Michigan study (*6*) recommended removal of flashing operation when the volume ratio is equal to or less than 4.

The traffic engineer survey indicated that there is widespread use of flashing operation. The majority of traffic engineers make decisions on flashing operation without guidelines, depending on field experience and engineering judgment. It is recognized that these two qualities are valuable, but comprehensive guidelines based on relevant research should be available to assist the engineer in making the decision. The survey found that many traffic engineers do not feel comfortable with selecting flashing signal operation because of concerns about motorist behavior at such signals.

TABLE 3 Typical Relationship Between 24-Hr Volume and Late-Night Traffic Volume

Hour of the Day	Percent of 24-Hour Traffic Volume
12:00 to 1:00 a.m.	0.8 %
1:00 to 2:00 a.m.	0.5 %
2:00 to 3:00 a.m.	0.4 %
3:00 to 4:00 a.m.	0.2 %
4:00 to 5:00 a.m.	0.2 %
5:00 to 6:00 a.m.	0.5 %
Total 12:00 to 6:00 a.m.	2.6 %

An evaluation of various types of traffic signal operation was made using the TEXAS and NETSIM models for isolated intersections and for signal systems. The following conclusions can be drawn:

- For isolated intersections, the volume ratio must less than 3 before the minor street affects total intersection delay. For signal systems, the volume ratio must be less than 4 before the minor street affects total intersection delay.
- For major-street approach volume less than 750 vph, yellow/red flashing produces approximately one-third the delay produced by pretimed operation for volume ratios greater than 3. For volume ratios less than 3, yellow/red flashing produces approximately one-half the delay produced by pretimed operation.
- For major-street volume over 750 vph, yellow/red flashing produces approximately one-third the delay produced by pretimed operation for volume ratios greater than 4. For volume ratios less than 4, yellow/red flashing approaches the same delay as pretimed operation.
- Red/red flashing operation produces a relatively constant amount of delay when the major-street volume is less than 750 vph, and becomes constant after reaching a volume ratio of 4 for major-street volumes greater than 750 vph.

The conclusions reached from the operational analysis considered only the traffic volume relationship of flashing operation and not its safety impacts. Consideration should be given to the potential increase in accidents that the literature review found may accompany flashing operation. In addition, the operational results reflect the conditions and assumptions in Table 2 (i.e., 5 × 4 intersection geometrics and three-phase leading-left phasing arrangement).

An analysis of late-night, low-volume periods indicated that, as a general rule of thumb, the total volume between midnight and 6:00 a.m. is 2.6 percent of the ADT. For a typical hour during this time period, the volume is 0.4 percent of the ADT. Traffic volumes drop off after 2:00 a.m. Roadways with higher ADTs do not experience higher late-night traffic as compared with roadways with lower ADTs.

On the basis of the operational analysis of flashing signal operation only, the following observations are presented. These findings accounted for the number of lanes used in the geometric analysis, and therefore delay per vehicle is expressed in per-lane terms.

- Yellow/red flashing operation can significantly reduce overall delay at intersections with pretimed or actuated signal controllers when all of the following conditions are met: volume ratio greater than 3, major-street approach volumes less than 250 vph per lane, and higher approach minor-street traffic volume less than 85 vph per lane.

- Red/red flashing operation produces less delay at intersections with pretimed signal controllers when the following conditions are met: volume ratio is less than 3 and major-street volume is less than 250 vph per lane.

ACKNOWLEDGMENT

This paper is based on research sponsored by the Texas Department of Transportation in cooperation with the Federal Highway Administration.

REFERENCES

1. *Manual on Uniform Traffic Control Devices,* FHWA, U.S. Department of Transportation, Washington, D.C., 1988.
2. *Manual on Uniform Traffic Control Devices for Streets and Highways*. American Association of State Highway Officials, National Conference on Street and Highway Safety, Washington, D.C., Nov. 1935, reprinted Sept. 1937.
3. *Texas Manual on Uniform Traffic Control Devices*. Texas Department of Transportation, Austin, 1980, revised 1988.
4. Benioff, B., C. Carson, and F. C. Dock. *A Study of Clearance Intervals, Flashing Operation, and Left-Turn Phasing at Traffic Signals*. Report FHWA-RD-78-48. FHWA, U.S. Department of Transportation, May 1980.
5. Feroz Akbar, M., and R. D. Layton. Accident Experience of Flashing Traffic Signal Operation in Portland, Oregon. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 24–29.
6. Barbaresso, J. C. Flashing Signal Accident Evaluation. In *Transportation Research Record 956*, TRB, National Research Council, Washington, D.C., 1984, pp. 25–29.
7. Barbaresso, J. C. Relative Accident Impacts of Traffic Control Strategies During Low-Volume Nighttime Periods. *ITE Journal,* Vol. 57, No. 8, Aug. 1987, pp. 41–46.
8. Lee, C. E., R. B. Machemaehl, T. W. Rioux, and R. F. Inman. *TEXAS Model Version 3.0 Documentation with Supplement for Version 3.10 and Version 3.11*. Center for Transportation Research, University of Texas, Austin, March 1992.
9. Radwan, A. E., F. Naguib, and J. Upchurch. Assessment of the Traffic Experimental and Analysis Simulation Computer Model Using Field Data. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 216–226.
10. Kell, J. H., and I. J. Fullerton. *Manual of Traffic Signal Design*. Prentice Hall, Englewood Cliffs, N.J., 1991.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Microscopic Simulation Modeling of Minimum Thresholds Warranting Intersection Signalization

ANTHONY A. SAKA

The subject of this paper is use of a microscopic simulation model to estimate the minimum thresholds that require the installation of traffic signals at intersections. A simulation modeling approach was used to evaluate the reasonableness of signal installation Warrants 1 and 2 documented in the *Manual on Uniform Traffic Control Devices* (MUTCD). The results obtained from the simulation experiment indicate that the MUTCD warrants are conservative for some situations and hence if rigidly applied can result in premature installation of traffic signals. It was deduced from the simulation experiment that the minimum thresholds that require the installation of traffic signals depend on the geometric configuration of the intersection, that is, four-leg versus T-intersections. For example, according to the results obtained from the simulation experiment, for the same traffic conditions, four-leg intersections will require lower thresholds than T-intersections. A reasonable minimum threshold was estimated for T-intersections. This threshold can be used to supplement the MUTCD warrants.

The *Manual on Uniform Traffic Control Devices* (MUTCD) is considered the authoritative reference manual for implementing a number of traffic control measures, including the warrants for traffic signal installation at intersections. Usually a traffic condition satisfying one or more of the warrants documented in the MUTCD is considered necessary for traffic signal installation. These warrants include minimum traffic volumes on the major and the minor streets.

Typical thresholds used range from 500 to 900 vehicles per hour (vph) for the major streets and a one-way volume of 75 to 200 vph for the minor streets (*1*). In addition, some states and localities have additional sets of guidelines to supplement those of the MUTCD.

Traffic signals have proven to be very effective in improving safety and traffic flow at intersections. However, experience has shown that traffic signals can be a nuisance if not properly timed or if used when not "warranted," or both. For example, poorly timed signals impede the flow of traffic by giving green time to approaches that do not have adequate demand. This type of problem is more prevalent at isolated intersections with pretimed traffic signals.

In order to minimize the likelihood of premature installation of traffic signals at intersections, it is very important to evaluate and validate commonly used guidelines (i.e., guidelines documented in the MUTCD).

Center for Transportation Studies, Morgan State University, Baltimore, Md. 21239.

For this paper, a microscopic simulation modeling approach is used to evaluate the minimum thresholds (traffic volumes) documented in the MUTCD that warrant traffic signal installation. A set of supplementary guidelines that could be used in conjunction with those of the MUTCD is suggested.

BACKGROUND OF THE PROBLEM

Historically, minimum vehicular volumes warranting traffic signal installation are generally applied to all intersections with little consideration to the geometric configuration of the intersection (*1*).

As documented in the 1985 *Highway Capacity Manual* (HCM) (*2*), traffic for the different lane-groups accepts different safe gaps to undertake turning maneuvers. For example, minor-street left-turning traffic accepts significantly larger gaps than minor-street right-turning traffic, and so on. Consequently, two intersections with the same total major- and minor-street traffic volumes may not operate at the same level of service (LOS). The intersection with the most vehicular movement conflicts is expected to operate at a lower LOS than the intersection with the least vehicular movement conflicts. Therefore, the geometric configuration of the intersection is a very important factor to consider in estimating the thresholds warranting signal installation. For example, T-intersections have fewer vehicular movement conflicts than do four-leg intersections. Consequently, it will be unreasonable to apply the same minimum thresholds to these two categories of intersections.

OBJECTIVES

The objectives of this paper are to evaluate, via simulation, the minimum thresholds warranting traffic signal installation as documented in the MUTCD and to provide supplemental guidelines.

ASSUMPTIONS

The building of the simulation model involved several assumptions, most of which are based on standard traffic engineering practices. The assumptions made are discussed under five major categories: (*a*) geometric configuration of the intersection used, (*b*) service prioritization of the different lane-

group traffic movements at the intersection, (*c*) traffic flow parameters used in the simulation experiment, (*d*) configuration of the simulation model, and (*e*) decision rules for the simulation experiment.

Geometric Configuration of the Intersection

The minimum thresholds warranting traffic signal installation depend on the configuration of the intersection. One of the primary purposes of traffic signals is to minimize vehicular conflicts at intersections. The severity of the conflicts depends on the number of different types of turning maneuvers at the intersection. For example, intersections of two-way streets have more vehicular conflicts than those of one-way streets, and four-leg intersections have more vehicular conflicts than T-intersections. Clearly, each of the above two cases of intersections requires a different set of minimum thresholds warranting traffic signal installation.

In the simulation experiment, two sets of intersections were considered. These were a four-leg intersection with two-way major- and minor-street traffic and a T-intersection with two-way major- and minor-street traffic.

Prioritization of Traffic Movement

Using the 1985 HCM (*2*) guidelines for "unsignalized intersection analysis," turning movement priorities were established (see Figure 1). Because both the major-street through movements and right-turn movements are usually unimpeded, these movements were designated Priority 1, the major-street left-turn movements Priority 2, the minor-street right-turn movements Priority 3, the minor-street through movements Priority 4, and the minor-street left-turn movements Priority 5. Priority 1 has the highest preference and Priority 5 has the lowest preference with regard to service.

Traffic Flow Parameters

As mentioned earlier, major-street through and right-turn traffic at unsignalized intersections is usually unimpeded. Lower-priority traffic movements will have to queue whenever there is major-street through or right-turn traffic, or both, at the intersection.

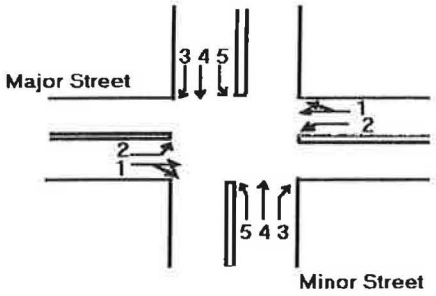


FIGURE 1 Movement categorization.

The service time—that is, the time taken by a vehicle to traverse the intersection—depends on two major variables: the width of the intersection and the average travel speed of the vehicle. Therefore, in the simulation experiment, exponentially distributed service times were assumed for all lane-groups (i.e., Lane-Groups 1, 2, 3, 4, and 5). Using the critical gap acceptance data documented in the "unsignalized intersection analysis" section of the 1985 HCM, average service times (Table 1) were assumed of 2, 5, 5.5, 6.5, and 7.0 sec/vehicle for Lane-Groups (movement priorities) 1, 2, 3, 4, and 5, respectively (*2*).

In the simulation experiment, vehicles were served one at a time on the basis of the aforementioned service priorities. However, in the real world, more than one vehicle can be served at a time. This is particularly true for traffic belonging to the Priority 1 lane-group category. For example, although arrival and departure at unsignalized intersections are generally considered random, vehicles belonging to the major-street through and right-turn lane-group category can arrive and be served from both directions at the same time. Therefore, the minimum threshold estimated from the experiment could be even less than the "actual" threshold.

A Poisson traffic arrival pattern, that is, exponentially distributed interarrival times, was used. The vehicular arrival rate, A_t, at the intersection was estimated as

$$A_t = (\Sigma v_i)/3{,}600 \tag{1}$$

where A_t is the average traffic arrival rate (in vehicles/second) at the intersection and v_i is the traffic volume at the intersection for lane-group i (in vehicles/hour).

In the simulation experiment, the expected traffic arrival rate (i.e., demand for a given lane-group) was estimated as

$$A_i = A_t(P_i) \tag{2}$$

where A_i is the expected traffic arrival rate for Lane-Group i and P_i is the probability that a given arrival at the intersection belongs to Lane-Group i; that is, $P_i = v_i/\Sigma(v_i)$.

As mentioned earlier, the interarrival times at the intersection are assumed to be exponentially distributed. Thus, the parameter of the exponential distribution is $1/A_t$.

Decision Rules

As mentioned earlier, one of the primary purposes of traffic signals is to reduce vehicular conflicts and hence to increase traffic flow at intersections. Without traffic signals, arrivals other than those belonging to the Priority 1 lane-group category would have to remain in the queue until a safe gap was available to undertake turning maneuvers. As the traffic on the major street or on the minor street, or both, increases, the likelihood of finding safe gaps for turning maneuvers reduces. Eventually, the expected available number of safe gaps will become less than the expected number of vehicles to be served.

This study assumes the minimum demand thresholds that warrant the installation of traffic signals at an intersection to

TABLE 1 Lane-Group Service Times

Critical Lane-Group	Average Service Time per Vehicle
1. Major Street Through & Right-turn	2.00 seconds
2. Major Street Left-turn	5.00 seconds
3. Minor Street Right-turn	5.50 seconds
4. Minor Street Through	6.50 seconds
5. Minor Street Left-turn	7.00 seconds

be the thresholds beyond which queues on one or more lane-groups at the intersection will no longer attain statistical equilibrium. In other words, the queue or queues will continue to increase with time as the expected demand exceeds the expected available safe gaps required to undertake turning maneuvers. Figure 2 shows when traffic signal installation is necessary. It can be seen that the queues for movements (MVTs) 1 and 2 are statistically stable, whereas the queue for MVT 5 increases with time. In other words, the queue for MVT 5 cannot attain statistical equilibrium, a condition warranting traffic signal installation.

CONFIGURATION OF THE SIMULATION MODEL

The SIMAN simulation package was used in building the model. Figure 3 shows the architecture of the simulation model, which contains four main blocks: vehicular arrival block, lane-group categorization block, service block, and queue inventory block.

Vehicular Arrival Block

The purpose of the vehicular arrival block is to create the traffic arriving at the intersection. Arrivals created in this block are considered generic. They do not have any lane-group identifications.

Lane-Group Categorization Block

The arrivals created are brought into the lane-group categorization block, where they are categorized and assigned to the appropriate lane-groups. Assignments are made on the basis of the aforementioned probability, P_i, assigned to the individual lane-groups at the intersection.

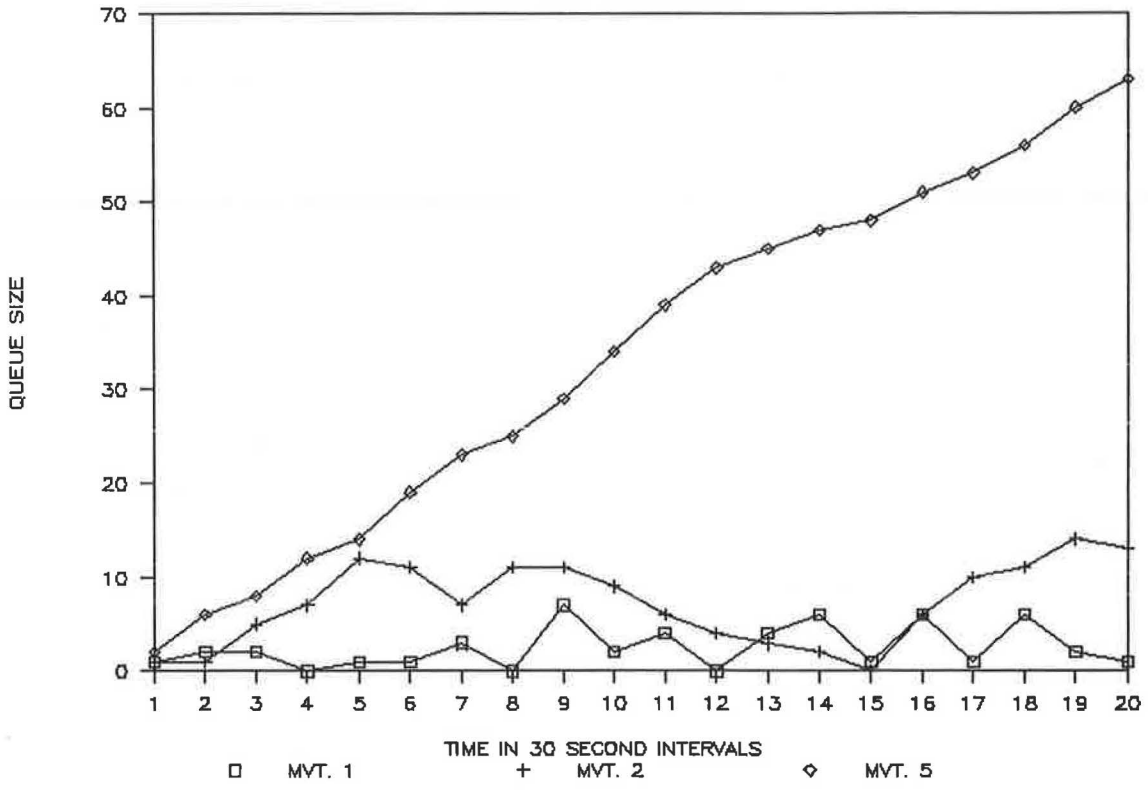


FIGURE 2 Minimum threshold requirements.

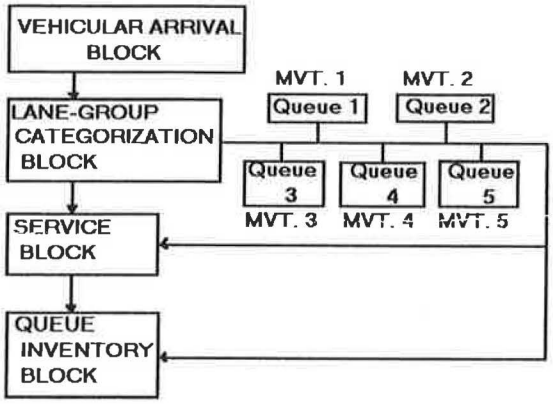


FIGURE 3 Configuration of simulation model.

Vehicular Service Block

Vehicles assigned to a given lane-group are either served or sent to a queue block. Services are rendered in accordance with the aforementioned lane-group service prioritization.

Queue Inventory Block

Unserved vehicles for a given lane-group are stored in the queue block reserved for that lane-group. The size of the queue for the individual lane-groups is continually updated in the queue inventory block. In addition, the queue size is reported at 30-sec intervals for a minimum observation period of 20 min or 600 sec. The status of the queue for the individual lane-groups during the period of observation was evaluated from output graphs. In other words, the graphs were used to easily identify lane-groups where queues do not attain statistical equilibrium. As mentioned earlier, the stability of the queues at the intersection is used as a guide to determine whether traffic signal installation is necessary.

DESCRIPTION OF THE EXPERIMENT

As stated earlier, the primary objective of this study is to evaluate the suitability of the MUTCD "minimum volumes warrants for traffic signal installation" for both four-leg and T-intersections. The study considered two cases, Scenarios 1 and 2. Scenario 1 involves the evaluation of the performance of the aforementioned MUTCD warrants for four-leg intersections and Scenario 2 those for T-intersections. The MUTCD warrants evaluated in the experiment are the minimum volumes for intersections with two or more traffic lanes for both the major and minor streets.

The experiment began by inputting into the simulation model the equivalent traffic flow parameters of the MUTCD "minimum thresholds." The resulting queue and delay data were evaluated for statistical equilibrium for Scenarios 1 and 2. Using an increment of 50 vehicles at a time, a sensitivity analysis was undertaken to determine whether thresholds higher than those of the MUTCD would result in stable traffic conditions at the intersection. The term "stable" refers to the status of the queues at the intersections. As mentioned earlier, the decision rule of the experiment is that traffic signal installation is necessary when queues in one or more lane-groups cannot attain statistical equilibrium (see Figure 2).

SUMMARY OF RESULTS

Table 2 shows the minimum thresholds obtained from the simulation experiment for the aforementioned scenarios (see also Figures 4–6).

Figure 4 shows the behavior of the queues at a four-leg intersection when the aforementioned MUTCD warrants were applied. It can be deduced from Figure 4 that queues for MVT 2 attained statistical equilibrium, whereas those for MVTs 3 and 5 did not attain statistical equilibrium. Therefore, according to the aforementioned decision rule, traffic signal installation is necessary. Using threshold values slightly lower than those of the MUTCD resulted in satisfactory traffic conditions for MVTs 3 and 5. Therefore, it was concluded that the MUTCD Warrants 1 and 2 traffic signal thresholds are optimum for four-leg intersections.

Figure 5 graphs the behavior of the queues at a T-intersection when the aforementioned MUTCD Warrant 1 threshold was applied. It can be deduced from Figure 3 that even the queues for the most "critical" lane-group, MVT 5 (minor-street left-turn movement), were determined to be statistically stable when MUTCD Warrant 1 was applied to a T-intersection. Similar results were obtained for MUTCD Warrant 2 and slightly higher thresholds. This implies that the current MUTCD signal Warrants 1 and 2 are not "optimum" for T-intersections.

A sensitivity analysis was undertaken to estimate a more reasonable minimum threshold for T-intersections. Figure 6 shows the queue behavior for MVT 5 based on the estimated minimum threshold. It can be deduced from Figure 6 that the

TABLE 2 Estimated Minimum Thresholds

Intersection Geometric Configuration	Major Street two-way hourly volumes	Minor Street one-way hourly volumes
Four-Leg Intersection	600 vph --- 900 vph	200 vph --- 100 vph
T-Intersection	1,000 vph	200 vph

MUTCD Warrants 1 & 2 minimum thresholds range from 500 to 600 vph and 750 to 900 vph (both directions) for the major street, and 150 to 200 vph and 75 to 100 vph (one direction) for the minor street, respectively. There are no separate guidelines for Four-leg intersections and T-intersections.

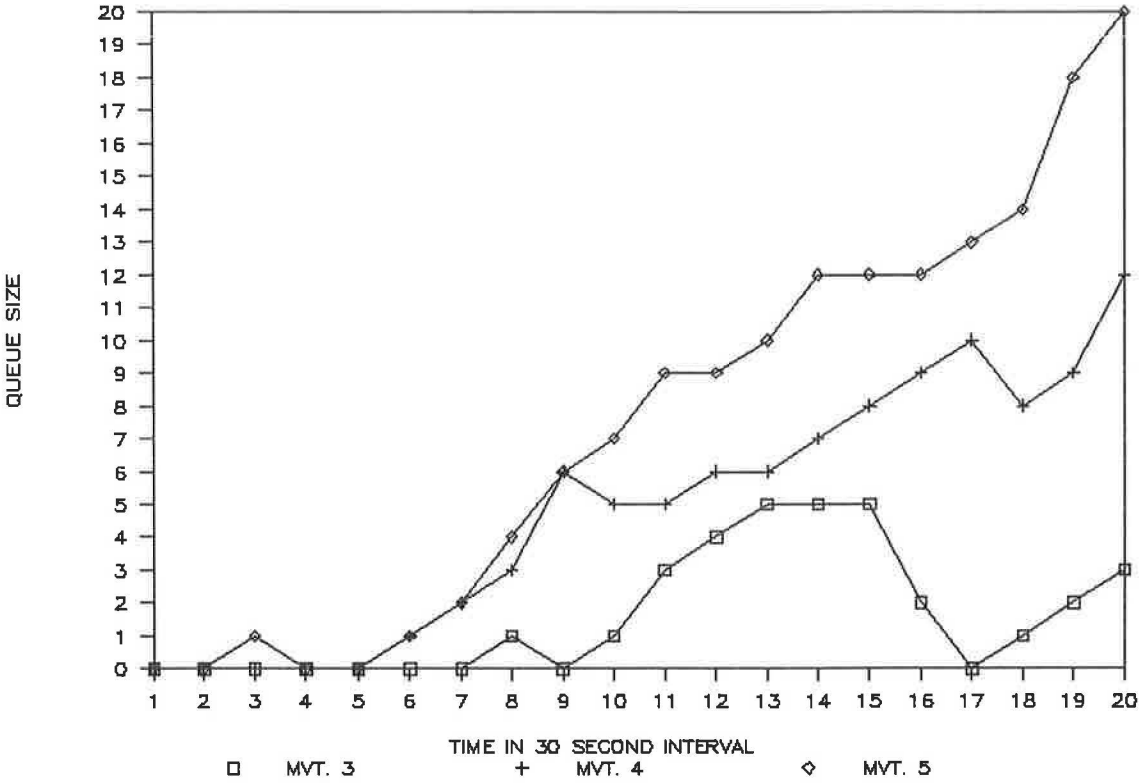


FIGURE 4 Queues for Scenario 1.

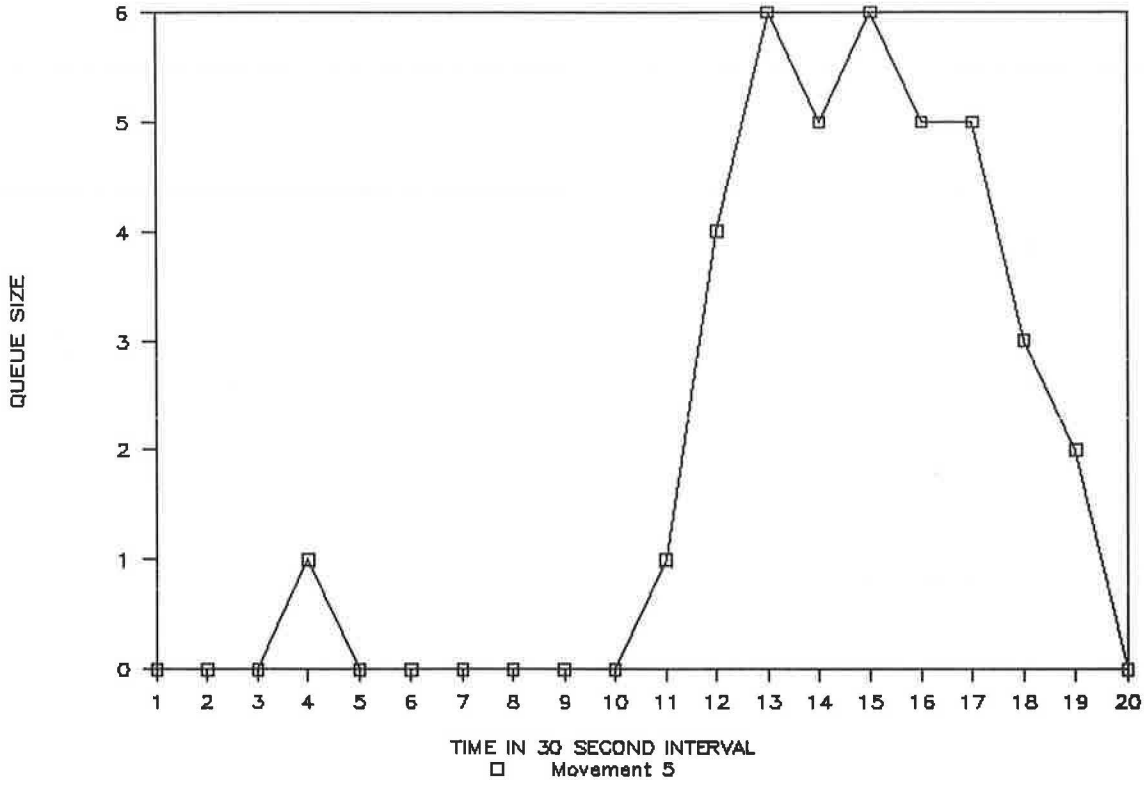


FIGURE 5 Queues for Scenario 2.

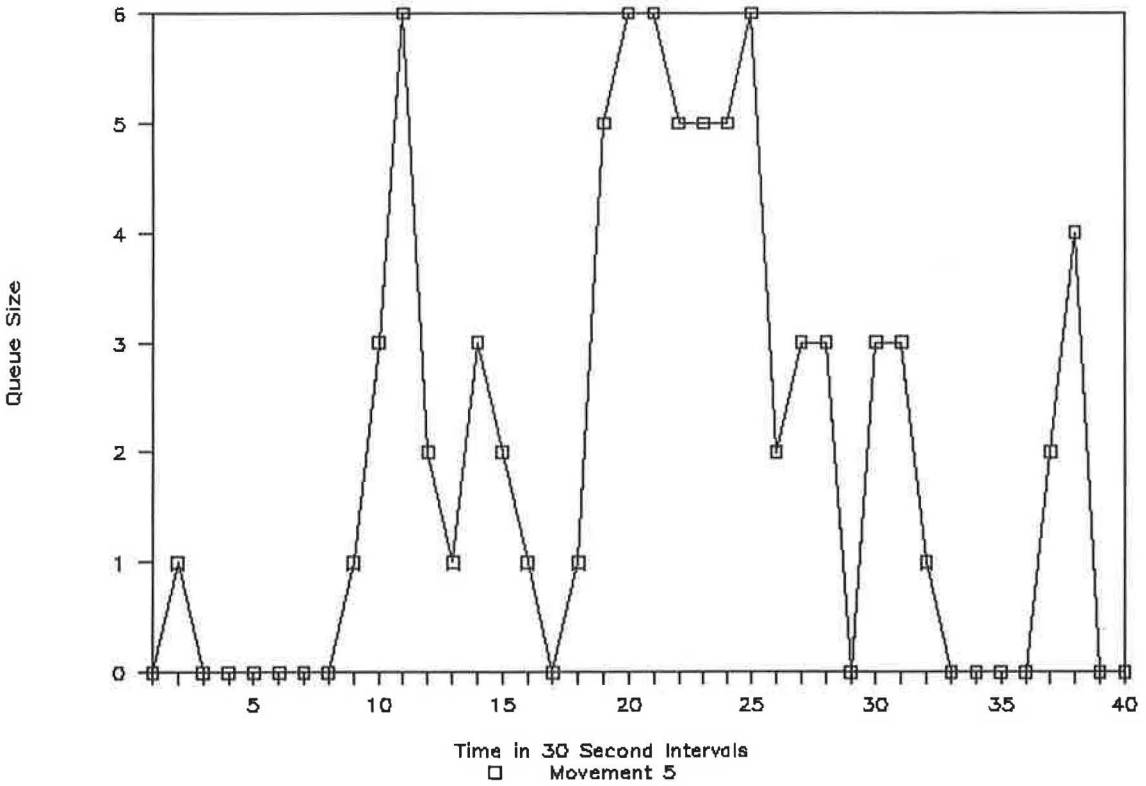


FIGURE 6 Minimum threshold queues for T-intersection.

queues for MVT 5 do attain statistical equilibrium. However, using thresholds higher than the estimated minimum threshold resulted in unstable queues at the intersection. Therefore, it was concluded from the experiment that the estimated minimum threshold was "optimum" for T-intersections.

CONCLUSIONS

Traffic signals have proven very effective in minimizing traffic conflicts at intersections. However, installation of traffic signals at intersections should be well timed to avoid unnecessary traffic delay to the major-street traffic. The MUTCD provides guidelines for determining when traffic signals are required at intersections. Traffic signal installation warrants of the MUTCD involving vehicular volumes are too generalized. These warrants, the aggregated traffic volumes for the major street and the minor street, do not give adequate consideration to the effects of the intersection's geometric configuration.

It has been shown in this paper that the minimum thresholds (volumes) warranting traffic signal installation vary for different intersection configurations. A minimum threshold was estimated for T-intersections. It is expected that this threshold would be useful as a supplement to the existing MUTCD warrants. However, this author strongly recommends that further studies be undertaken to validate this threshold as well as those MUTCD warrants not considered in this study.

REFERENCES

1. *Manual on Uniform Traffic Control Devices for Streets and Highways.* FHWA, U.S. Department of Transportation, 1978.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Development of an Emergency Zone Sign

MARTIN T. PIETRUCHA

The concept of having a separate category of traffic sign to control traffic in emergency situations is advanced. This category is referred to as an emergency zone sign (EZS). A rationale is provided for having this category of sign and for developing a family of emergency zone signs. Detail is provided on the sign messages, shapes, and colors.

One of the functions of a traffic control device (TCD) is to warn drivers of unexpected hazards in the roadway. Usually these hazards are permanent features of the roadway or environment, but often hazards are temporary, as in the case of construction and maintenance. The construction and maintenance function has become so pervasive and is viewed as so different a hazard that a special class of TCD was developed for use in work zones. Attention should now be focused on another on-street operation area that poses a hazard to the motoring public and the parties involved—the emergency zone (EZ).

The emergency zone can be defined as an area in which, because of some incident, a special hazard exists that necessitates emergency services such as those provided by police, fire, and emergency medical professionals. Traffic accidents, downed power lines, and building and automobile fires are a few examples of incidents in an emergency zone. These incidents can cause serious safety problems for those attending to the emergency situation and for motorists attempting to pass by or through the EZ. Although most emergency vehicles are equipped with some type of special lighting, these lighting devices alone do not give the motorist enough information to pass the EZ without causing additional problems. Therefore, it is proposed that an emergency zone sign (EZS) or family of signs, similar to the work zone signs, be developed for use by emergency personnel to control traffic in and around the EZ. The objective of this paper is to report on the development of such a sign.

LITERATURE REVIEW

The literature contains many reports that deal with traffic control during emergencies. The Maryland Police Training Commission (*1*) has produced a nine-part instructional series on collision management procedures for police trainees. One part deals exclusively with controlling the accident scene. Flares, cones, and emergency vehicle lighting are all recommended as advance warning devices, but the use of signing of any type is not suggested. In a report prepared by Wilbur Smith and Associates (*2*) for the Highway Safety Division of Virginia, flares, cones, lighting, and signing are advocated as aids to secure a traffic accident scene. Guidelines for placement of these devices are given, but there is no mention of what specific signing is to be used.

Although signing is usually mentioned as a traffic control alternative, the use of vehicle lighting in emergency situations is cited most frequently. According to some state vehicle codes, the use of particular colors on certain types of vehicles at specific times constitutes a specific type of warning, but there appears to be no uniformity among these conventions from state to state (*3*). Another problem cited is the often extreme difficulty for a motorist of determining whether an emergency vehicle is moving or stationary when it is using lights or light bars. One study investigated the possibility of removing the roof-mounted lighting devices from police vehicles as a means of saving energy and to improve surveillance capabilities (*4*).

Changeable or variable message signs have long been recognized as an effective part of a freeway incident management system (*5–7*). Often the effectiveness of these systems is compromised by information that does not reflect actual roadway conditions because of the time lag between a change in the status of the incident and a change in the message to the driver.

This problem of time lags in the reporting system was addressed in a project by 3M Company and the Minnesota Highway Patrol (*8*). A vehicle-mounted changeable message sign was developed by 3M Company and field-tested by the Highway Patrol. The sign used a continuous scroll of eight different messages to warn motorists of various hazards. It was mounted flat on the roof of the vehicle and could be raised while the vehicle was still in motion. Use of this device reduced the time needed to attend to an incident and reduced the number of secondary collisions as well.

Since TCDs for emergency zones do not exist, there is no discussion of the placement for such a device. Placement of many standard TCDs is based on prevailing speed and conditions as well as the time necessary for drivers to comprehend and react to the TCD and alter their driving accordingly (*9*). Methods for determining stopping sight distance and decision sight distance take these factors into consideration (*10*), whereas for placement of flares or other warning devices currently used by police or other emergency personnel, distances are based on vehicle braking distances only (*1*).

Since the EZS is a new type of device, not only can the message be novel, but also colors and shapes can be used that are not bound to currently used forms. In the *Manual on Uniform Traffic Control Devices* (MUTCD), several colors have been reserved for future use in addition to the standard colors already in use (*9*). One of the few studies of motorists' understanding of traffic signing shape and color coding was done by the Virginia Highway Research Council (*11*). The

Pennsylvania Transportation Institute, The Pennsylvania State University, Research Office Building, University Park, Pennsylvania 16802.

study showed that singular and combined uses of color and shape did not effectively communicate to drivers what type of message they were to receive from a sign. Although much has been done to study the recognizability and legibility of various sign shapes and colors (*12*), little has been done to study driver knowledge of the MUTCD color and shape coding conventions.

EMERGENCY ZONE SIGN DEVELOPMENT

The first step in developing the EZS was to determine the needs of the groups who would be using the device and the information requirements of motorists. To accomplish this, the aid of several public agencies was sought to provide information about "on the street" conditions. The author rode with county police traffic units (Montgomery County, Maryland), state trooper units (Maryland), and large urban area fire and rescue crews (District of Columbia Fire Department). These experiences provided insight into the potential uses for an EZS, possible means of deployment, and the nature of the traffic such a device would have to control.

This variety of emergency service agencies provided opportunities to observe a wide range of activities and incidents in EZs. Riding with the county police traffic units provided occasions to observe accidents and stopped-vehicle situations in low- to medium-speed conditions on arterial streets, collectors, and rural country roads. While traveling with the state troopers, the researchers had many opportunities to observe emergency situations on higher-speed limited-access facilities. Riding with fire and rescue squads in the District of Columbia, the author experienced many different emergency situations in an urban setting.

The major advantage of riding in the police and fire vehicles was the speed with which the vehicles arrived on the scene. This allowed observations to be made for the full time period in which an EZS would be deployed, used, and picked up. To facilitate the analysis of each incident, a videotaped record of the emergency was made. The records were limited to views of the traffic approaching the emergency zone and verbal descriptions of the actual hazard.

In analyzing the videotapes, it became apparent that although the exact nature of every incident was different, there were several common elements. These elements were given generic names: recovery time, closure type, and control strategy.

Recovery time is the total time period from when the incident first occurs until roadway conditions return to normal. Recovery time has a great bearing on whether an EZS is to be used. If the time to deploy and take up the EZS is equal to or greater than the recovery time, it is impractical to use it. To assess the impact of the recovery time element, it is necessary to find the point at which the added risk of placing and retrieving the device is outweighed by the added protection afforded by the device. The question still to be answered is "At what point does the break between liability and benefit occur?" This subject was beyond the scope of this study.

Closure type is a description of what part of the roadway is no longer available to the motorist because of the incident. On the basis of the field observations, there were six self-descriptive types of closure: shoulder, single-lane undivided roadway, multiple-lane undivided roadway, single-lane divided roadway, multiple-lane divided roadway, and full roadway. Each of these closure types can be treated with specific control strategies.

The control strategy is the means by which the traffic is redirected past the specific closure type. There are three basic control strategies. The first is to direct the traffic around a hazard utilizing the same side of the road as the affected motorists' direction of travel. The second is to direct the traffic around a hazard utilizing the side of the road opposite the affected motorists' direction of travel. The third is to completely close off the area to traffic at the nearest junction and reroute the traffic. These control strategies were the basis for the design of the message on the EZS.

Message Content

When the actual sign messages were developed, several things were kept in mind. One was that symbolic messages appear to offer several advantages over word messages, and the current preference by the National Committee on Uniform Traffic Control Devices and FHWA is symbolic signing. Another is that emergency personnel cannot keep an entire sign shop in the trunks or equipment bays of their vehicles; therefore, a limited number of designs with a wide variety of uses would be desirable. Last, it would be advantageous to use conceptual elements already in use on other TCDs in order to facilitate comprehension and learning of the new signs.

The candidate signs were designed by a team of traffic engineers, human factors specialists, and graphic artists.

The first series of signs, designed to execute the first control strategy, moving the traffic around a hazard using the same side of the road, was designated the E1 series (Figure 1). The

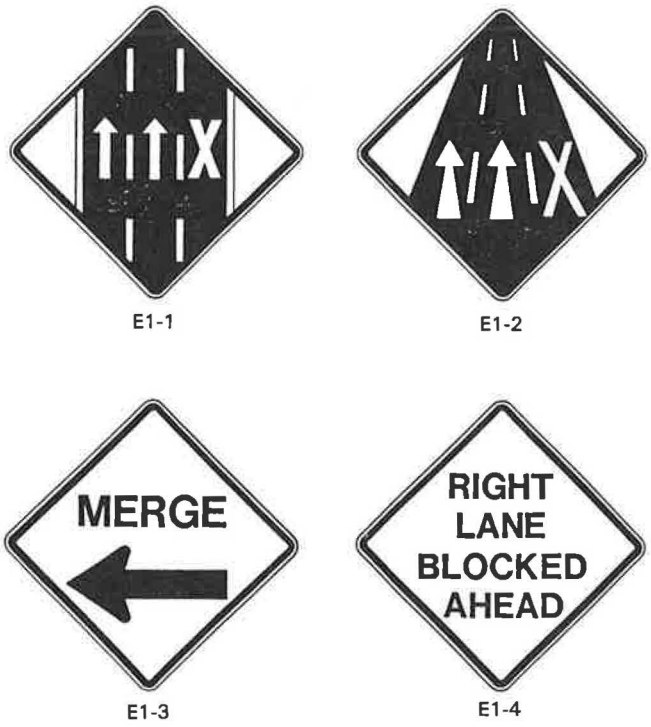


FIGURE 1 E1 series signs.

design incorporates the use of arrows for allowed through movements and X's for closed lanes, following the conventions for many existing signs and lane-use control signals. The signs would be fabricated so the arrows and X's could be moved from lane to lane to provide applicability for all situations. The perspective view used for Sign E1-2 is a variation based on experimental issues raised by Pietrucha and Knoblauch (*13*) in their study of sign comprehension. Signs using only word messages were also tested. These signs were designed to allow the MERGE arrow to point right or left or to have Sign E1-4 read RIGHT/LEFT LANE BLOCKED AHEAD.

The second series of signs, designated the E2 series, would be used to direct moving traffic around a hazard using the opposing flow lanes, the second control strategy (Figure 2). The designs again use the familiar arrows, X's, and merging elements of other TCDs. The signs could be modified to depict any situation. Within this series there are two types of signs. One shows the road condition to a driver who is approaching the hazard and would have to cross over to a contraflow lane (E2-1, 3, 7, and 9). The other type shows the road condition and would restrict the driver approaching the contraflow situation to a certain lane or lanes (E2-2, 4, and 6).

The third series of signs, designated the E3 series, is to be used for roadway closures. The signs use a variety of symbols—some familiar, some new to communicate the meaning of "no entry" (Figure 3). These symbols may be supplemented by a word message as part of a hybrid word-symbol sign (E3-1a).

Laboratory Procedures

The EZS went through a two-phase laboratory test. The first phase of testing was a screening procedure to winnow down the large number of EZS candidates. The second phase was a device-selection procedure to designate the specific device messages.

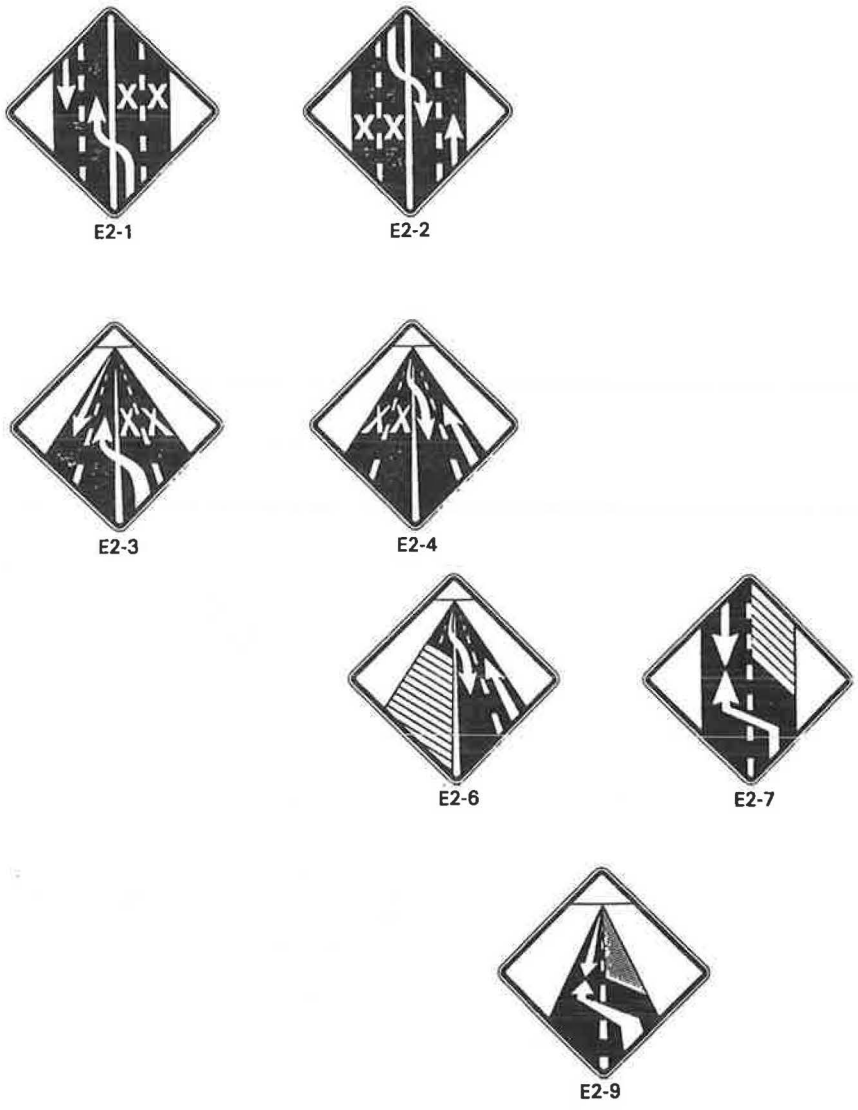


FIGURE 2 E2 series signs.

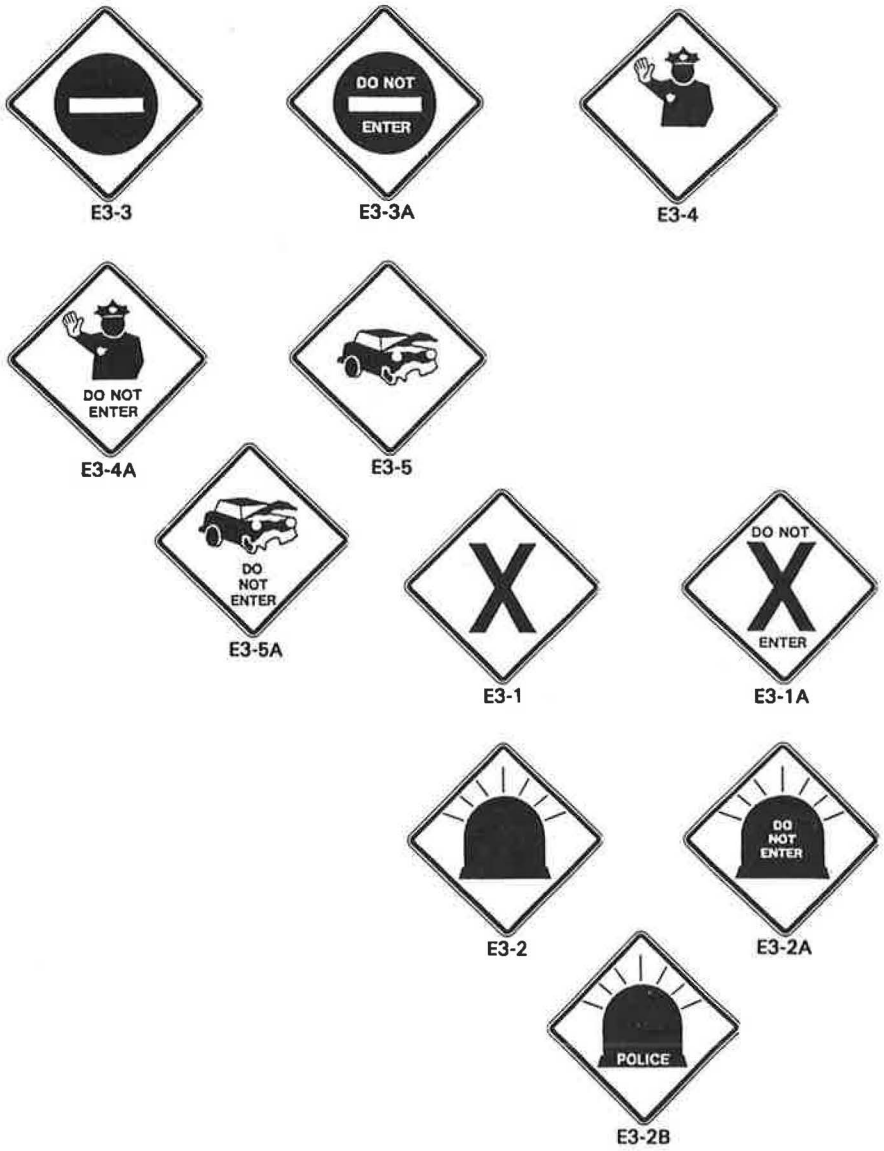


FIGURE 3 E3 series signs.

The primary measure of effectiveness (MOE) was made by administering paper-and-pencil tests to determine the accuracy of the subjects' interpretation of each design. This was done by presenting a stimulus (a picture of a traffic sign) and asking the simple open-ended question "What do you think this sign means?" (Figure 4).

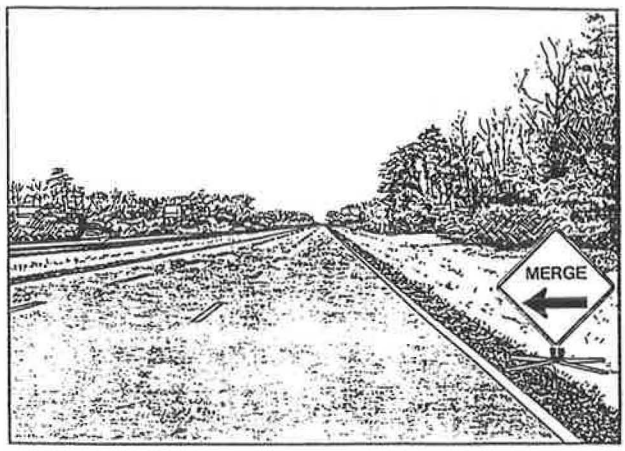
Test booklets containing the EZSs and other traffic signs were prepared. Each page included a picture of the sign and the question "What do you think this sign means?" The subjects were given as much time as they needed to complete the test booklet.

Screening Procedure

Test Subjects The subjects were selected from among individuals who were renewing their driver's license at a local office of the Department of Motor Vehicles (DMV). Test subjects were selected from an urban area (Baltimore, Maryland), a densely populated suburban area (Arlington, Virginia), a less densely populated suburban area (Fairfax, Virginia), and a rural area (Warrenton, Virginia). There were three age categories (<30, 30 to 50, and >50) for both sexes. A target cell total of 10 subjects was set. By testing 10 subjects in each age and sex category from each of the four geographic areas, a total of 240 subjects was tested. This guaranteed that each sign would be interpreted by at least 30 subjects.

Results A numerical coding scheme was created so that the subjects' answers could be tabulated and analyzed. The coding scheme attempted to preserve the essence of the original responses while giving the flexibility to cluster the data in several different categories and not lose the ability to expand and contract the data into new tabulations.

A two-part code was assigned to each response. The first part assigned the response to a general answer category. The second part identified individual responses within each cat-



What do you think this sign means ?

What do you think this sign means ?

FIGURE 4 Sign presentation in test booklet.

egory. Every distinct response was given its own code, and all similar replies were assigned the same code. The subject response code is as follows:

Code	*Category*
RIGHT	
0X	Correct
1X	Nearly correct
WRONG	
2X	Conceptually close
3X	Incorrect
4X	Bizarre
5X	Dangerously incorrect
6X	Confused with existing sign
7X	Overflow from other categories
8X	Overflow from other categories
9X	Unknown
00	No response/Don't know

In an attempt to facilitate decision making, a superhierarchy was established for the categories. Any answer considered correct or nearly correct was grouped into a "right" supercategory and all other responses (e.g., incorrect, bizarre, unknown) formed the "wrong" supercategory. Although the categories were useful for noting trends in responses and breaking ties among promising sign candidates, the decision to use a sign was based on how many people (expressed as a percentage) could give a functionally correct (right) interpretation of the sign. A chi-square test of independence was used to determine if there was a relationship between the sign candidates and the subject responses.

It was originally intended to use only one test procedure. When conducting the (screening) test and analyzing the results, researchers identified problems with the test method and analysis procedures. These identified problems were used to redesign the test and to modify the method of analysis.

The written responses from the screening procedure yielded answers that could have been interpreted in many ways. After the tests were completed, the subjects were no longer available to explain any ambiguous answers, so it was decided that the laboratory procedure would be repeated. In the new (selection) procedure, after the subjects filled out the test booklets, they were debriefed about their replies. Nondirective questions to clarify vague responses or to elicit additional information provided more information for analysis.

The screening procedure resulted in the elimination of several of the original EZS designs. Signs E1-1 and E1-2 were the only signs from the E1 series tested in the screening phase. It was thought that the word message signs (E1-3 and E1-4) would be fairly well understood, so they were defaulted to the selection procedure. A statistical analysis of the results showed the relationship between the signs and the subject responses to be significant at the 0.05 level. Although Sign E1-2 was interpreted correctly more often by the subjects (94 percent correct), it was decided to use Sign E1-1 (78 percent correct) for further testing because of the problems caused by perspective view signs in another sign comprehension study (*13*). In the previous study, there was no consistency in the performance of perspective view signs. For some types of sign, a perspective view version of the standard sign was very convincing; however, for other types perspective view versions performed poorly. Rather than introduce perspective view signing as part of a new sign category, it was decided to continue with standard plan view representations.

The large black area on Sign E1-1 also caused some concern about potential visibility problems, so a negative version of this sign was designed for subsequent testing along with Signs E1-3 and E1-4.

In the E2 series, the "crossover" signs, there were two subcategories, the four-lane crossover and the two-lane crossover. The two-lane crossover is a situation similar to that of a one-lane road. The pictographs for Signs E2-7 and E2-9 were tested as part of a set of One Lane Road Sign (W20-4) candidates in the previously referenced study by Pietrucha and Knoblauch (*13*). The results of the four-lane crossover signs, which were not statistically significant, were as follows:

Sign No.	*Percent Correct*
E2-1	71
E2-2	58
E2-3	74
E2-4	71
E2-6	69

Since perspective has been shown to cause cognitive problems, Signs E2-3, E2-4, and E2-6 were eliminated from further testing. The potential visibility problems caused by the large black areas on Signs E2-1 and E2-2 necessitated a change to a negative version for these signs.

The E3 series of signs consisted of symbols only and hybrid word-symbol signs. It was decided to test only the symbol signs, since it was believed that the hybrid signs would be more easily understood and the real interest was to see what responses the different symbols would elicit. The results for this group, which were statistically significant, were as follows:

Sign No.	*Percent Correct*
E3-1	14
E3-2	0
E3-3	22
E3-4	51
E3-5	42

Although the "wrecked car" (E3-5) was the second most often correctly identified sign, it was decided to eliminate it from further testing because the E3 series signs are envisioned as being used at all types of street closures (e.g., fires, crime scenes) rather than just for motor vehicle accidents. The police and fire dome light performed poorly and was eliminated from further testing. The remaining signs (E3-1, E3-3, and E3-4) were retained for testing in the next phase together with their hybrid counterparts (E3-1a, E3-3a, and E3-4a).

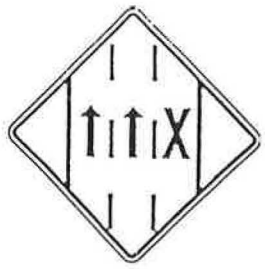
Selection Procedure

Test Subjects Subjects were selected from the age and sex categories previously described. Again, drivers from DMV offices were used. Results of the screening procedure showed that there was no significant variation among the test results at the four testing locations. Therefore, it was decided to test at only one location for this procedure. The Fairfax, Virginia, location was used. To ensure that at least 30 subjects saw each sign, a minimum of 240 subjects had to be tested.

Results The same coding scheme used to tabulate the data from the first procedure was used to analyze the results of the second procedure. Information gathered from the debriefings was used to clarify subjects' written responses. This allowed the subjects' individual responses to be assigned to specific response codes with greater confidence than was the case in the first procedure. Upon probing subjects about some answers that were considered "incorrect" in the first procedure analysis, it was found that these subjects had a functionally correct interpretation of the sign but failed to express it in writing. Therefore, many of the answers previously considered incorrect were counted as correct answers.

The selection procedure results were used to choose the signs to be recommended for use as actual EZSs. The reformatted signs, which were tested in the selection procedure, are shown in Figures 5, 6, and 7.

For the E-1 series the results were significant at the 0.05 level. Sign E1-4 was correctly identified by all of the test subjects (100 percent correct). Signs E-1 and E-3 performed about the same, scoring 77 percent and 73 percent, respectively. Sign E1-3 caused a problem for some of the test sub-

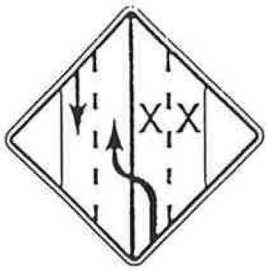


E1-1

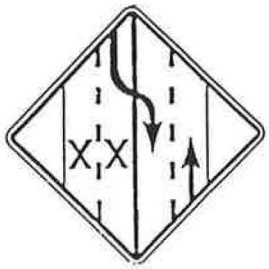
E1-3

E1-4

FIGURE 5 E1 series signs: selection procedure.



E2-1



E2-2

FIGURE 6 E2 series signs: selection procedure.

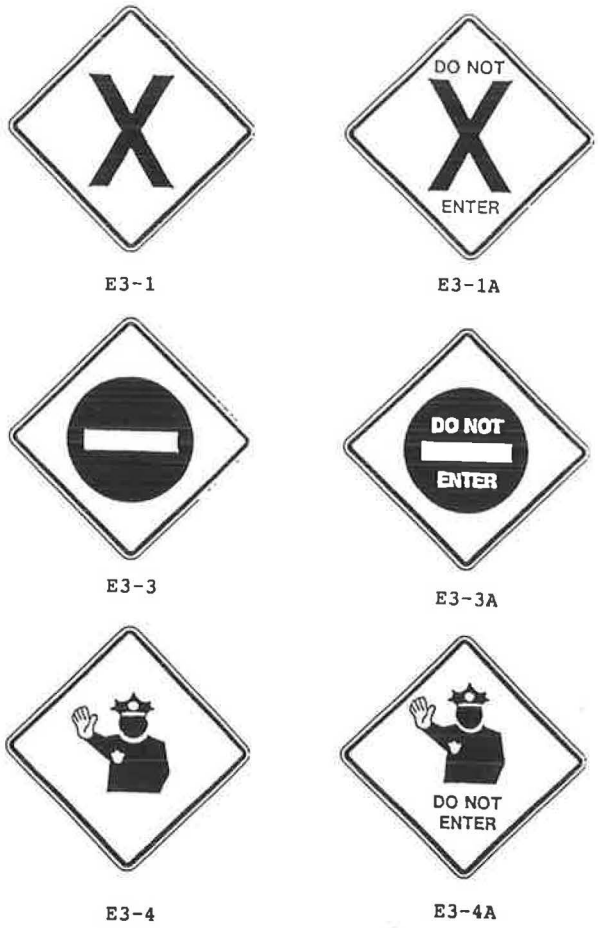


FIGURE 7 E3 series signs: selection procedure.

jects. The message to "merge" was clear, but many subjects did not know how many lanes were closed. That type of confusion did not occur with Signs E1-1 or E1-4. Sign E1-3 does not have the flexibility to warn of a center-lane closure as Signs E1-1 and E1-4 do: the X on Sign E1-1 can be moved from lane to lane to show the closure, and the word RIGHT on Sign E1-4 can be changed to CENTER or LEFT. Detachable arrows, X's, and words would make the use of either of these signs very flexible.

As mentioned previously, the E-2 series of signs, the two-lane crossover subcategory, was tested as part of the One Lane Road Sign (W20-4) candidates in a separate study by Pietrucha and Knoblauch (*13*). The four-lane crossover signs were actually two different signs. One (E2-1) shows the traffic pattern for the driver who is crossing over the centerline, and the other (E2-2) shows the conditions for the driver who is sharing the first driver's side of the road. Sign E2-1 was correctly understood by 94 percent of the subjects, whereas Sign E2-2 was understood by 63 percent of the test group. There seems to be no explanation for the fact that although the signs were similar in concept, there was such a wide disparity in their interpretation.

In the E-3 sign series, all of the hybrid signs (symbol and word messages) performed very well. All three signs were understood by over 96 percent of the test subjects. The symbol-only signs did not perform quite as well. Sign E3-1 was correctly understood by 79 percent of the test group, Sign E3-3 by 77 percent, and Sign E3-4 by 64 percent; however, Sign E3-4 was often misunderstood as involving traffic control by a police officer. The results were considered statistically significant.

Sign Shape and Color

As part of this research project, a specific shape and color for the EZS was considered. On the basis of past research, there were some doubts about motorists' understanding of the shape and color code currently in use (*12*). Since no work has been done to check or update the results of the testing done by Ferguson and Cook (*11*), it was decided to do some limited testing on sign color and shape by repeating their 1967 test to verify their results.

In the Ferguson and Cook technique, blocks of color or a colorless outline of a traffic sign was presented. The test subjects were asked to write down the message or type of information they would expect to see on a sign of the given color or shape. The test subjects were drivers from Virginia DMV offices, high school students, employees of industrial concerns, and members of civic and service organizations. Included in the sample were drivers who had stopped at rest areas along Virginia Interstate highways. There was no apparent effort to control the sample for age or sex. The results showed that, overall, only a few colors and shapes were very well recognized by the population sample.

The testing done as part of this project was an attempt to update the Ferguson and Cook findings, as well as to test other shapes and colors not tested as part of the 1967 study, to determine the comprehension levels associated with these shapes and colors.

In a technique similar to that used by Ferguson and Cook, a group of upper-level engineering undergraduate students at the University of Maryland formed the test sample. Since the Ferguson and Cook results showed relatively low recognition, it was thought that the interpretations of a well-educated, system-oriented audience might yield higher results. Surprisingly, the results were essentially the same. Since the number of subjects tested by Ferguson and Cook was so large, their data were assumed to be the population data or the expected results. The data from the University of Maryland tests were considered the sample or the observed results. A Z-test of statistical significance was performed between the observed and expected results. The Z-test results were considered significant at the 0.05 level.

Table 1 shows the results of the shape test. For the sign shapes tested in both procedures, there are no differences between the results. The octagon was correctly identified most often (89 percent correct). The regulatory rectangle (long axis vertical), the guide rectangle (long axis horizontal), and the pentagon were not tested in the Ferguson and Cook study. In the University of Maryland results, the pentagon was the only shape to have a less than 70 percent recognition level (38 percent correct).

Table 2 shows the results of the color testing. In most cases in which there were comparative data, the results again did not differ significantly. The only exceptions to this were for blue and green. Although orange was tested in the Ferguson and Cook study, it was not considered appropriate to compare

TABLE 1 Subjects Correctly Identifying Sign Shape

Shape	Percent Correct: Ferguson and Cook (n = 1163)	Percent Correct: University of Maryland (n = 37)	Difference Statistically Significant?
Diamond	71	70	No
Rectangle (long axis horizontal)	n/t	73	—
Octagon	89	89	No
Pentagon	n/t	38	—
Triangle	85	84	No
Rectangle (long axis vertical)	n/t	73	—

NOTE: n/t = not tested.

their results with the University of Maryland results because in 1967 orange was not the standard construction and maintenance colorit was used to denote school areas. In the University of Maryland results, the strong yellow-green (SYG) was the only color currently reserved by the MUTCD that was tested. In the past, it had been proposed to use SYG as the background color for the EZS and to remove it from its "reserved" status. Therefore, "correct" results for this color were those that included references to emergency vehicles (8 percent of the responses), general warning (5 percent of the responses), and special route information (3 percent of the responses).

CONCLUSIONS AND RECOMMENDATIONS

It would appear that an EZS would be a useful TCD for emergency situations. The laboratory procedures show that Signs E1-1, E1-4, and Signs 3-1 and 3-3 and their hybrid counterparts have the greatest potential for near-term use on the basis of the levels of understanding associated with these signs. However, before any field deployment under actual operating conditions is contemplated, it is recommended that these signs undergo further testing regarding visibility under closed field conditions.

All of the testing procedures showed that the shape and color coding scheme is not well understood. Of all the concepts tested, the only strong relationships appear to be between the "stop" concept and the use of an octagon and red. Strong recognition also occurred when the customary shapes and colors used for guide signs and motorist services were tested. Some other relationships exist, but they are much weaker.

TABLE 2 Subjects Correctly Identifying Sign Color

Color	Percent Correct: Ferguson and Cook (n = 1163)	Percent Correct: University of Maryland (n = 37)	Difference Statistically Significant?
Red	85	84	No
Orange	n/a	32	—
Yellow	86	76	No
White	48	54	No
Blue	26	54	Yes
Brown	n/t	51	—
Strong yellow-green	n/t	16	—
Green	24	49	Yes

NOTE: n/a = not applicable. n/t/ = not tested.

The question to be answered is "What would be the best shape and color for the EZS?" Since there are no reserved sign shapes and there does not appear to be any strong relationship between the emergency concept and any sign shape, it is recommended that a diamond shape be used for drivers who understand the shape codethe diamond indicates a hazard warning.

It was also important to determine the significance of the recognition and visibility of various colors. It is well known that the reserved SYG color is the best color at night, whereas yellow has the best daytime visibility. Since SYG is currently reserved, it should come as no surprise that the test subjects did not associate it with any traffic sign use.

This is not to say that individuals cannot learn to recognize SYG as representing an emergency situation to a degree, just as they moderately recognize other colors, excluding red. The conclusion is that there appears to be no cognitive reason for using or not using SYG as the EZS color. It is recommended that if a more appropriate use of SYG cannot be found, then on the basis of its superior visibility characteristics, it should be considered for use as the EZS color.

ACKNOWLEDGMENT

The research reported in this paper was sponsored by FHWA, U.S. Department of Transportation. The contents reflect the views of the author, and do not necessarily reflect the official views or policies of the U.S. Department of Transportation.

REFERENCES

1. *Collision Management Procedures.* Maryland Police Training Commission, Pikesville, Md., 1975.
2. *Administrative Guide for Post Traffic Crash/Incident Procedures.* Wilbur Smith and Associates, Columbia, S.C., 1972.
3. Post, D. V. *Signal Lighting System Requirements for Emergency, School Bus and Service Vehicles.* Highway Safety Research Institute, University of Michigan, Ann Arbor, Nov. 1978.
4. Stocia, T. L. *Evaluation of Semi-marked Police Vehicles.* Bureau of Planning and Development, Illinois Department of Law Enforcement, Springfield, April 1983.
5. Weckesser, P.M. Evolution and Use of Hazard Warning Signs on the New Jersey Turnpike. In *Proceedings of the International Symposium on Traffic Control Systems,* Vol. 2C: *Driver Communication and Information,* Institute of Transportation Studies, University of California, Berkeley, Aug. 1979.
6. Hanscom, F. R. *NCHRP Report 235: Effectiveness of Changeable Message Displays in Advance of High-Speed Freeway Lane Closures.* TRB, National Research Council, Washington, D.C., Sept. 1981.
7. Roper, D. H. *NCHRP Synthesis of Highway Practice 156: Freeway Incident Management.* TRB, National Research Council, Washington, D.C., Dec. 1990.
8. Owen, H. Q. Vehicle-Mounted Traffic Safety Sign. *Better Roads,* Sept. 1973.
9. *Manual on Uniform Traffic Control Devices for Streets and Highways.* U.S. Department of Transportation, 1988.

10. McGee, H. W., W. Moore, B. Knapp, and J. H. Sanders. *Decision Sight Distance for Highway Design and Traffic Requirements*. BioTechnology, Inc., Falls Church, Va., 1978.
11. Ferguson, W. S., and K. E. Cook. *Driver Awareness of Sign Color and Shapes*. Virginia Highway Research Council, Charlottesville, Va., May 1968.
12. Markowitz, J., C. W. Dietrich, W. J. Lees, and M. Farman. *An Investigation of Design and Performance of Traffic Control Devices*. FHWA, U.S. Department of Transportation, Dec. 1968.
13. Pietrucha, M. T., and R. L. Knoblauch. *Motorist Comprehension of Regulatory, Warning, and Symbol Signs*. FHWA, U.S. Department of Transportation, 1985.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Evaluation of the Federal Vision Standard for Commercial Motor Vehicle Operators

LAWRENCE E. DECINA AND MICHAEL E. BRETON

A reassessment was made of the adequacy of the current federal interstate vision standards for commercial motor vehicle operators. The technical approach included a critical review of existing literature, development of draft recommendations, delphi-approach surveys, a workshop to review draft recommendations with expert truck industry and vision panelists, and a report with final recommendations. No compelling evidence was found in the research literature on the vision performance of passenger and commercial drivers to warrant substantial change to the current standard. However, a number of problems in the current standard were identified during the literature review and at the workshop. The requirements for distant visual acuity remain at least 20/40 in each eye without corrective lenses or visual acuity separately corrected to 20/40 or better with corrective lenses and distant binocular acuity of at least 20/40 in both eyes with or without corrective lenses. The requirement for field of vision was revised to at least 120 degrees in each eye measured separately in the horizontal meridian. The standard also now states that a driver should have the ability to respond safely and effectively to the color of traffic signals and devices showing standard red, green, and amber, although no test for color vision is required. The instructions to perform and record the visual examination were extensively revised as were the identification of the type of equipment, specification of stimuli needed to conduct testing, and instructions on how to perform tests. In addition, revisions were made to the list of visual disorders and impairments to be noted on the exam form.

There is widespread agreement that vision plays an essential role in the driving task. However, the level of vision that is necessary for safe driving continues to be a contentious issue. The reason for this is the continuing unavailability of definitive empirical evidence upon which to base a clearly defensible visual performance standard. The purpose of setting vision standards for drivers of commercial motor vehicles (CMVs) is to identify individuals who will represent an unreasonable and avoidable safety risk if allowed to drive CMVs. The research objective in support of a vision standard has been to identify the level of seeing, based on empirical evidence in place of a consensus, that has to be met so that CMV drivers will not be a safety risk to themselves or to the motoring public.

Driving safety is maintained through a constant stream of small decisions and less frequent larger decisions that require a high rate of accurate visual information about the driving environment. The level of vision required to support success in the decision-making process and driving safety depends on the level of complexity of the projected driving task (i.e., high-speed, wide-open highway compared with congested urban or suburban roadway environments). It also depends on the consequence of encountering an error, or series of errors, in the decision stream that will lead to a catastrophic outcome for the driver and others in the driving environment. For drivers of CMVs, the consequence of error is likely to be much greater in terms of loss of life and property than the result of a similar error made by the driver of a private motor vehicle. This fact is supported by the statistics accumulated from 1979 to 1986 on the disproportionately high rate of heavy-vehicle involvement in fatal crashes. For all types of accidents (adjusted for exposure mileage), combination trucks (tractor-trailer combinations) have slightly less than 50 percent of the accident involvement rate of passenger cars but have a fatality involvement rate that is nearly twice that of passenger cars (*1*). In fact, in 1990, 4,061 people died in tractor-trailer crashes. However, only 12 percent were truck occupants; the majority of these fatalities were passenger vehicle occupants (*2*).

Driving errors that might not produce a crash in a smaller motor vehicle may well lead to a crash in a heavy vehicle because of its more limited maneuverability. The appreciation of this fact motivates the effort to define visual standards for driving that are most likely to lead to safer driving. In addition, the apparently greater difficulty of the CMV driver's vehicle control task and the obviously greater adverse consequences of heavy-vehicle crashes lead to the presumption that the visual requirements for the driver of a CMV should be more stringent than those thought to be appropriate for smaller vehicles. This view is reflected in the existing federal interstate vision standard for CMV operators.

The current need to reassess the bases for the federal vision standard for CMV operators was motivated by a number of factors, including inaccuracies in the current standard, claims that current standards unfairly discriminate against some drivers, and emerging trends in vision assessment technology and vision-driver performance evaluation methods. The initiative for this research was set by the Federal Highway Administration's Office of Motor Carriers.

The technical objectives for the reassessment of the federal vision standard for CMV operators were

- Critical review and evaluation of the current federal vision standard (*3*) scientific information and data sources pertaining to driver vision testing requirements for operating CMVs that weigh more than 10,000 lb,
- Development of preliminary recommendations for revising vision test and testing requirements and testing procedures,

L. E. Decina, KETRON Division, Bionetics Corporation, Great Valley Corporate Center, 350 Technology Drive, Malvern, Pa. 19355-1370. M. E. Breton, Scheie Eye Institute, University of Pennsylvania, Myrin Circle, 51 North 39th Street, Philadelphia, Pa. 19104.

- Conducting a delphi-approach opinion survey with vision and industry experts to assess the most important visual functions for critical CMV driving tasks,
- Conducting a workshop to review draft recommendations with panelists representing industry and the visual science community, and
- Summarizing project findings, including final recommendations for the vision test requirements and testing procedures.

LITERATURE REVIEW

The literature review covered a comprehensive analysis of the history of the standard, published research, and selected unpublished project data on the relationship between driving and vision performance, identification of state and international standards, and published recommendations from the medical community.

History of Standard

The federal government began regulating vision standards for motor carriers in interstate commerce during the late 1930s. At that time, the standard was based on a consensus of experts in the fields of vision and driver safety. The vision standard has been changed steadily in the direction of requiring more stringent visual capability (Table 1). The standard (*3*) as currently stated calls for "distant visual acuity of at least 20/40 (Snellen) in each eye without corrective lenses or visual acuity separately corrected to 20/40 (Snellen) or better with corrective lenses, distant binocular acuity of at least 20/40 (Snellen) in both eyes with or without corrective lenses, field of vision of at least 70 degrees in the horizontal meridian in each eye, and the ability to recognize the colors of traffic signals and devices showing standard red, green, and amber." Along with the lack of an empirical base for the visual measures used for the standard, there were problems with major inaccuracies of the visual field requirement. The current standard states that a 70-degree field of view is the minimum requirement for each eye. This is obviously erroneous since the field of view in a normal healthy adult is closer to 140 degrees for each eye. In addition, problems were found with the color vision requirement, which on a practical basis is probably unenforceable. The color requirement as now stated would not exclude red-green color-defective drivers since the standard does not provide adequate instruction on requirements for color vision testing. It is also doubtful that the standard intended to exclude typical red-green color-defective drivers since these drivers are currently on the road and there is a lack of evidence that their driver safety record is worse than the record of those without such color vision defect. In addition, one of the major problems with the standard is the lack of an adequate description of the specificity of testing stimuli, lighting conditions, equipment, or uniformity of testing procedures. The standard also does not provide any direction on uniformity of testing procedures.

Empirical Evidence: Driving and Vision Performance

A comprehensive literature review was undertaken to identify research that reported measurements of the relationship between many aspects of visual performance and accessible indicators of driving safety. The studies identified were primarily *post hoc* analyses of data already accumulated through routine driver registration testing and record keeping. However, some studies introduced into the driver testing routine novel controlled vision testing methods designed to obtain data on a broad scale that could then be correlated with the driving record over time. The literature search found numerous research projects that examined the relationship between vision test results for operators of motor vehicles and their driving performance record (i.e., accidents and violations), dating back to the mid-1950s. Most of these studies were initiated to determine what visual skills best correlate with driving performance. The results were used to recommend to state licensing agencies the most practical vision tests to administer to license applicants and renewals. Many of the studies focused on vision tests that were easily accessible through commercial vision screening devices. However, some of the

TABLE 1 History of the Visual Standard for CMV Operators

	Visual Acuity			Visual Fields		Color Vision			
Year	One Eye	Other Eye	Binocular	All Meridians	Horizontal Meridians	Red, Green	Yellow	Amber	Other Notes
1937 *(4)*	"Good eyesight in both eyes (either without glasses or by correction with glasses) including adequate perception of red and green colors"								
1939 *(5)*	20/40	20/100	--	45 degrees	--	Yes	Yes	--	--
1944 *(6)*	20/40	20/100	--	45 degrees	--	Yes	Yes	--	--
1964 *(7)*	20/40	20/40	--	--	140 degrees (Binocular)	Yes	Yes	--	Drivers requiring correction by glasses are required to wear them while driving.
1970 *(8)*	20/40	20/40	20/40	--	70 degrees (each eye)	Yes	---	Yes	
1985 *(3)*	20/40	20/40	20/40	--	70 degrees (each eye)	Yes	---	Yes	If driver wears contacts, evidence to indicate good tolerance.

studies involved developing customized vision testing apparatus, and some used clinical testing equipment known to be impractical for mass vision screening in a licensing bureau environment. Most of the research identified and reviewed focused on the passenger vehicle operator and only a few studies investigated the visual and driving performance of the CMV operator.

Passenger Vehicle Operators

One of the earliest, most comprehensive studies on the relationship between vision and the driving performance record was conducted by Burg (*9–12*) on more than 17,500 drivers over a 3-year period in the 1960s. Driving habits (annual mileage reported), age, and gender were reported in addition to information on their vision test performance for dynamic visual acuity, static visual acuity, lateral visual field, low-light recognition thresholds, glare recovery, and sighting dominance. Of the vision tests analyzed in relation to traffic convictions and accidents (reported), very weak statistically significant correlations were found between vision and the driving performance record. Like other researchers from the 1960s (*13,14*), Burg reported that mileage and age were the most powerful predictors of traffic accidents and convictions. Further analysis of the Burg data by Hills and Burg in 1977 (*15*) revealed a small but significant correlation between static and dynamic visual tests and glare recovery tests and accident rates for drivers over age 54.

In the early 1970s, the U.S. Department of Transportation (DOT) was interested in the results of the Burg studies. DOT initiated a series of investigations designed to develop a battery of vision tests that were more functionally related to driver performance and safety and that could lead to the development of a vision testing device for use in screening driver's license applicants or renewals. In this study, Henderson and Burg (*16*), after reviewing prior literature and analyzing earlier data, provided a systematic analysis of the visual requirements for driving. The initial phase of the study identified important visual functions: static visual acuity (normal illumination), central angular movement, central movement-in-depth, useful peripheral vision, static acuity (low-level illumination), field of view, eye movement and fixation, dynamic visual acuity, accommodation faculty, and glare sensitivity. These visual functions were incorporated into a prototype vision testing device (the MARK I Vision Tester). Over 600 license renewal operators were screened on the device. Accident statistics were collected for the preceding 3 years for each operator. Results showed a moderate, consistent, age-related decline for all the visual functions. Significant age-related loss in visual ability was reported for static acuity under normal and low illumination, glare, and dynamic acuity. However, the correlational analyses conducted to assess the potential predictive validity of the MARK I showed many significant correlations in the direction of poor visual performance statistically related to a good driving record.

DOT, encouraged by some of the results of the MARK I study, decided to continue this research in an effort to establish a generally valid vision screening device for motor vehicle department use. Further testing by Shinar (*17–19*) on 890 licensed operators revealed very low correlations between accident rate measures and visual performance. In fact, no significant correlation existed between vision and driving records for the 25 to 54 age group. Additional testing indicated that poor dynamic and static visual acuity under low levels of illumination was most consistently related to accidents; poor static acuity under low levels of illumination was related to nighttime accidents. There was also a relationship between central angular movement and accident involvement. In addition, none of the single vision tests was significantly associated with accident involvement for all age groups, but each test was significantly associated with accident involvement for one or more of the age groups. Results for the battery of vision tests and the driving statistics did not establish a clear-cut relationship between specific visual tests and the driving record.

Another important effort conducted around the same period by Hofstetter (*20*) correlated the visual acuity test scores of 13,700 drivers with self-reported accidents during the previous 12-month period. Data were collected nationally over a period of 10 years by means of a survey form given out in a variety of settings and populations, with support from the Auxiliary to the American Optometric Association, using commercial vision screeners. Accident rates for persons with acuity in the lower quartile of the measurements were compared with rates for persons with acuity above the median measurement. Drivers in the lower visual acuity group were found to be twice as likely to have had three accidents in the previous year as those with acuity above the median, and 50 percent were more likely to have had two accidents. No significant differences were found between the lower-acuity and higher-acuity drivers when only one accident was used as the criterion of comparison. This study provided some evidence for the connection between poor visual acuity and increased accident frequency. However, these results applied only to the very poor visual performers compared with the best in the driver cohort.

Studies on visual fields and glare were also conducted in the 1970s. Council and Allen (*21*) compared horizontal visual field measurements with accident rates for more than 52,000 drivers and found that only 1 percent of the drivers recorded a horizontal field of 120 degrees or less and that the accident rate for these drivers was no higher than the rate for those whose fields were greater than 120 degrees. Studies on glare sensitivity incorporated into other vision testing using the MARK I and MARK II (*17*) devices were also unable to show any significant relationship. Wolbarsht (*22*) conducted a study of glare sensitivity using a modified commercial vision screener with a customized overlying glare source of controllable intensity. He tested 1,500 driver's license applicants and renewals for glare sensitivity at three veiling glare ratios (background:target) of 2:1 (high glare), 4:1 (medium glare), and 8:1 (low glare). The results showed no significant correlation between glare sources and driving performance, although the average glare sensitivity scores did increase with age.

Research on assessing visual and driving performance continued in the 1980s. Keltner and Johnson (*23*) used automated static perimetry to screen more than 500 drivers for any evidence of visual field loss in 1980. With this technique it was found that approximately 5 percent of the motorists had significant visual field loss compared with only 1 percent found

to have a noticeable deficit in the study by Council and Allen (*21*), who tested only in the horizontal meridian. In addition, Keltner and Johnson reported that subjects over age 65 had four to five times the incidence of visual field deficits of younger persons. For the Keltner and Johnson study, field loss was defined as substantial depression of all or part of the peripheral visual field or an inability to detect two or more adjacent visual field points (scotoma), or both. This project was extended (*24*) to compare the visual field loss of 10,000 volunteer drivers with accident and conviction histories. For this larger study, it was found that drivers with visual field loss in both eyes had accident and conviction rates that were twice as high as those for drivers with normal visual fields. The results were statistically significant. It was suggested that decreased performance on a visual field test probably results from age-related decreases in retinal illumination and other acquired vision impairments (such as glaucoma, degenerative myopia, diabetic retinopathy, and retinal detachment) that are more common in older age groups.

Another study, conducted by Davison (*25*) in 1985, examined 1,000 motorists who were randomly stopped in and around a town in England and asked to volunteer for a vision test and provide information on driving record, vision examination history, and other demographic information. Significant positive associations were found between accidents and right-eye or left-eye visual acuity and binocular acuity for all drivers, and a relationship was found between accidents and heterophoria for drivers who were over 55. Decina et al. (*26*) recently completed a study for the Pennsylvania Department of Transportation to determine the value and feasibility of periodic vision screening during license renewal. The study examined the relationship of three vision measures (static visual acuity, horizontal visual fields, and contrast sensitivity) to accident and violation records for over 12,400 licensed operators, who were unaware that they would be tested. It was found that drivers who failed the Pennsylvania Department of Transportation visual standard or scored below "normal" on the contrast sensitivity test were at a significantly higher risk for accidents in only the two oldest age groups (66 to 76 and 76+). However, the researchers found no significant relationship between poor vision performance on each of the vision tests analyzed separately with accident and violation records.

For the most part, significant statistical relationships between specific vision test scores and driver performance records (for passenger vehicles) were not clearly established in the literature. Many researchers found it difficult to relate driving performance to visual capabilities; some of the more important difficulties were as follows:

- Vision is only one of many factors influencing driving performance,
- Some vision tests may not relate closely to visual requirements of driving,
- Reliability of criteria used to measure driving performance may be low,
- Samples of the driving population may be unrepresentative, and
- Individuals with visual difficulties often place self-imposed limits on their driving, reducing their exposure to the risk of an accident and biasing statistical sampling.

CMV Operators

In 1973, Henderson and Burg attempted to relate CMV driving skills to the visual tests included in the MARK I Vision Tester (*16*). Their goal was to establish a sound scientific basis for minimum visual standards for the Office of Motor Carriers. The relative importance of different aspects of the driving task was established by examining literature, interviewing truck drivers, observing truck drivers in action, and conducting a systematic examination of the driving task. The researchers established a hierarchy of importance for the visual functions selected as most important. Weights were assigned to various driving behaviors and to each visual function according to its judged importance to driving behavior. Those visual functions judged to be most important to the truck driving task and necessary to an analysis comparing visual performance and accidents and violations were static visual acuity; dynamic visual acuity; perception of angular movement; perception of movement-in-depth, visual field, movement-in-depth and steady, saccadic, and pursuit fixations; glare sensitivity; and angular movement. Significant relationships between accidents and poor visual performance were found only with measures of perception of movement and dynamic visual acuity. No correlation was found between static visual acuity or field of view and accident frequency for commercial drivers.

In a more recent attempt to correlate visual performance with accident record, Rogers et al. in 1987 (*27*) compared the driving records of visually impaired heavy-vehicle operators with the records of a sample of visually nonimpaired heavy-vehicle drivers. The purpose of the project was to determine whether the federal vision standard could be justified on the basis of the traffic safety record of these drivers. The records of more than 16,000 heavy-vehicle operators registered by the California Department of Motor Vehicles were examined. Measures of driving performance consisted of 2-year total accidents and convictions associated with incidents involving commercially registered vehicles. Visually impaired operators were categorized into two subgroups of substandard static acuity: (*a*) moderately visually impaired (corrected acuity between 20/40 and 20/200 in the worse eye and 20/40 or better in the other), and (*b*) severely visually impaired (corrected acuity worse than 20/200 Snellen in the worse eye and 20/40 or better in the other). Nonimpaired drivers met current federal acuity standards (corrected acuity of 20/40 or better in both eyes). Analysis results, adjusted for age, showed the following:

- Visually impaired drivers had a significantly higher incidence of total accidents and convictions and commercial-plate accidents and convictions than did nonimpaired drivers.
- Moderately impaired drivers had a significantly higher incidence of commercial-plate accidents than did nonimpaired drivers.
- The incidence of total accidents did not significantly differ between the nonimpaired and moderately impaired drivers.
- Severely impaired drivers had a significantly higher incidence of commercial-plate convictions than did nonimpaired drivers.
- Nonimpaired and moderately impaired drivers did not significantly differ on commercial-plate convictions.

- Drivers licensed to operate any combination of heavy vehicles had a higher incidence of total accidents and convictions and commercial-plate accidents and convictions than did those licensed to operate single vehicles having three or more axles.

These findings led to qualified support for the current federal visual acuity standard, particularly regarding exclusion from driving of the severely impaired (visual acuity below 20/200 in the worse eye and 20/40 or better in the other). Less support is offered regarding the restriction of the moderately visually impaired heavy-vehicle operator (visual acuity between 20/40 and 20/200 in the worse eye and 20/40 or better in the other).

Another recent study identified in the literature assessing the relationship between vision and truck operator performance was conducted by McKnight et al. (*28*), who examined monocular and binocular visual and driving performance of tractor-trailer drivers. On the visual measures, the monocular drivers were significantly deficient in contrast sensitivity, visual acuity under low illumination and glare, and binocular depth. However, monocular drivers were not significantly deficient in static or dynamic visual acuity, visual field of individual eyes, or glare recovery. In addition, no differences were shown between monocular and binocular drivers on driving measures of visual search, lane keeping, clearance judgment, gap judgment, hazard detection, and information recognition. The one exception was sign-reading distance, which was defined as the distance at which signs could be read during both day and night driving in a controlled road test. The binocular drivers were first able to read road signs at significantly greater distances than were the monocular drivers in both daytime and nighttime driving, and this decrement correlated significantly with the binocular depth perception measure. McKnight also reported a large variation in visual and driving measures among monocular drivers and several significant differences between them and binocular drivers, suggesting the need to assess the monocular drivers' visual functioning capabilities more closely and to continue research in identifying visual performance measures that significantly correlate with measures of safe driving skills.

Summary of Literature Results

The studies reviewed represent a substantial accumulation of data on the relationship of vision to driver (passenger and heavy vehicle) performance. No single study provided support for definitive changes to the current federal commercial motor vehicle vision standard. Nevertheless, it was equally apparent that changes in terms of both more and less stringent requirements in several performance areas should be evaluated at this time with the minimum aim of encouraging further empirical work. In addition, it is apparent that a large gap exists between the current standard and its uniform and effective implementation at the level of routine practical testing. Even though little evidence appears to exist to support a substantial and direct relationship between vision and driver safety, much evidence has been accumulated to support the hypothesis that vision, in interaction with other factors, contributes in a critical way to influence highway safety.

State and International Visual Standards

State CMV vision standards applying only to intrastate driving were reviewed. The requirements for each state are generally less stringent than the current federal CMV standard. The binocular visual acuity requirement in almost 80 percent of the states is 20/40, but less than 10 percent of the states deny a license for monocularity. Less than 40 percent of the states have visual field standards comparable with the federal standard, and only 24 percent have a color standard (*29*). Review of vision standards for CMVs in other industrialized countries revealed wide variances. Most countries require a visual acuity level for each eye separately that is more stringent than the current U.S. standard of 20/40 in each eye. Only a few countries have a binocular acuity requirement, and when specified, it is more stringent than the U.S. requirement. For visual fields, most other countries state that the driver must have "normal" or "full" fields. Most other countries do not have a requirement for color vision. In addition, the driving privilege in many countries may be denied because of stereopsis, aphakia, diplopia, high myopia, night blindness, and nystagmus. Many countries also require periodic vision checks.

Medical and Government Recommendations

The American Medical Association (AMA) has participated in setting vision standards for CMV operators and has provided guidelines (*30*) for vision testing to its members. The guidelines published in 1986 differ from the federal vision standard in excluding high-power spectacle lenses (10 diopters or greater) and in requiring visual acuity in each eye of 20/25 or better compared with 20/40 for the CMV standard. In addition, other visual disorders are discussed, including stereopsis, nighttime vision, diplopia, and oscillopsia, but specific recommendations for excluding drivers with these conditions are avoided.

The National Highway Traffic Safety Administration of the U.S. Department of Transportation, in cooperation with the American Association of Motor Vehicle Administrators, published a booklet in 1980 (*31*) that presented a set of recommendations for all drivers otherwise medically capable of operating commercial vehicles, including heavy trucks. The recommendation for visual acuity differs from the federal vision standard but is the same as that proposed by the AMA (i.e., 20/25 or better is required in each eye, not 20/40 as specified in the federal standard). The recommendation for visual fields is specified as 140 degrees for each eye in the horizontal meridian. The recommendation for color vision is the same as the federal vision standard and AMA recommendations (i.e., ability to distinguish red, green, and yellow/amber). The booklet provides recommendations for visual acuity, visual field, ocular motility, color discrimination, depth perception, dark adaptation, refractive states, and strabismus (crossed eyes).

EXPERT OPINION SURVEY

An expert opinion survey was conducted because of the dearth of reliable data relating visual assessmenteither clinical ex-

amination or screening by a Department of Motor Vehicles protocolto the driving record. Accordingly, using a delphi-type approach with a panel of visual and truck industry experts, specific visual functions deemed most important for safely performing each of seven critical CMV driving tasks were initially identified. This information established minimum acceptable performance levels for each visual function for each driving task.

The approach used an iterative process in which the most frequent response for visual functions ranked by order position (most important, second most important, third most important, etc.) was tabulated for each driving task; this information was then made available to each panel member, and further responses from each person were requested as needed to resolve ties and achieve consensus for all rankings. Three iterations of this process were required, resulting in the collective judgments. Panelists also provided subjective (rating scale) evaluation of the relative safety of matched monocular and binocular drivers with respect to the seven critical CMV driving task response capabilities. Table 2 presents the results of these two surveys. Finally, panelists were able to express their opinion on visual disorders and ocular conditions that should be noted on a physical examination form and that should require a follow-up exam by a vision specialist.

WORKSHOP CONSENSUS

A workshop was conducted to review and provide a consensus on preliminary draft recommendations. The panel represented the truck industry and the visual science community and consisted of licensed doctors of medicine, ophthalmologists, optometrists, professors of ophthalmology, and traffic and safety professionals in private industry. Focused discussion was held on the most vital points at issue, including the need to exclude monocular drivers or those with substantial visual loss in one eye only, the statement of the visual field requirement, the need for more complete and accurate testing of visual field (more in accord with the medical diagnostic procedure), the benefit of including newer tests of vision, the intent and effectiveness of the current color vision standard, and the basis of a risk analysis model that could be used to evaluate changes to the standard. The workshop panelists concluded that there were no compelling reasons to change the current binocular visual acuity standard of 20/40, that there was a need to measure horizontal visual fields using a more rigorous method than that currently employed in commercial vision screening equipment, and that the current color vision requirements are unenforceable and do not meet the intent of not excluding red-green color-defective individuals from the driving privilege. Most panelists agreed that the testing procedures for measuring acuity and visual field needed to be more comprehensive. Visual acuity optotypes, background illumination, and target luminance should follow the procedures recommended by the National Academy of Sciences (*32*). Specifying visual field target size and luminance was recommended, and the need for a test procedure that would provide a repeatable and accurate measure of field limits in the horizontal meridian was discussed. In addition, doubt was expressed about risk, if any, presented by drivers who are color blind, since traffic signing has been standardized and drivers have many other cues for the operation of a vehicle in a safe and effective manner. Panelists generally believed that it was important to note visual disorders and ocular conditions and that individuals with specific conditions should be referred to ophthalmologists.

Panelists participated in post-workshop evaluation of visual acuity, visual field, and color vision standards. Panelists were asked to select specific alternative wording for each requirement of the standard. The wording of the final recommended standard conforms to the majority choice for each requirement.

FINAL RECOMMENDATIONS

On the basis of the review of the literature, delphi exercise, and workshop views of the panelists, the recommended changes to the CMV standard were amended as follows. The statement of the visual acuity standard was found to be adequate. More specific wording to rule out below-standard performance in one eye was added to the Instructions for Performing and Recording Physical Examinations. Extensive revisions were made to this section to specify more completely the testing conditions and procedures to be used when measuring acuity, including light level, stimulus type, and specific test procedures. The statement of the visual field standard was changed to require at least a 120-degree field of view in each eye measured separately in the horizontal meridian. Extensive revisions were also made to the Instructions section to specify minimum stimulus conditions and an acceptable procedure

TABLE 2 Visual Functions Judged Most Important for Safely Performing Seven Critical CMV Driving Tasks

	Visual Function by Order of Importance			
Driving Task	1	2	3	Binocularity Critical
Maintaining safe speed for conditions	Visual fields	Motion Perception	Contrast Sensitivity	Yes
Maintaining safe following distance	Depth perception	Motion Perception	Visual Fields	No
Staying in lane/steering control	Visual fields	Static acuity	Contrast Sensitivity	No
Merging/Yielding in traffic conflict situations	Visual fields	Visual search/Attention	Motion Perception	Yes
Changing lanes and passing	Visual fields	Depth perception	Motion Perception	Yes
Complying with traffic control devices	Static acuity	Visual fields	Contrast Sensitivity	Yes
Backing up/Parking operation	Depth perception	Visual fields	Contrast Sensitivity	Yes

for testing in the horizontal meridian. The statement of color vision was changed to require only a "safe and effective response" to colored traffic signals and devices, without a specific test of color vision. Under this statement, red-green color-deficient individuals who can otherwise respond safely and effectively (virtually all) will be allowed the driving privilege.

PROPOSED STANDARD

If all recommendations are accepted as visual standards for CMV operators, they could be incorporated into the Code of Federal Regulations as follows (proposed changes in bold type):

391.41 Physical qualifications for drivers.

(b) A person is physically qualified to drive a motor vehicle if that person . . . (10) Has distant visual acuity of at least 20/40 in each eye without corrective lenses or visual acuity separately corrected to 20/40 or better with corrective lenses, distant binocular acuity of at least 20/40 in both eyes with or without corrective lenses, **field of vision of at least 120 degrees in each eye measured separately in the horizontal meridian, and the ability to respond safely and effectively to colors of traffic signals and devices showing standard red, green, and amber. No test for color vision is required**.

391.43 Medical examination; certificate of physical examination.

(a) Except as provided in paragraph (b) of this section, the medical examination shall be performed by a licensed doctor of medicine or osteopathy.

(b) A licensed optometrist may perform as much of the medical examination as pertains to visual acuity, field of vision and the ability to respond appropriately to traffic signals and devices as specified in paragraph (10) of 391.41(b).

(c) The medical examination shall be performed, and its results shall be recorded, substantially in accordance with the following instructions and examination form.

INSTRUCTIONS FOR PERFORMING AND RECORDING PHYSICAL EXAMINATIONS

Head-Eyes

The recommended procedure for testing visual acuity is based on the standard procedures recommended for clinical measurement as reported by the Committee on Vision of the National Academy of Sciences (1980). The standard optotype is the Landolt ring. However, other equivalent optotypes, such as the Sloan letters as a group, are acceptable. Logarithmic sizing should be used (i.e., successively larger sizes should be 1.26 times larger than the preceding size). Optotype letters should be black on a white background of 85 to 120 cd/m². Under these conditions, acuity should be defined as the smallest size at which 7 out of 10 (or 6 out of 8) letters are correctly identified at a given distance. Effective viewing distance should not be less than 4 meters. Regardless of viewing distance, acuity should be specified in terms of a fraction with 20 as the numerator and the smallest type that could be read at 20 feet as the denominator (i.e., 20/20 or 20/40). Although the Snellen chart departs from the standard in several ways, it is acceptable if no practical means of following the recommended procedure is available. If the applicant wears corrective lenses, these should be worn while applicant's visual acuity is being tested. If appropriate, indicate on the Medical Examiner's Certificate by checking the box, "Qualified only when wearing corrective lenses." **The recommended procedure for testing visual fields requires equipment that is able to present a round, luminous stimulus of 0.15 to 0.25 degrees in angular extent on a low photopic background of 1 to 10 cd/m². Stimulus luminance should be 50 to 100 cd/m² and duration should be in the range of 100 to 200 msec. Subject fixation should be verifiable. Multiple presentation in random sequence under monocular test conditions must be possible. This will normally require separate test stimulus positions for determining temporal and nasal field limits. Testing must be monocular with one eye blocked. The test procedure should present the nasal and temporal limits (70 degrees to 80 degrees temporal and 50 degrees to 40 degrees nasal) a minimum of 3 times each in a random alternating sequence. Responses are best recorded automatically. If the applicant wears corrective lenses, these are not required to be worn while applicant's visual fields are being checked.**

Note aphakia, cataract, corneal scar, exophthalmos, **glaucoma, macular degeneration**, ocular muscle imbalance, ptosis, **retinopathy**, strabismus uncorrected by corrective lenses, and **any other conditions deemed important. Individuals with no vision in one eye or vision below standards in one eye as specified in paragraph (1) of 391.41(b) are disqualified to operate commercial motor vehicles under existing federal Motor Carrier Safety Regulations**. If the driver habitually wears contact lenses, or intends to do so while driving, there should be sufficient evidence to indicate that the individual has good tolerance and is well adapted to their use. The use of contact lenses should be noted on the record.

ACKNOWLEDGMENT

This research was sponsored by the Office of Motor Carriers, FHWA. The contracting officer's technical representatives were Eliane Viner and Richard Schwab.

REFERENCES

1. *Special Report 223: Providing Access for Large Trucks*. TRB, National Research Council, Washington, D.C., 1989.
2. Tractor-Trailers. In *Fatality Facts 1991*, Insurance Institute for Highway Safety, Washington, D.C., 1991.
3. *Federal Register*, 49, Section 391.43(b)(10), Oct. 1, 1985, p. 408.
4. *Federal Register*, 57, Section 40, Feb. 28, 1992, p. 6793.
5. *Federal Register*, 4, Section 1.22, June 7, 1939, p. 2295.
6. *Federal Register*, 9, Section 192.2(b), 1944.
7. *Federal Register*, 29, Section 191.2(b), July 3, 1964, p. 8420.
8. *Federal Register*, 35, Section 391.41(h)(10), April 22, 1970, p. 6463.
9. Burg, A. Vision and Driving: A Summary of Research Findings. In *Highway Research Record 216*, HRB, National Research Council, Washington, D.C., 1968.
10. Burg, A. Relation Between Vision Quality and Driving Record. In *Proceedings, Symposiums on Visibility in the Driving Task*,

Highway Research Board, Illuminating Engineering Research Institute, and Texas A&M University, 1968, pp. 5–16.
11. Burg, A. Vision and Driving: A Report on Research. *Human Factors*, Vol. 13, 1971, pp. 79–87.
12. Burg, A., and R. S. Coppin. Visual Acuity and Driving Record. In *Highway Research Record 122,* HRB, National Research Council, Washington, D.C., 1966, pp. 1–6.
13. Crancer, A., and P. A. O'Neal. *Comprehensive Vision Tests and Driving Record.* Report 28. Washington State Department of Motor Vehicles, 1969.
14. Forbes, T. W. *Human Factors in Highway Traffic Safety Research*. Krieger Publishing Company, Florida, 1972.
15. Hills, B. L., and A. Burg. *A Re-analysis of California Driver-Vision Data—General Findings.* TRRL Report 768. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1977.
16. Henderson, R. L., and A. Burg. *Vision and Audition in Driving.* Final Report. NHTSA, U.S. Department of Transportation, 1974.
17. Shinar, D. *Driver Visual Limitations: Diagnosis and Treatment*. Institute for Research in Public Safety, Indiana University School of Public and Environmental Affairs, Bloomington, 1977.
18. Shinar, D. Driver Vision and Accident Involvement: New Findings with New Vision Tests. In *Proceedings of the American Association for Accident and Traffic Medicine VII Conference* (D. F. Herelke, ed.), American Association of Automotive Medicine, Ann Arbor, 1978, pp. 81–91.
19. Shinar, D., and J. W. Eberhard. Driver Visual Requirements: Increasing Safety Through Revised Visual Screening Tests. In *Proceedings of the 20th Conference of the American Association for Automotive Medicine*, 1976, pp. 241–252.
20. Hofstetter, H. W. Visual Acuity and Highway Accidents. *Journal of the American Optometric Association,* Vol. 47, 1976, pp. 887–893.
21. Council, F. M., and J. A. Allen, Jr. *A Study of the Visual Relationship to Accidents*. Highway Safety Research Center, University of North Carolina, Chapel Hill, 1974.
22. Wolbarsht, M. L. Tests for Glare Sensitivity and Peripheral Vision in Driver Applicants. *Journal of Safety Research*, Vol. 9, 1977, pp. 128–139.
23. Keltner, J. L., and C. A. Johnson. Mass Visual Field Screening in a Driving Population. *Ophthalmology,* Vol. 87, 1980, pp. 785–792.
24. Johnson, C. A., and J. L. Keltner. Incidence of Visual Field Loss in 20,000 Eyes and Its Relationship to Driving Performance. *Archives of Ophthalmology*, Vol. 101, 1983, pp. 371–375.
25. Davison, P. A. Interrelationships Between British Drivers' Visual Abilities, Age, and Road Accident Histories. *Ophthalmic Physiological Optics*, Vol. 5, 1985, pp. 195–204.
26. Decina, L. E., L. Staplin, and A. S. Spiegel. *Correcting Unaware Vision Impaired Drivers*. No. 730009. Pennsylvania Department of Transportation, Harrisburg, Oct. 1990.
27. Rogers, P. N., M. Ratz, and M. K. Janke. *Accident and Conviction Rates of Visually Impaired Heavy-Vehicle Operators—Final Report*. California Department of Motor Vehicles, Sacramento, Jan. 1987, 46 pp.
28. McKnight, A. J., D. Shinar, and B. Hilburn. The Visual and Driving Performance of Monocular and Binocular Heavy-Duty Truck Drivers. *Accident Analysis and Prevention*, Vol. 23, No. 4, 1991, pp. 225–237.
29. State Provincial Licensing Systems—Comparative Data. In *Guidelines for Motor Vehicle Administrators*, NHTSA, U.S. Department of Transportation, and American Association of Motor Vehicle Administrators, 1989.
30. *Medical Conditions Affecting Drivers*. American Medical Association, Chicago, Ill., 1986.
31. Functional Aspects of Driver Impairment: A Guide for State Medical Advisory Boards. In *Guidelines for Motor Vehicle Administrators*, NHTSA, U.S. Department of Transportation, and AAMVA, Oct. 1980.
32. *Recommended Standard Procedures for the Clinical Measurement and Specification of Visual Acuity*. Committee on Vision, National Research Council, Washington, D.C., 1980.

Publication of this paper sponsored by Committee on Visibility.

Entrance Angle Requirements for Retroreflectorized Traffic Signs

MICHAEL S. GRIFFITH, JEFFREY F. PANIATI, AND RICHARD C. HANLEY

The primary objective of this study was to examine the validity of the maximum specification (30 degrees) for entrance angles of retroreflective traffic signs, which is considered to be the widest angle for signs. However, the 45-year-old specification is not substantiated by empirical data. Accurate data are necessary to evaluate the need for a new specification. The amount of light returned from a sign to a driver determines retroreflectivity; therefore, research was conducted from the driver's perspective. Measurements of sign entrance angles were made and their distribution was analyzed. A customized computer software program, SEAMS (Sign Entrance Angle Measurement System), was used to measure entrance angles for over 1,100 in-service traffic signs on several roadway types. After examination of previous research and consideration of other factors, it was decided to take sign entrance angle measurements at 30.5 and 61.0 m (100 and 200 ft). Using the 61.0-m (200-ft) distance for freeways and the 30.5-m (100-ft) distance for nonfreeways provided a conservative estimate of sign entrance angles. The empirical distributions show that approximately 95 percent of the sign entrance angles measured are less than 21 degrees and approximately 99 percent are less than 27 degrees. The study results indicate that the current 30-degree specification covers nearly all signs and provides a margin of safety to compensate for signs that are twisted, bent, or leaning out of plumb. However, the data also show that a lower specification (20 degrees) would cover 99 percent of the freeway signs and 96 percent of all signs measured.

Traffic signs are designed to provide the motorist with the warning, regulation, and guidance necessary to move safely and efficiently through the highway network. To meet this goal, these signs must be clearly visible to the driver both during the day and at night. Nighttime visibility of most traffic signs is provided through the use of retroreflective sheeting. Retroreflection occurs when light rays from an automobile's headlamps strike the surface of a sign and are redirected back toward the driver (see Figure 1). The measure of retroreflectivity is termed the coefficient of retroreflection (R_A).

The amount of light reflected back to the driver varies, depending on two important angles: the entrance angle and the observation angle. The entrance angle is that between a light beam striking the surface of the sign and a line perpendicular to the sign surface [see Figure 2 (*top*)]. There are two components of the entrance angle: β_1 corresponds to the horizontal part of the angle and β_2 corresponds to the vertical part of the angle. The horizontal component of the entrance angle is shown in Figure 2 (*top*). The entrance angle ß may be derived from the expression $\cos \beta = \cos \beta_1 \cos \beta_2$ (*1*). Figure 2 (*bottom*) shows the vertical component of the entrance angle.

The observation angle is that between a light beam striking the surface of the sign and the line of sight of the driver. This angle is a function of the height of the driver's eyes with respect to the vehicle headlamps. Both the entrance angle and the observation angle change as the distance between the vehicle and the sign changes (*2*). This study did not examine observation angles.

Current specifications for minimum R_A values for new sign sheeting are contained in ASTM D 4956-89. These specifications are given for different sign colors at two entrance angles and two observation angles. The entrance angles specified are −4 degrees and +30 degrees. The −4-degree angle is intended for signs that are close to a straight road but turned slightly away from traffic to avoid glare from the smooth sign surface. The +30-degree angle has traditionally been considered to be the widest angle at which signs would commonly be seen on curved roadways. Recently, the basis for the +30-degree entrance angle requirement has been questioned. Investigation into this specification has revealed that it is 45 years old and not substantiated by empirical data.

Presented here are the results of a research study to collect empirical data to evaluate the need for a new maximum specification for sign entrance angles. This study was conducted using a customized computer software program, SEAMS (Sign Entrance Angle Measurement System) (*3*), developed for use with the Connecticut Department of Transportation (ConnDOT) photolog laser videodisc (PLV) retrieval system. This program allowed the measurement of entrance angles for a large sample of in-service traffic signs in an office environment.

APPROACH

The amount of light returned from a sign to a driver determines retroreflectivity; therefore, research was conducted from the driver's perspective. Measurements of sign entrance angles were made and their distribution was analyzed. The implementation of this type of approach required several key components:

1. An efficient method to collect sign entrance angle data for a large group of signs;
2. A definition of the "last-look distance," the distance before the sign after which the driver no longer obtains information from the sign; this is the distance at which the maximum entrance angle would be measured; and

M. S. Griffith and J. F. Paniati, Office of Safety and Traffic Operations R&D, Federal Highway Administration, McLean, Va. 22101. R.C. Hanley, Office of Research and Materials, Connecticut Department of Transportation, Rocky Hill, Conn. 06067.

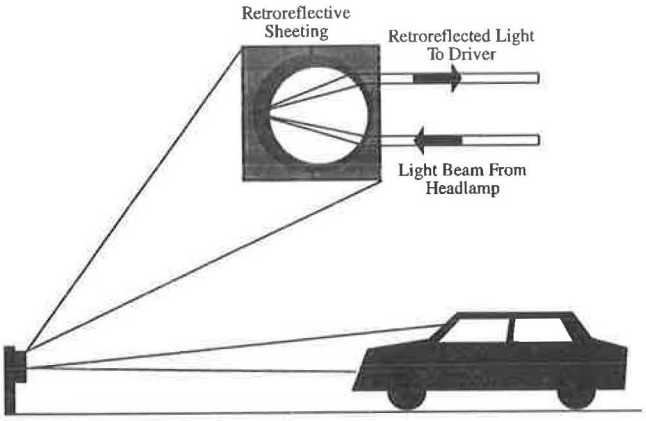


FIGURE 1 Principle of retroreflection.

3. A sampling plan that provides a representative sample accounting for differences in sign classes, sign placements, roadway types, and so forth.

The current literature on the last-look distance and how this research was applied to this study are outlined in the next section, followed by a discussion of the PLV retrieval system and the development of the SEAMS software to allow collection of entrance-angle data in an office environment. Then the sampling plan and data collection and the analysis of the data are presented. Last, the results of the field data collection and a validation analysis of the results are discussed.

LAST-LOOK DISTANCE

Measurement of sign entrance angles requires a distance specification. Entrance angles are a function of the distance between the driver and the sign. On a straight road, the entrance angle of a sign increases as a driver gets closer to a sign. Last-look distance is defined as the distance from the sign to the point at which the driver moves his or her eyes from the sign and does not look at it again (*4*). This is the last distance at which the driver acquires information from the sign. It is not the only point at which a driver looks at a sign. Generally, a driver will look at a sign several times before his or her last look. Figure 3 shows an example for an urban street sign with

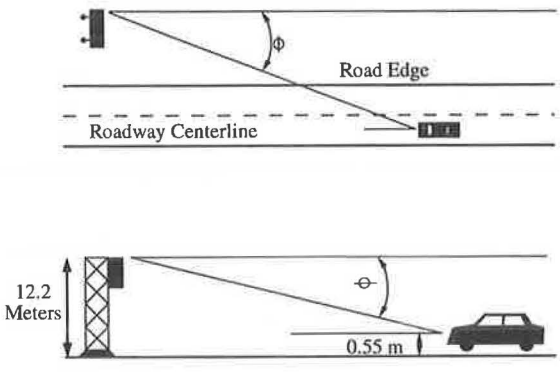


FIGURE 2 *Top*: Entrance angle; *bottom*: vertical component of entrance angle.

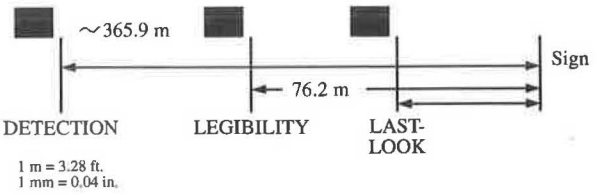


FIGURE 3 Last-look distance.

a 101.6- to 152.4-mm (4- to 6-in.) legend. This sign is detected on average at a distance of 365.9 m (1,200 ft) by the driver, and it becomes legible to the driver on average at 76.2 m (250 ft) (Douglas Mace, unpublished data). Between 76.2 m (250 ft) and a last-look distance of less than 76.2 m (250 ft), the driver may look at the sign several times. Most drivers last look at a nonfreeway sign at a distance of less than 76.2 m (250 ft) (*4,5*).

Study of the current literature on last-look distance was required to determine the distance at which entrance angles would be measured. The objective of the first of two studies by Zwahlen (*4*) was to determine the effectiveness of the STOP AHEAD sign in warning drivers of an upcoming, unexpected, partially concealed STOP sign and intersection during daytime and nighttime conditions. The driving performance and eye-scanning behavior of 39 subjects were studied as they approached an intersection of two-lane rural roads where they were required to stop. The objective of another study (*5*) was to determine the effectiveness of advisory speed signs used in conjunction with curve warning signs in Ohio. A total of 40 drivers were used to drive an unfamiliar test route on a two-lane rural road that included two typical curves equipped with curve warning signs. Eye-scanning data ("first- and last-look distances") were collected for stop signs with and without the STOP AHEAD sign and for curve signs with and without the advisory speed sign. Each study collected this data to identify any differences in driver eye-scanning behavior.

Detailed eye-scanning results for individual subjects and groups for both of the aforementioned studies are given by Zwahlen (*6*). The combined number of last-look distance measurements collected in the studies was 240 under daytime conditions and 141 under nighttime conditions. In both experiments, subjects performed tasks in a group. There was a total of 44 subject groups. The average operating speeds of the subjects ranged from 48.6 km/hr to 89.1 km/hr (30 to 55 mph). Means and standard deviations of last-look distance were computed to find the 99 percent confidence interval for both conditions. The 99 percent confidence interval for the population mean, μ, is (63.4 m, 78.4 m) (208 ft, 257 ft) for daylight conditions and (49.4 m, 61.3 m) (162 ft, 201 ft) for nighttime conditions. Minimum last-look distance results were computed to examine the statistics of the shortest distances for all 44 subject groups. The 99 percent confidence interval for the population mean of minimum last-look distances, μ, is (23.8 m, 33.8 m) (78 ft, 111 ft) for daytime conditions and (27.7 m, 37.8 m) (91 ft, 124 ft) for nighttime conditions.

Last-look distances for signs vary depending on driver characteristics; the function of the sign (signs that require lane changes with merging activity and those that require a complete stop must be detected and read at considerable distances

from the sign); environmental conditions (signs are read at distances very close to the sign under nighttime inclement weather conditions); and the placement of the sign (signs that are further from the shoulder line have longer last-look distances). For example, older drivers as a group exhibit a significant decrease in perceptual, cognitive, and psychomotor abilities, all of which are related to safe driving performance. The U.S. population is aging. By 2030, the number of people older than 65 will more than double (7). The night legibility distances for older drivers are significantly lower than those for younger drivers for all sign types, as shown by the following (1 m = 3.28 ft; 1 mm = 0.04 in.):

		Legibility Distance (m) by Age (years)	
Letter Size (mm)	*Sign Type*	*<40*	*>65*
101.6	Street name	73.2	36.6
152.4	Regulatory	106.7	54.9
203.2	Warning	137.2	73.2
304.8	Guide	213.4	109.8

Placement of signs is generally greater on freeways than on nonfreeways. Vehicle speeds also tend to be greater on freeways, especially in rural areas. Generally, longer last-look distances can be found for signs on freeways because of the combination of the greater placement of the signs and the higher speeds. A longer last-look distance will usually result in a lesser entrance angle.

On the basis of the results of the analysis of the Zwahlen data and consideration of these other factors, it was decided to collect sign entrance angle measurements at 30.5 and 61.0 m (100 and 200 ft). It was believed that using the 61.0-m (200-ft) distance for the freeways and the 30.5-m (100-ft) distance for the nonfreeways would provide a conservative estimate of sign entrance angles.

SIGN ENTRANCE ANGLE MEASUREMENT SYSTEM

ConnDOT, in cooperation with the Federal Highway Administration (FHWA), has developed the software SEAMS for the measurement of entrance angles for in-service highway signs. SEAMS allows measurement of these angles in an office environment using the PLV retrieval system to access highway images stored on laser videodiscs. A photolog is a series of sequential images taken from a moving vehicle at approximately driver's eye level to provide a permanent record of the state-maintained roadway network. ConnDOT uses two automated vehicles to annually film the entire state highway system in both directions on 35-mm color film. A photograph is taken every 0.016 km (0.01 mi) or 16.1 m (52.8 ft).

During the filming, on-board sensors simultaneously collect and store an array of data including route number, direction of travel, cross slope, compass reading, date, time, horizontal and vertical curvature, long-term and short-term roughness, grade, side friction, and vehicle speed. Currently, Connecticut is the only state that collects and stores the photolog images and the corresponding geometric data. The 35-mm film used to collect the images is developed, edited, and recorded onto videotapes, which are then shipped to a videodisc mastering facility where the images are transferred to double-sided laser videodiscs. The final product is a library of 15 videodiscs, each disc side containing 429 km (265 mi) of highway images. The advantages of the videodisc over film are random accessibility, storage density, durability, and the ability of the player to accept computer input.

Measuring entrance angles with SEAMS is fairly simple for the user. The process requires that the selected sign be visible in at least two photolog images, at least two corresponding points on the sign be visible in each image, and the corresponding highway geometric data be available. The user first uses the menu-driven software to move to the image closest to the sign. Then by operating a five-button cursor, the user places four points on the outer edges of a sign. The task of placing points on the sign is repeated one image farther back from the sign, which enables SEAMS to reconstruct the path of the vehicle. SEAMS then calculates the coordinates of these points in relation to the road alignment provided by grade, cross slope, and azimuthal geometric data acquired by the instrumented photolog vehicle. This process is a complex form of parallax. An algorithm within the software computes the entrance angle for a sign at different distances. The distance from the photolog van to the sign is measured along the centerline of the roadway. The output from SEAMS shows entrance angles for Points 1 and 2. Point 1 is at the top edge of the sign and Point 2 is at the edge of the sign furthest from the driver. Taking the conservative approach, Point 2 was selected as the one at which all sign entrance angle measurements would be collected. In the majority of cases, measurements at Point 2 will result in a greater angle than those at Point 1. It is also important to note that the legend of a sign never reaches this outer edge (Point 2). It is the sign legend that contains the information the driver must acquire.

Connecticut's photolog van generally takes all pictures and measurements from the right lane for the nonfreeway system and from the right lane or one of the center lanes for the freeway system. Measurements of signs were taken from the lane in which the photolog van traveled at each location. Clearly, all motorists do not drive in the right lane on the nonfreeway system or in one of the center lanes on the freeway system. Therefore, taking the conservative approach, adjustments were made to all measurements to compute the entrance angle in the lane furthest from the sign except in certain cases. These cases include locations where the same sign is mounted on both sides of the highway (the majority of these cases exist on the freeway system) and for particular signs (e.g., MERGE) where the sign is intended primarily for the motorist in the right lane. For each lane of travel further in distance from a sign, 3 degrees was added to the initial angle for measurements taken at 61.0 m (200 ft) and 7 degrees was added for measurements taken at 30.5 m (100 ft). These adjustments are based on measurements taken in the field.

Using SEAMS has many benefits over conventional taping or surveying techniques. Since measurements are done in an office environment, the need for field work (aside from the collection of the photolog images) is eliminated, thus saving time and field trips and removing personnel from the hazardous highway working environment. The actual measurement operation can be performed by one person, reducing personnel costs associated with a survey crew. Measurements using SEAMS do not impair traffic flow.

Measurements with SEAMS are based on the following assumptions:

1. The photolog van always tracks in the center of the lane, with no erratic maneuvers;
2. The photolog camera is located in the center of the truck;
3. The plane of a sign is always at a right angle to a point on the roadway shoulder line; and
4. The vertical component of the entrance angle is insignificant.

Assumptions 1 and 2 are believed to be reasonable and representative of the conditions when the photologs are obtained in the field. Assumption 3 is valid if signs are installed and maintained in accordance with the *Manual on Uniform Traffic Control Devices* (MUTCD), which states that signs should be mounted approximately at right angles to the direction of, and facing, the traffic that they are intended to serve (*8*). Additional discussion concerning the validity of this assumption in the real world is included later in the paper.

Assumption 4 was made because the horizontal component of the entrance angle dominates the vertical component in its effect on the overall entrance angle, as can be demonstrated through the following example. Figure 2 (*bottom*) illustrates a worst-case scenario for a freeway sign. The driver is approaching a sign 12.2 m (40 ft) high (approximate maximum height of a sign from the roadway surface) and in most cases last looks at the sign 61.0 m (200 ft) or more before it. The distance of 0.55 m (1.8 ft) represents the average height of passenger-car headlamps above the road surface (49 CFR, Section 571.108, Table II, Oct. 1991). The software used in this study (SEAMS) only calculated entrance angles based on the first component. Referring to the bottom part of Figure 2, the second component of the entrance angle is $\tan^{-1}$ 38.2/200 = 10.8 degrees. If the first component is 30 degrees, then entrance angle ß = (cos 30 degrees)(cos 10.8 degrees) = 31.7 degrees. Therefore, in the worst case the entrance angle estimate will be off by less than 2 degrees. On average, it is expected that the "error" will be less than 1 degree.

In the nonfreeway scenario, a driver approaches a sign no greater than approximately 4.6 m (15 ft) high and in most cases last looks at the sign at a distance of 30.5 m (100 ft) or more before it. In this case, the second component of the entrance angle is $\tan^{-1}$ 13.2/100 = 7.5 degrees. If the first component is 20 degrees, then entrance angle β = (cos 20 degrees)(cos 7.5 degrees) = 21.5 degrees. The computations indicate that the first component of the entrance angle (the horizontal part) has much greater influence on the overall entrance angle than the vertical part. Therefore, disregarding the second component (vertical part) of the entrance angle in this study did not have a significant impact on the final results.

SAMPLING PLAN AND DATA COLLECTION

A sampling plan was developed to collect sign entrance angle data with the SEAMS software. The plan was based on Connecticut's system of approximately 12 636 bidirectional km (7,800 mi) of state-maintained highways. Connecticut's system has a range of terrain conditions including hilly and flat, and other topographical features. Entrance angle data were collected for essentially all types of permanent signs including regulatory, warning, and guide signs located on the right side of the roadway. Data were not collected for temporary work zone devices, overhead guide signs, milepost signs, street name signs, and NO PARKING signs.

SEAMS was not validated for construction work zone devices such as drums, cones, and A-frame barricades. Therefore, it was not reasonable to make entrance angle measurements on work zone devices. In addition, the sample of work zone services available on Connecticut's videodisc photolog system is too small to capture the cumulative distribution of entrance angles for these devices. Given the unique characteristics of work zone devices, it may be appropriate to have a separate (possibly higher) specification for these materials.

It was unnecessary to collect and examine the distribution of entrance angles for overhead guide signs. This is because physical limitations such as windshield cutoff and dynamic visual acuity cause an overhead sign to become illegible approximately 53.4 m (175 ft) before the vehicle reaches the sign (*9*). Also, since overhead signs are located directly above the roadway, the effect of the first component of the entrance angle, β_1, is minimal. Therefore, the second component (the vertical part), β_2, dominates the overall entrance angle, β, for overhead signs. The maximum height of an overhead guide sign above the headlamps of a passenger vehicle is assumed to be 11.6 m (38.2 ft) if the sign is 12.2 m (40 ft) high. The entrance angle of this overhead guide sign at a distance of 53.4 m (175 ft) is $\tan^{-1}$ 38.2/175 = 12.3 degrees. This angle does vary depending on the driver characteristics and vehicle type. However, the maximum entrance angle requirement for overhead signs is well below that needed for roadside signs.

Sample size was determined to estimate the number of signs required for data collection. Data had to be collected on a sufficient number of signs to capture a valid cumulative distribution of sign entrance angles. Sample size was estimated by drawing inference on the population mean (population average of sign entrance angles) as the parameter of interest. If the desired accuracy of the sample mean is denoted by d and the test level of significance by α, the formula for sample size (n) is (*10*)

$$n = \frac{(Z_{1-\alpha/2})^2}{d^2} \tag{1}$$

The value of z is a probability extracted from the standard normal probability table. To estimate the population mean to within 15 percent with a probability of .95, the required sample size is

$$n = \frac{(z_{.975})^2}{(.15)^2} = \frac{(1.96)^2}{.0225}$$

$$n = 171 \text{ signs} \tag{2}$$

The data collection goal for each roadway type was 200 signs.

A stratified random sampling plan was developed to collect sign entrance angle data for all roadway types across the entire state of Connecticut. Specific sampling schemes were developed for five roadway types: Interstate, other freeways, principal arterial, other urban (urban arterials and collector roads), and other rural (rural arterials and collector roads). Entrance angle data were collected in 1.62-km (1-mi) samples. The number of miles and the average number of traffic signs per

mile were calculated for each roadway type. Dividing 200 signs by the average number of signs per mile indicated roughly the number of 1-mi samples to collect for each roadway type. The overall sampling plan implemented provides a representative sample accounting for differences in sign types, sign locations, roadway types, and so forth.

A special collection effort was completed for signs situated to the left side of the roadway on freeway facilities. These facilities generally have the most signs situated on the left side of the roadway and were believed to provide a reasonable worst-case scenario for left-mounted signs.

The total number of signs collected is shown below:

Roadway Type	*No. of Signs*
Interstate	212
Other freeways	195
Freeways (left)	192
Principal arterial	182
Other urban	187
Other rural	174
Total	1,142

DATA ANALYSIS

The focus of the analysis was to examine the upper percentiles of the sign entrance angle data. This examination would indicate how the highest entrance angles compare with the current specification (+30 degrees) and provide an indication of the impact of changing this specification. The data were analyzed separately for each roadway type. The sign locations at the 75th, 85th, 90th, 95th, and 99th percentiles for the different roadway types at 30.5 m (100 ft) for nonfreeways and at 61 m (200 ft) for freeways are shown in Tables 1 and 2, respectively.

The average sign location for freeway roads at the 99th percentile is 20 degrees and that for nonfreeway roads is 27 degrees. Table 1 shows for the nonfreeway system that approximately 95 percent of all sign entrance angles are less than 21 degrees and approximately 99 percent are less than 27 degrees. Table 2 shows for the freeway system that approximately 95 percent of all sign entrance angles are less than 16 degrees and approximately 99 percent are less than 20 degrees. The percentage of signs that have entrance angles greater than 20, 25, and 30 degrees for each roadway type is as follows:

Roadway Type	*>20 degrees*	*>25 degrees*	*>30 degrees*
Interstate	0.5	0	0
Other freeways	1	0	0
Interstate and other freeways (left)	0	0	0
Principal arterial	13	3	0
Other urban	5	1	0
Other rural	3	1	0

None of the 1,142 signs measured have entrance angles greater than 30 degrees and only 10 of the signs measured have entrance angles greater than 25 degrees.

Other signs were studied that have the potential of having high entrance angles. In particular, ONE WAY and DO NOT ENTER signs positioned at the end of one-way freeway exit ramps were examined. These signs are needed to prohibit traffic from the cross road that intersects the exit ramp from entering the restricted road section (*8*). Sign entrance angles of eight ONE WAY signs mounted on both sides of the roadway were measured. The angles at 30.5 m (100 ft) range from 10 degrees to 24 degrees. DO NOT ENTER signs on these exit ramps that do not directly face the traffic on the cross road were not measured. The DO NOT ENTER signs that face in this direction are typically supplemented with a ONE WAY sign that faces the traffic on the cross road. The ONE WAY sign displays the required information to the driver on the cross road.

The results show that the effect of using a lower requirement (20 degrees) would not be significant on freeways (0 to 1 percent of the entrance angles are greater than 20 degrees) and moderate on nonfreeways (3 to 13 percent of the entrance angles are greater than 20 degrees) but would not include a margin of safety. The results also show that there is little benefit to be gained from using a higher maximum entrance angle.

TABLE 1 Percentiles of Entrance Angle Measurements at 30.5 m

Roadway Type	75th	85th	90th	95th	99th
Principal Arterial	17°	19°	22°	24°	29°
Other Urban	14°	16°	16°	20°	26°
Other Rural	12°	13°	15°	19°	26°
Average	14°	16°	18°	21°	27°

1 m = 3.28 ft

TABLE 2 Percentiles of Entrance Angle Measurements at 61 m

Roadway Type	75th	85th	90th	95th	99th
Interstate	12°	14°	14°	16°	19°
Other Freeways	12°	14°	15°	18°	24°
Interstate and Other Freeways (Left)	11°	12°	13°	14°	17°
Average	12°	13°	14°	16°	20°

1 m = 3.28 ft

FIELD DATA COLLECTION AND VALIDATION ANALYSIS

Entrance angle data were collected in the field to verify the accuracy of SEAMS. It was determined that measurements on 75 signs would be sufficient to do a valid statistical analysis. Signs were selected for all roadway types in a preferred area around Rocky Hill, Connecticut, to minimize travel time. (ConnDot's Office of Research and Materials is located in Rocky Hill.) Before data collection, color prints of each sign were produced with the color video printer, a component of the ConnDot photolog laser videodisc system. This allowed for easy identification of the selected signs to be measured in the field. Measurements were attempted on over 100 signs by a survey crew. For various reasons, such as roadway safety, new sign replacement, and sign elimination, data were collected on only 77 signs at 30.5 and 45.7 m (100 and 150 ft).

Entrance angle data were collected with a device designed by an engineer at FHWA. The entrance angle instrument has a telescope that is mounted to an aluminum base with a level and computer-generated protractor attached to the top of the base. A handle is attached to the bottom of the base, which is held when using the instrument.

The measurement process used in the field to collect entrance angle data was fairly simple. First, distances of 30.5 and 45.7 m (100 and 150 ft) from the selected sign were measured with a measuring wheel and marked. Cones were then placed at the lane line and shoulder line at the marked distances. After the required distances from a selected sign were measured and marked, the lane width was measured with a tape. Angle measurements were taken with the entrance angle device by first supporting it level against a sign. The telescope was then turned until the target (a cone) was viewed at the intersection of the crosshairs. At this point, an angle was read from the protractor. The farther the telescope was turned, the greater was the sign entrance angle. This measurement process was completed for all four cones. Through interpolation between the two angles obtained at the lane line and shoulder line, the entrance angle of each sign was calculated at the point 1.4 m (4.5 ft) from the lane line. Assuming that the motorist drives in the middle of a 3.7-m (12-ft) lane and the width of the vehicle is approximately 1.8 m (6 ft), 1.4 m (4.5 ft) to the right of the lane line is the average position where the driver is located. This is the position on the roadway where a motorist views signs at different entrance angles.

Figures 4 and 5 show the cumulative distribution of the SEAMS data and field data for all roadway classifications (Interstate, principal arterial, etc.) combined at 30.5 m (100 ft) and 61 m (200 ft), respectively. The distance measured in the field for the freeways was 45.7 m. This is because at the time of field data collection it was believed that 45.7 m was the most reasonable distance to represent the last-look distance for freeways. After further consideration of all the factors that affect the last-look distance (driver characteristics, function of the sign, environmental conditions, placement of the sign, etc.), it was decided that a distance of 61 m (200 ft) is more appropriate. Both Figures 4 and 5 show that the distribution of the SEAMS data is more conservative than the distribution of the field data. The entrance angles from SEAMS are greater from the lowest percentile to approximately the 80th percentile. In the upper range, 80th percentile and above,

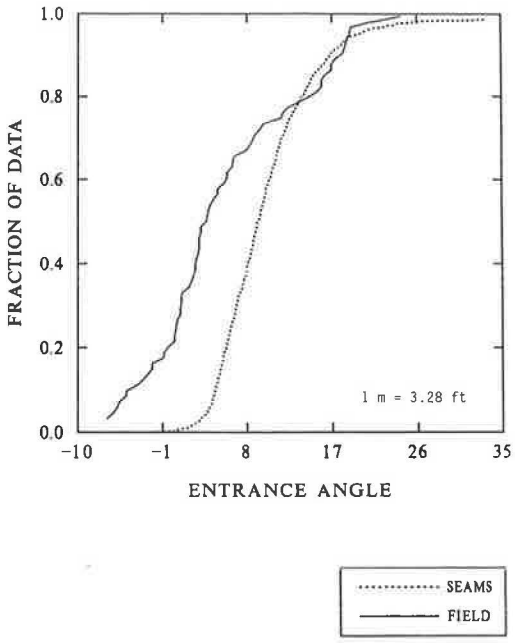


FIGURE 4 SEAMS versus field data at 30.5 m.

the SEAMS and field distributions parallel one another. Since this is the range of interest for the maximum entrance angle requirement, the SEAMS data reasonably represent the greatest entrance angles.

The assumption of SEAMS that the plane of a sign is at a right angle to a point on the roadway shoulder line was examined in the field. Normally, signs should be mounted approximately at right angles to the direction of, and facing, the traffic that they are intended to serve. They should be turned slightly away from the road to avoid the specular reflection (in which drivers would see their headlights by mirror reflec-

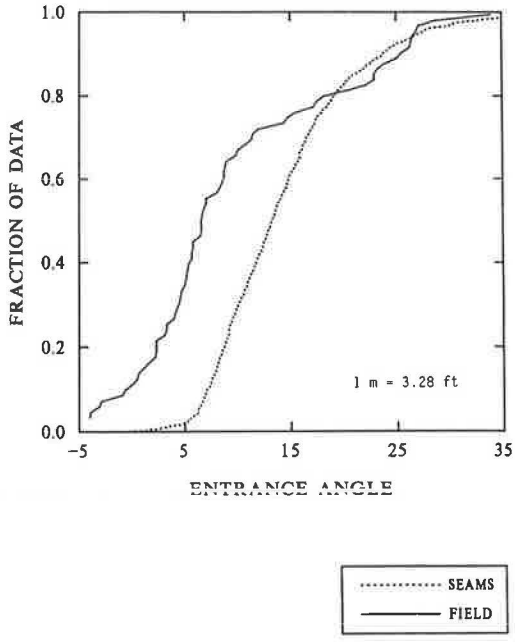


FIGURE 5 SEAMS versus field data at 61 m.

tion on the front surface of the sign sheeting) (*8*). An estimate of the degree of skewness from the perpendicular was made for each sign in the field. The SEAMS assumption that signs are at a right angle was found to be incorrect for a significant number of signs in Connecticut. Twenty-three signs were estimated to be skewed greater than ±10 degrees to the perpendicular of the roadway. Table 3 shows a comparison of the data at 30.5 m (100 ft) and 61 m (200 ft) between all signs and all signs estimated in the field to be skewed 10 degrees or greater to the perpendicular of the roadway. The majority of the differences between the field and SEAMS data are 5 degrees or less. The largest discrepancies exist between the signs that were measured in the field to be skewed greater than ±10 degrees to the perpendicular of the roadway. This signifies that when the assumptions of SEAMS are satisfied, one can have greater confidence that the results from SEAMS truly represent actual entrance angles that can be found on the nation's highways.

In addition to verifying SEAMS in the field, the repeatability of the SEAMS data was studied. Entrance angle data for 1989 and 1990 were compared to see if measurements can be reproduced over time. Connecticut collects new pictures and highway geometric data each year. Fifteen signs were examined, and discrepancies were found to be 2 degrees or less.

CONCLUSION

The study reported here obtained empirical data that can be used to establish a maximum specification for entrance angles of retroreflectorized traffic signs. Data were collected for a wide range of urban and rural conditions. SEAMS provided a quick and easy method to collect entrance angles on over 1,100 signs. The study results indicate that the current maximum entrance angle requirement, 30 degrees, includes a margin of safety to compensate for signs that are twisted, poorly placed to alignment, bent, or leaning out of plumb. As a general entrance angle specification, 30 degrees is valid. However, the data indicate that a lower specification, 20 degrees, could be used for signing on freeways with no adverse effect. Only 3 freeway signs of the 599 freeway signs measured (0.5 percent) have entrance angles greater than 20 degrees. There are also cases on nonfreeways where a 30-degree requirement is unwarranted. It is important that a jurisdiction examine the signing on their roadways to determine the potential for very high entrance angles before deciding on what maximum specification to use.

The research results from SEAMS are conservative and reasonable for the following reasons:

1. The results are based on a large sample of signs (1,142);
2. The sample selected is representative of the nation, accounting for differences in sign classes, sign placements, roadway types, and so on;
3. All measurements were calculated at the average minimum last-look distance;
4. All measurements were calculated at the point on the sign furthest from the roadway;
5. All measurements were calculated from the lane furthest from the sign except in special cases; and
6. Comparison of the SEAMS and field distributions shows that SEAMS parallels the upper range of the field data, which is the range of interest in this study.

It is believed that the data collected in Connecticut reasonably represent that which can be found in other states. Highway design and geometric characteristics of freeways across the nation are relatively standard. Although the nonfreeway system is not as standard as the freeway system, the results are based on measurements taken at distances very close to the sign and therefore are believed to be representative of the conditions found elsewhere. In addition, since a conservative approach was taken in the measurement of the entrance angle, a factor of safety exists to account for greater sign entrance angles that might exist on other states' nonfreeways.

Although it is believed that the results reported here are representative of the overall conditions encountered by the driver, it is recognized that there are cases where sign entrance angles are over 30 degrees and in special situations they are significantly above 30 degrees. In the authors' opinion, it is not a prudent approach to expect the specification to cover all of these cases. It is incumbent upon the engineer to find other site-specific solutions, such as the installation of supplementary signing.

Although SEAMS was developed for use in the measurement of sign entrance angles, numerous other applications of the technology are envisioned. Currently, two efforts are under way. A videodisc-based sign inventory is being developed to allow users to relate signs to photolog images and use a mod-

TABLE 3 Differences Between the SEAMS and Field Data

		Difference			
Distance	Signs	0°-5°	5°-10°	10°-15°	15° - +
30.5 meters	All signs	48	18	8	3
	All signs > 10° to perpendicular of roadway	8	6	6	3
61 meters	All signs	50	19	4	4
	All signs > 10° to perpendicular of roadway	10	7	3	3

1 m = 3.28 ft

ified version of the SEAMS program to measure sign sizes directly from videodisc. A second effort is generalizing the SEAMS concept for use in measuring heights, offsets, and longitudinal distances. This generalized measurement system has the potential for measuring sign sight distance, passing sight distance, vertical clearances, roadside hazard locations, and many parameters.

ACKNOWLEDGMENTS

The authors would like to acknowledge Willard J. Kemper, FHWA, for his significant assistance in collecting entrance angle data for this study. They also express special thanks to John Arens, FHWA, for designing the field entrance angle device. Their gratitude is also extended to Peter Van Dyne, who developed the mathematical principles employed in SEAMS, and David Burns, who developed the SEAMS software programs.

REFERENCES

1. *Guide to the Properties and Uses of Retroreflectors at Night.* Publication 72. International Commission on Illumination (CIE), Vienna, Austria, 1987.
2. McGee, Hugh W., and D. L. Mace. *Retroreflectivity of Roadway Signs for Adequate Visibility: A Guide.* Report FHWA/DF-88/001. Federal Highway Administration, U.S. Department of Transportation, 1987.
3. *Sign Entrance Angle Measurement System (SEAMS).* Connecticut Department of Transportation, 1991.
4. Zwahlen, H. T. Stop Ahead and Stop Signs and Their Effect on Driver Eye Scanning and Driving Performance. In *Transportation Research Record 1168*, TRB, National Research Council, Washington, D.C., 1988, pp. 16–24.
5. Zwahlen, H. T. Advisory Speed Signs and Curve Signs and Their Effect on Driver Eye Scanning and Driving Performance. In *Transportation Research Record 1111*, TRB, National Research Council, Washington, D.C., 1987, pp. 110–120.
6. Zwahlen, H. T. *Warning Signs and Advisory Speed Signs—Re-evaluation of Practice.* Report FHWA/OH-84/003. Ohio Department of Transportation, Columbus, June 1983, 185 pp.
7. Paniati, J. F., and R. N. Schwab. Research on the End of Life for Retroreflective Materials: A Progress Report. In *Transportation Research Record 1316*, TRB, National Research Council, Washington, D.C., 1991, pp. 13–17.
8. *Manual on Uniform Traffic Control Devices for Streets and Highways.* Federal Highway Administration, U.S. Department of Transportation, 1988.
9. Stein, A. C., Z. Parseghian, R. W. Allen, and C. E. Wolf. *Overhead Guide Sign Visibility Factors.* Volume I: Final Report. Report FHWA-RD-88–196. Federal Highway Administration, U.S. Department of Transportation, 1989.
10. Desu, M. M., and D. Raghavarao. *Sample Size Methodology.* Academic Press, Inc., San Diego, 1990.

Publication of this paper sponsored by Committee on Visibility

Exact Road Geometry Output Program for Retroreflective Road Sign Performance

KENNETH D. UDING

The angles used for the laboratory testing of retroreflective sign sheeting that are set forth in specifications are well defined and are well understood in the laboratory test setting. Not well known is exactly what values of these angular parameters, especially observation angle, occur for actual signs on the roadway. The mathematics for a complete vector structure incorporating the location data for headlamps, driver's eye, the sign, and the vehicle-to-sign (road) distance has been set up. These inputs define all locations exactly; there are no assumptions. The mathematics has been incorporated into a computer program, ERGO (Exact Road Geometry Output), which computes the exact angles at which the sign is actually seen by approaching drivers. The observation angle for actual signs is shown to be two separate values—one for each headlamp—and not simply the eye height over the headlamps. Eye setback is also shown to be a critical factor under some conditions. The ERGO data demonstrate that observation angle is a direct function of road distance: as road distance becomes less, observation angle becomes greater. Specific observation angles correspond to specific road distances. Graphs of the observation and entrance angle correlates are given for STOP and near-roadside signs, overhead guide signs, and signs at a large offset from the road edge. The effect of different size vehicles on observation angle is shown. The relationship of time to observation angle is demonstrated: as the approaching driver observes a sign at observation angles greater than 0.5 degree, these angles are traversed in fractions of a second. Both the ERGO program and its mathematical basis will be made available to others so that it can be applied to other experimental and theoretical data to better correlate laboratory test values and actual road performance of retroreflective material.

Road signs at night require some minimum level of luminance in order to be seen effectively and in time by an approaching driver who is dependent on reading such signs for certain essential information. Retroreflective sheeting is used on road traffic signs as a means to provide this luminance in the absence of internal or external illumination. Specifications for minimum reflectivity values in laboratory tests of this sheeting have the ultimate objective of providing the required level of effective performance (i.e., luminance) for such signs at the distances and for the time that the driver requires. The efficiency of retroreflective sheeting is specified by setting reflective efficiency (coefficient of retroreflection) values (R_A) for certain laboratory test points determined by designating values for those angular parameters that have been carefully defined for laboratory tests (*1*; ASTM Standard E808-91; AASHTO Standard Method of Test T257-86). The angular parameters that are recognized as the primary determinants of reflective efficiency and are set forth in every specification are observation angle and entrance angle. These angular parameters must be carefully determined and accurately set up in the laboratory to ensure accurate and valid measurements of the reflectivity (as well as to achieve correlation between laboratories). An R_A-value associated with a certain reflective sheeting has meaning only in the context of the *pair* of these angular values at which it is measured: one exact observation angle and one exact entrance angle. (The "one entrance angle" may be defined by the two angular coordinates B_1 and B_2, which, taken together, define a single entrance angle condition.) These angles are well defined and well understood for laboratory test purposes. What is not well understood is what the actual observation and entrance angles are that occur on the roadway in specific sign situations. How are these angles determined for actual road signs seen by the driver of a vehicle? How do these angles change as the vehicle approaches the sign? What angular values are important to a driver approaching a particular sign? How are these values different for drivers of different types of vehicles?

In a recently published National Cooperative Highway Research Program (NCHRP) Report (*2*), it was pointed out that the federal standard (then FP-85) does provide minimum specific intensity per unit area (SIA) (R_A) standards for new material. These standards, however, were developed by sheeting manufacturers as purchase specifications, not based on drivers' needs. Therefore, in order to set a new standard for sheeting based on the driver's actual minimum visibility requirements, the report stated:

> The FHWA project on 'Minimum Visibility Requirements for Traffic Control Devices,' is to determine the minimum visibility distances for signs and markings. Based on these minimum visibility requirements, it will be possible to determine the retroreflectivity necessary to make a sign or marking visible at a given distance.

A parallel effort is under way by the European Committtee for Coordination of Standards CEN to develop a standard based on the luminance requirements for reading actual road signs. This effort is equally dependent on accurate values for the observation and entrance angles at which those signs are actually seen by drivers if it is to arrive at specification values that correlate with the actual performance as planned.

A method is needed that can readily provide accurate observation and entrance angle correlates for given sign situations and at different viewing distances. These measurement parameters can then easily be included in study data and considered in arriving at conclusions relating reflectivity levels and other variables. Although some researchers may be computing these values accurately, the basis of such values is never certain and almost never is it adequately described.

Stimsonite Corporation, 7542 N. Natchez, Niles, Ill. 60648.

Thus computational error or the dimensional basis cannot be determined or verified. A search of the literature reveals no data on any such method generally available to accurately determine these angular values. There does not appear to be any detailed compilation of these values for actual road sign locations and viewing distances.

Consequently, the computer program ERGO (Exact Road Geometry Output) has been developed to determine the exact observation and entrance angle correlates that actually occur for signs and traffic control devices at different locations and distances, as well as values for several other angular parameters of retroreflectivity. ERGO is available at no charge to qualified personnel interested in the use or study of retroreflective sheeting. Its structure and formulas are available for analysis and proof. It can be used as a common reference for other experimental data.

The data obtained with ERGO are discussed in this paper. Summaries in the form of graphs of the observation and entrance angle correlates for some typical signs are shown. These "applications" of ERGO data illustrate how the data can be useful in studying how the angular parameters change in real road sign situations. In turn, this may be essential in the determination of valid measurement values for minimum levels of retroreflectivity. Given a sufficient range of retroreflectivity data, the program may be used to accurately compare the efficiency of different retroreflective materials for particular applications.

ERGO OPERATION

To calculate the exact road geometry, a simple mathematical vector structure was created together with the formulas to compute all angles precisely and accurately. The computer program was written to accommodate all dimensions in the three coordinate planes that prescribe the locations of the different elements and, in turn, the retroreflective geometry of a specific road sign situation. The program then computes the defined angles determined by those inputs. (The complete mathematical analysis of the vector structure and the derivations of the defined angles by D. Couzin are available on request from the author.)

ERGO easily determines exact, not approximate, retroreflective geometry for any set of input data desired. Each set of correlates generated by this program is specific to the one corresponding set of input dimensions that exactly locate the eye, headlamps, and sign relative to each other. The values are absolute; the only subjectivity is in the selection of the input dimensions, or range of dimensions, to represent any generic designation such as typical STOP sign, large-offset guide sign, standard car, and so forth.

The parameters of location for car headlamps, driver's eye, and sign location that are inputs for ERGO are as follows (all input dimensions are entered in meters or, in an alternative menu choice, a version is available in which all inputs are in feet):

ROAD DISTANCE to Sign: _____
SIGN: Offset from Road Edge: _____
Height above Roadway: _____

VEHICLE DIMENSIONS (STANDARD CAR):	*Meters*	*Feet*
Separation between headlamps	1.042	3.42
Headlamp height over roadway	.661	2.17
Eye height over headlamps	.466	1.53
Eye setback behind headlamps	2.057	6.75
Eye displacement left of vehicle centerline	.330	1.08

Dimensions were measured for a wide variety of vehicles in many models. The dimensions above are the mean of a relatively narrow range of data for compact and mid-size cars and thus well represent the universe of such cars. These mean values have been dubbed the "standard car" and are included as default values in ERGO. Of course, any values can be entered to override the defaults. Separate dimensions were determined for such vehicle groups as large cars, small vans, large vans [recreational vehicles (RVs)], and large trucks with maximum eye-headlamp displacements, dubbed "MAX trucks."

OBSERVATION ANGLE: LABORATORY AND ROAD DEFINITIONS

An accurate understanding of the observation angle is critical both to understanding the effects of the different inputs on the geometry and to using the output data correctly. For laboratory test purposes, observation angle can be defined as follows:

> The angle that is formed at the reference center on the test sample between a line to the light source (the illumination axis) and a line to the receiver (observation axis) (see Figure 1a).

It is useful to observe that its measure is a function of the displacement distance from the light source to the receptor (measured perpendicular to the illumination axis) and the distance measured along the illumination axis. This is as defined (in slightly different terminology) in ASTM E808-91.

In the actual road situation where the driver of an approaching vehicle observes a road sign illuminated by the car's headlamps, determining the observation angle at which the sign is seen is a bit more complicated. In the literature, even as recently as the reports by Black et al. (*2*) and by McGee and Mace (*3*), the observation angle for the driver has consistently been presented as if it were simply the vertical displacement of the eye above the level of the headlamps. However, this is not correct.

Note that in Figure 2a (which diagrams the eye-headlamp relationships for the standard car), the vertical distance down from the eye is a dimension from the eye to nothing.

In truth, the observation angle, whether it is measured in the laboratory or on the road, is the angle intercepted (at the particular road or test distance) by the straight-line displacement of the receptor (or eye) from the light source illuminating the reflector sample or sign. In the road situation, the receptor is the human eye; the two eyes are sufficiently close together that they can be considered one point at the centerline of the driver. The light source is a headlamp.

Unfortunately for the cause of simplicity, (*a*) there are two headlamps, that is, two separate light sources illuminating signs; (*b*) the headlamps are unequally displaced on either

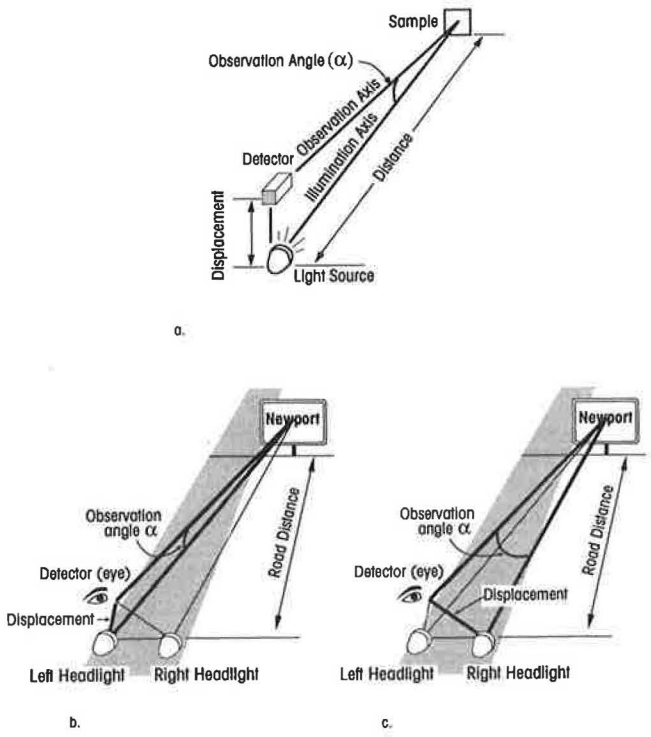
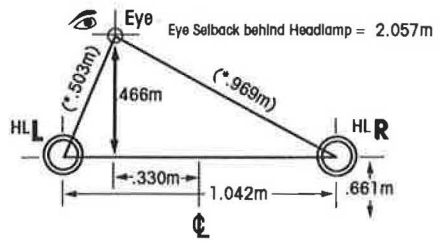


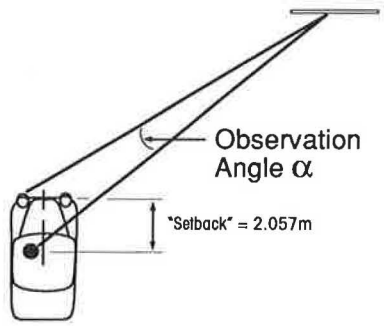
FIGURE 1 Observation angle: (*a*) laboratory test setup per specifications; (*b*) road to left headlamp; (*c*) road to right headlamp.



a.

*Not an ERGO input value but shown for the "Standard Car."

(1 meter = 3.28 ft.)



b.

FIGURE 2 Vehicle dimensions in ERGO (primarily affecting observation angle): (*a*) standard car dimensions; (*b*) setback dimension and effect on observation angle.

side of the driver; and (*c*) the headlamps are at a substantial separation. This means that in order to be accurate and mathematically specific, two different observation angles are in effect for each road sign situation. One is the angle intercepted by the displacement to the left (driver's side) headlamp and the other is the angle intercepted by the displacement to the right headlamp. [The necessity of computing the angles to each headlamp separately has also been recognized by Johnson (*4*).]

Figure 1b (to left headlamp) and Figure 1c (to right headlamp) illustrate how these two observation angles are formed between the respective headlamp, the eye, and the sign. Note the correspondence of the elements to the laboratory setup (Figure 1a).

Another factor, which appears to have often been overlooked, can also substantially complicate observation angle calculations under certain conditions. It is automatically included in ERGO's complete computation. This involves the setback distance. In all vehicles the eye is set back behind the headlamps by a substantial distance; the standard car setback averages 2.06 m (6.75 ft). For a driver viewing signs at very small offsets (off the road) and at substantial distance, the setback occurs more or less parallel to the observation axis and thus does not enter into the determination of the observation angle.

However, as a vehicle approaches close to a sign that is substantially offset to one side or the other, or as it turns away from the sign even slightly, as on a curved approach, the setback becomes an increasingly significant component of the eye-headlamp displacement that produces significantly larger observation angles. The rate of increase accelerates at very close distances, producing extremely large observation angles, sometimes even completely out of the effective range of retroreflected light. The principle is illustrated in Figure 2b.

APPLYING ERGO DATA AND DATA SUMMARY GRAPHS

Analysis of data from ERGO can yield substantial information on how retroreflective signs are actually seen. It can be used to determine which observation and entrance angles should be specified for any particular application. As a part of studies of the parameters of effectiveness of signs—legibility distance, detection distance, and so on—it can be useful in indicating where effective retroreflectivity is required, namely, at the observation angle that corresponds to some determined distance having an important function.

The few graphs summarizing ERGO-derived data shown here illustrate how some useful principles can be deduced. Following are a few guidelines for interpreting the ERGO data presented here.

First, plots of angular data can apply to many different sizes and types of signs if the signs are mounted at offsets and heights similar to those shown. However, the point at which a particular sign is usefully seen or detected may not start or end at the endpoints of the plot shown on a given graph. The beginning and end of the useful viewing time for a sign must be independently determined and then the angles corresponding to those distances should be noted.

In using ERGO to evaluate the performance of a given material for a particular type of sign, certain criteria should be separately evaluated: (*a*) the distance at which a particular sign needs to be detected, (*b*) the distance at which it should be read, (*c*) the span of distance during which it continues to be usefully read (and thus also the reading "time" based on a given approach speed), and (*d*) the distance between the approaching driver and the sign when the driver can no longer be expected to read the sign or no longer needs the information. Then, using ERGO, the observation and entrance angle correlates should be noted for those various determined distances.

In the opinion of this author, a determination should also be made in using the above series of data points, although its evaluation is not strictly a part of geometry-limited ERGO. It applies to reflectivity values (and luminance data, if available) at the "far" and "near" limits, especially in the selection between materials of different characteristics. This determination is whether the sign luminance at the far distance limit is less than, more than, or equally as important as that at the near distance limit (at which the sign viewing is actually terminated). Note that the near distance limit is reached after the sign has been observed and read during the entire span of time after the initial reading until reaching the near limit.

ERGO DATA OUTPUTS

The ERGO program output provides values for all the defined angles of retroreflectivity that the input values determine. An example of the actual output for a single road sign situation as reported by ERGO is shown below:

	Left Headlight	*Right Headlight*
Alpha (α)	0.33	0.52
Beta (β)	2.97	2.42
Gamma (γ)	39.29	121.09
Epsilon (ε)	29.98	−56.89
Omega (ω)	69.24	64.23
Beta 1	2.30	−1.25
Beta 2	1.88	2.07
Beta V	1.05	1.05
Beta H	2.78	2.18

In this example, the inputs defined a point on a sign that is offset 2.5 m (8.2 ft) from the edge of the road at a height of 2.5 m (8.2 ft) viewed from the standard car at a road distance of 100 m (328 ft). Output values of ERGO are given for the geometric parameters, which are defined for the laboratory test setup by ASTM E808-91. Two other parameters have been created for the road situation only. The geometric parameters output by ERGO are as follows:

Alpha: Observation angle
Beta: Entrance angle
Gamma: Presentation angle
Epsilon: Rotation angle (ASTM E808-91)
Omega: Orientation angle (ASTM E808-91)
Beta 1: Entrance angle component as defined
Beta 2: Entrance angle component as defined
Apply to ROAD environment ONLY: (not ASTM, not prior CIE)

Beta V: Entrance angle vertical component
Beta H: Entrance angle horizontal component

The analyses in this paper are principally concerned with the observation and entrance angle values, but other values can be important to laboratory tests or specifications intended to correlate with actual road performance, or to both.

The ERGO data presented in the balance of this paper consist of the observation and entrance angle outputs for particular signs over a range of approach distances. These data present the range of these angles that actually occurs for given signs as seen by drivers of approaching vehicles. The data are presented in graphs that plot observation angle (on the vertical scale) against road distance (on the horizontal scale). Since road distance is linear, it is very important to note that the plot is also one of time, given a specific vehicle approach speed. Since the driver's information and decisions are primarily defined by time, this type of graph best represents the rate of change in observation angle as the observing vehicle approaches the sign. Shown below the graphed data is the time in seconds before the vehicle passes the sign at various distances and for speeds of 50 kph (30.1 mph), 75 kph (46.6 mph), and 100 kph (62.1 mph).

Of course, there is an entrance angle correlate for every observation angle in the data. In the lower right-hand corner of most of the graphs is a separate plot of entrance angle against the common horizontal road distance scale. The scale for increments of entrance angle is the short vertical scale along the lower right side of the graph. To avoid confusion with the observation angle plots, the entrance angle output is only plotted when it exceeds 5 degrees. This is acceptable because any entrance angle of 5 degrees or less is considered equivalent to zero degrees. In fact, sheeting is actually tested at 4 or 5 degrees to avoid front surface reflection. The entrance angle plots are short because entrance angle does not exceed 5 degrees until the approaching vehicle is very close to the sign.

An alternative type of graph would plot observation angle against entrance angle. Thus, that plot is the compilation of specific sets of correlates of observation and entrance angles. The correlates of any specification can also be shown as specific points. This type of graph provides the best comparison of specification test points with the actual geometry, especially for unusual or extreme situations.

Road distance can be marked on the actual plots on this second type of graph but it is very nonlinear, with increments of distance very compressed for the longest distances and then increasing to longer and longer spans as the vehicle approaches the sign.

ERGO DATA GRAPHS

Several graphs summarizing ERGO-produced data are shown. The examples were selected to demonstrate a variety of circumstances in which the data from this program can be useful. Analysis of the summarized and plotted data for particular signs can reveal important relationships between the angular

variables (primarily observation angle) and other variables (primarily road distance).

STOP Signs and All Near-Roadside Signs

STOP signs have very specifically defined locations, one component of which is very small offsets from the road edge. In Figure 3, the solid-line plots represent a minimal urban offset of 0.6 m (2 ft). The dashed-line plots represent a large rural offset of 3.65 m (12 ft). The majority of right-edge roadside signs are located within this range.

The plots, virtually identical for either offset, demonstrate that this difference in offsets has essentially no effect on the observation angle value. The data also show that the mean entrance angle is only 13 degrees, even for the larger offset when the approach distance is only 25 m (82 ft) and the mean observation angle is 2.2 degrees.

Zwahlen (*5*) has shown that drivers approaching a STOP sign looked away from their final viewing ("last look") of the STOP sign at a mean distance of 47.6 m (156 ft) for the worst condition studied and this was after they had viewed the sign for 148 m (484 ft); presumably they no longer needed the information. Thus the effective distances for STOP signs involve observation angles of 1 to 0.5 degrees and entrance angles of 10 degrees or less.

Overhead Sign

Figure 4 is the plot for an overhead sign centered over the driver's lane at a height of 7 m (23 ft). Overhead signs are mounted at a very limited range of locations relative to the driving lane, which requires them to be considered a significant and separate category. They are always designed to be read at significant distances, and their fixed position above observing vehicles and perpendicular to the vehicles' direction determines that the observation and entrance angles differ very little from site to site. Assuming initial detection in the range of 200 to 300 m (656 to 984 ft) for a very significant span of time after initial detection and viewing, the observation angle in effect for the approaching driver is very small; the entrance angle is negligible. More than any other type, this sign functions during its useful viewing distance at very

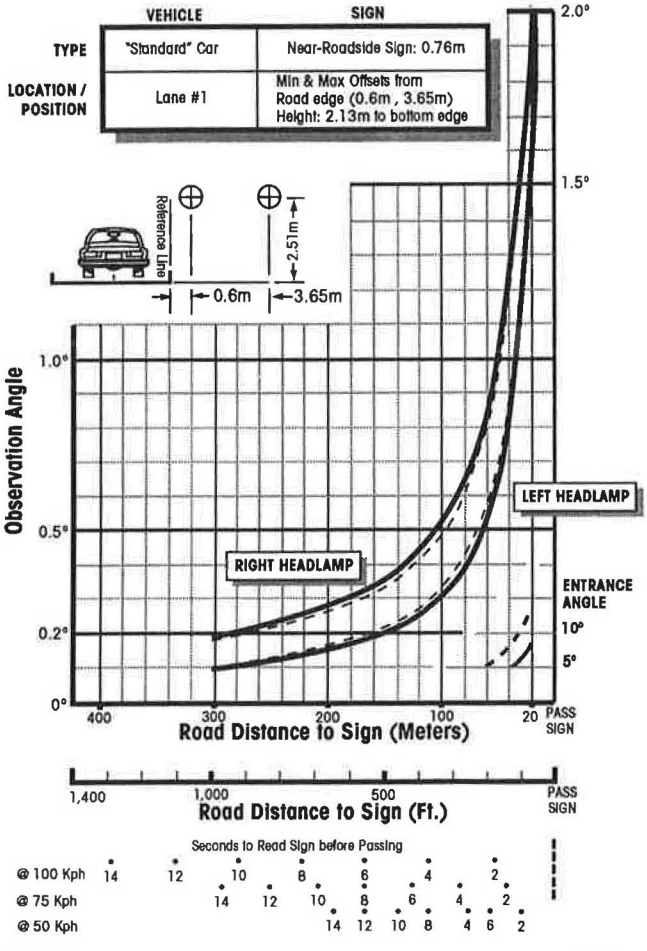


FIGURE 3 Observation angle versus road distance for STOP signs and all near-roadside signs, regulatory and other. Entrance angle shown where 5 degrees or more (to lower right vertical scale).

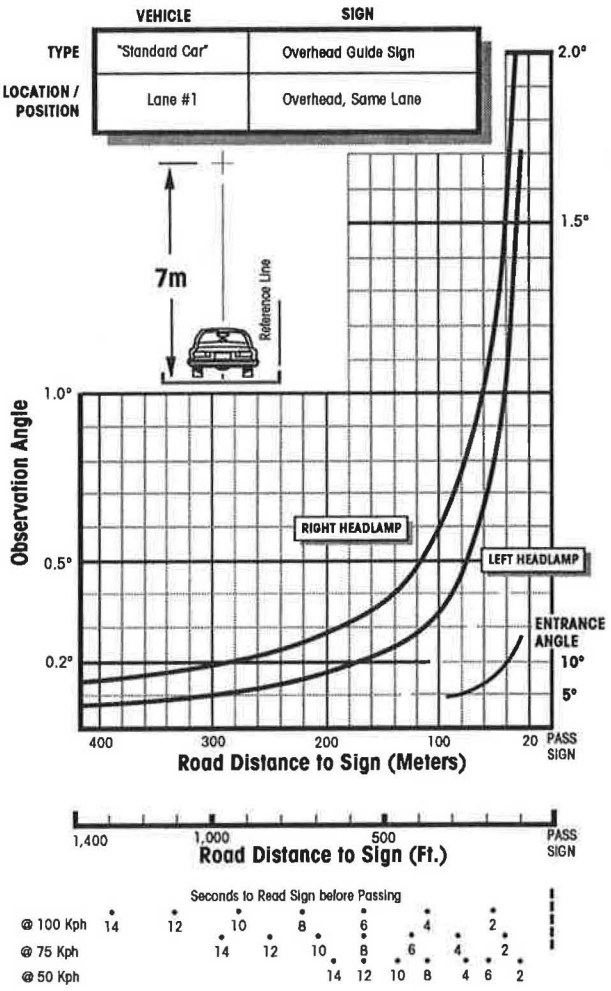


FIGURE 4 Observation angle versus road distance for overhead guide signs. Entrance angle shown where 5 degrees or more (to lower right vertical scale).

small observation and entrance angles. As close as 30 m (98 ft), where the observation angles are about 1.5 degrees and the driver has probably terminated viewing of the sign, the entrance angle is still only 13 degrees.

Large-Offset Signs

Figure 5 shows the observation and entrance angle correlates both to the center of a very wide sign [5.5 m (18 ft)] and at a very large offset from the road edge [9.1 m (30 ft)]. Thus the entrance angle computations are to a point offset 14.6 m (48 ft) from the road edge at a height of 3.34 m (11 ft). Note that for signs at such a large offset, the observation angle curves for left and right headlamps cross over as the line of sight passes over the right headlamp when the vehicle is close to the sign.

Comparison of Vehicles with Different Eye-Headlamp Displacements

Figure 6 is a comparison of the differing observation angles at which drivers of certain types of vehicles with different eye-headlamp displacements view typical road signs at successive approach distances. The vehicles represented are as follows:

1. "Standard car" represents the mean dimensions of the compact, mid-size, and medium cars. The majority of cars fit well into this category.
2. "Large car" represents the mean dimensions of cars such as the Lincoln Town Car, the Chevrolet Caprice (mid-1980s style), and similar cars. Cars in this category are rapidly disappearing.
3. Large vans (RVs).
4. "MAX truck" represents the approximate dimensions of the largest truck-tractors with a maximum eye height of about 2 m (6.6 ft) above the headlamps.

Actually, of course, there is a continous range of trucks having various eye-headlamp displacements so as to create a continuum of plots of observation angle from that shown for large vans to that shown for the MAX truck.

[Note: To avoid having to present excessive data on one graph, only the observation angle data for the left headlamp are shown. The sign position used for the ERGO computations is offset from the road edge by 6.08 m (20 ft) and at a height of 2.13 m (7 ft). Changes in vehicle parameters have no effect on the entrance angle; therefore no entrance angle data are shown.]

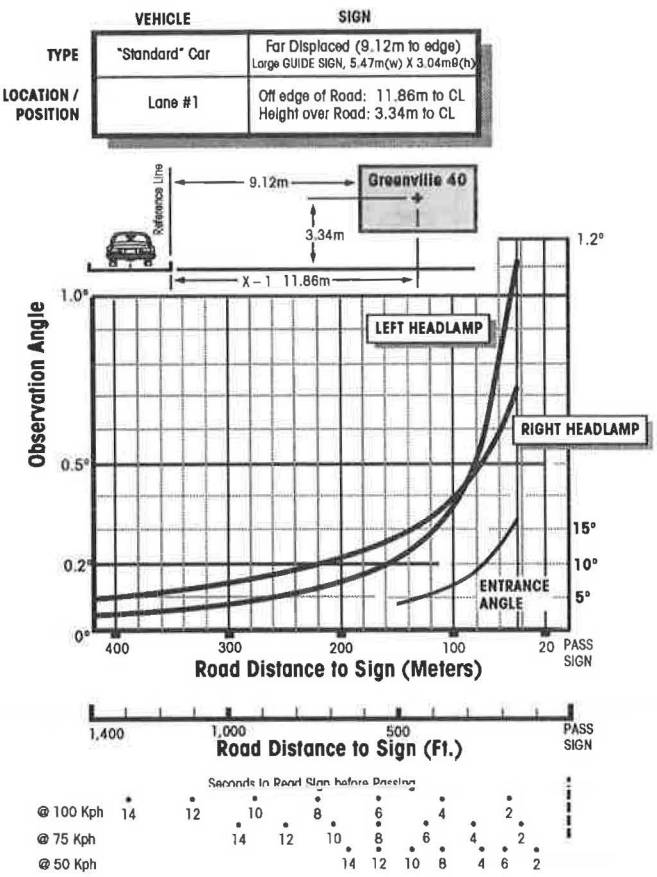


FIGURE 5 Observation angle versus road distance for signs at large offset from highway. Entrance angle shown where 5 degrees or more (to lower right vertical scale).

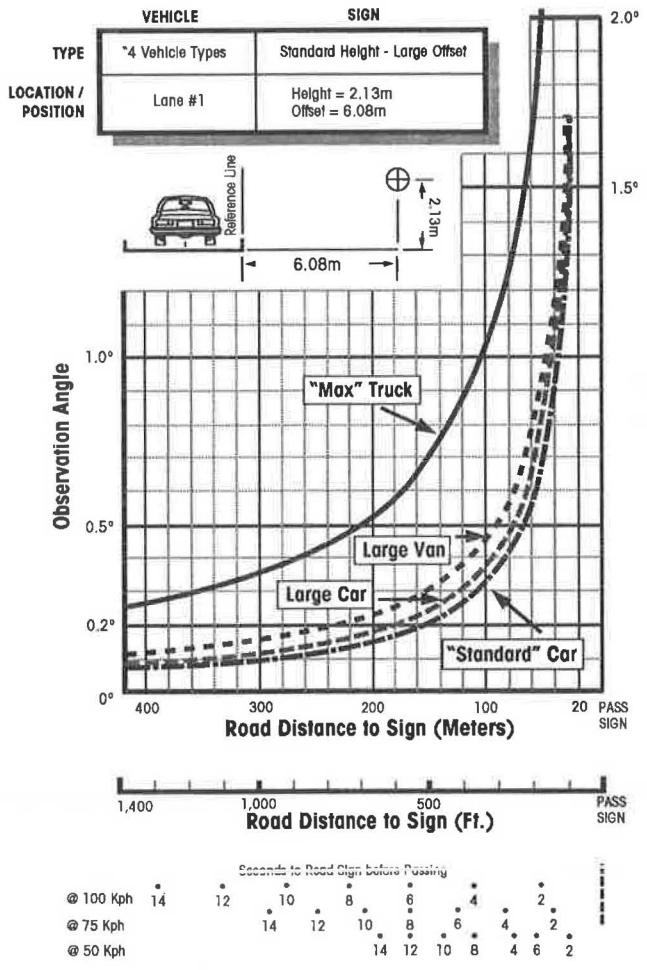


FIGURE 6 Observation angle versus road distance from different vehicles (left headlamp only) to same sign; right side, 2.13 m high.

RETROREFLECTIVE SIGNS: EFFECTIVE PERFORMANCE PRINCIPLES

Some significant principles relating to the effective performance of retroreflective sheeting on road signs can be noted in or deduced from ERGO data. The data developed in ERGO can also be a valuable addition to other experimental data and observations in correlating particular charateristics of reflective sheeting performance with the effectiveness of road signs at night. The four ERGO graphs relating to the geometry of signs (Figures 3–6) apply to and demonstrate the points discussed below. The graphed data are from ERGO calculations.

The most important determinant in the performance of retroreflective sheeting materials for any given application is the direct (although nonlinear) relationship of observation angle and road distance:

1. A specific observation angle value equals a specific road distance (for a given vehicle).
2. Observation angle increases as road distance decreases.

As shown on the three plots of ERGO data for different signs (Figures 3–5), the observation angle is very small (about 0.1 degree) when the road distance is as long as 280 m (919 ft) or more. As the vehicle approaches the sign from a great distance, the observation angle changes very slowly. Depending on the vehicle speed and the size of the sign, the approaching driver may see the sign for 10 sec at 0.1-degree observation angle. The observation angle begins to change more quickly as the driver passes 0.2 degree at 156 m (512 ft) to the sign. If the rate of travel is 75 kph, only 3 sec will elapse before the driver reaches 69 m to the sign and a 0.5-degree observation angle. Now the angle is changing rapidly: in only 1.5 sec the driver is at a 1.0-degree observation angle and in 0.8 sec at a 2.0-degree observation angle. ERGO computations show that at 2.0 degrees nominal, the driver passes through 1.3 degrees of observation angle in 0.5 sec (1.5 to 2.8 degrees).

Table 1 gives carefully computed road distance values corresponding (exactly, for the left headlamp) to the specific observation angle test points of various specifications. Also shown is the time before passing the sign and the time before the next given test point. The data are computed for geometry to a point on a sign offset at 2.5 m (8.2 ft) and at a height of 2.5 m (8.2 ft) viewed from the standard car.

TABLE 1 Observation Angle Nominal Points (Exact for Left Headlamp) Versus Road Distance and Travel Time at 75 kph

Observation Angle Specification Points (degrees)	Distance from Sign (m)	Time Before Passing[a] (sec)	Distance and Time to the Next Observation Angle	
			Distance (m)	Time[a] (sec)
0.1	306	14.6	148	7.1
0.2	156	7.5	57	2.7
0.333	99	5	30	1.4
0.5	69	3.3	30	1.5
1.0	38.7	1.8	10.5	0.5
1.5	28.2	1.3	5.5	0.3
2.0	22.6	1.1		

[a]At 75 kph.

The purpose of the summary in Table 1 is to provide an easy reference for the sign distances that correspond to particular observation angle test points. The left headlamp is used because it is generally the primary source, the observation angle for the right headlamp being generally larger. Table 1 is applicable to virtually all sign displacements (both offset and height).

All the plots of observation angle versus road distance (Figures 3–6) display this reality: observation angle is a direct function of road distance. Since the displacement distance from a headlamp to the eye is relatively fixed until approach distances quite close to the sign are encountered, the relationship is direct and occurs within a very narrow range for all cars and even for small trucks. In other words, a specific observation angle defines a specific distance to a sign. It follows that studies attempting to accurately correlate distance with some measure of actual sign performance (detection, legibility, etc.) for any or for several retroreflective materials and to draw conclusions from their data must recognize this relationship: different road distances involve different observation angles. For example, an R-value at 0.2 degree cannot correlate with performance at 60 m (197 ft), since at that distance the observation angle is about 0.6 degree. Review of test data (*6*) for different materials reveals that the relationship between materials at 0.2 degree is substantially different from the relationship between them at 0.6 degree. The objective of arriving at valid conclusions from these studies requires taking into account the actual observation angles that occur on the road as revealed by ERGO. The effective application of this method would require having reflectivity (R_A) data at all observation angles, that is, an observation angle curve (0.1 to 2.0 degrees at least).

Therefore, it would also follow that different retroreflective materials cannot be accurately characterized or referenced by a single R_A-value, as if the reflected light was an amorphous, uniform blob centered around the light source. This implies that the ratios between these single-number values hold for all road distances. However, test data studied (Stimsonite photometric laboratory data, 1992, unpublished) demonstrate that Material A can be substantially lower than Material B in its 0.2-degree observation angle laboratory values but nevertheless produces higher sign luminance than Material B at certain distances (*7*). Nevertheless, the use of single values is quite common in characterizing the relative retroreflectivity of materials.

Reflectivity values at the very small observation angles correspond to the longer distances and thus determine initial detection and overall reading time for most signs. Equal sign luminance is not equally important at all road viewing distances. In order to provide adequate reading time and so forth, which includes all the considerations that are used to select the size of a sign and its legend for daytime viewing, the distance at which initial detection and subsequent "primary" reading of a sign occur is the most important distance to see a sign, day or night. This distance involves correlates of very small observation and entrance angles. Subsequent continued reading as the driver approaches close to the sign at large observation angles must be, in the author's opinion, far less "necessary." Note the distances given by Zwahlen at which drivers no longer looked at the signs.

In addition to the common angles (observation and entrance angles), which have been emphasized in the data derived from ERGO and presented in this paper, ERGO data can also demonstrate what changes do or do not actually occur for the other angles of retroreflective geometry at different distances for a variety of sign locations. These other angular parameters also have an effect on the effective R_A-value and on actual road performance. ERGO data collected for this study for signs at various locations show that if a given material that is rotationally nonuniform is mounted with a uniform predetermined material orientation for all signs, it will simply result in maximum performance of the material at some sign locations and minimal performance at others. True orientation values range from -90 degrees to +90 degrees for actual signs.

CONCLUSIONS

The ERGO program data developed by users in support of their study requirements can contribute to the knowledge of the characteristics of effective performance of retroreflective sheeting on road signs. Since the values are absolute, they can contribute to valid conclusions from data relating study variables and retroreflectivity. Thus, soundly based and accurate data can replace blanket applications of simplistic beliefs about what geometries fully characterize the effective performance of retroreflective sheeting for road signs.

ERGO can be applied by users to any road viewing situation and any viewing distance. Successive increments of selected variables can be entered, the output geometry determined, and the resultant change in the correlates of observation and entrance angle can be evaluated. To that data, R-values resulting from laboratory tests at exactly the sign correlates given by ERGO can be added for a more accurate comparison of different retroreflective materials.

It is hoped that ERGO will be useful to those studying the application of retroreflective materials and that it will contribute to the accurate use of retroreflectivity data with other visibility parameters to promote the development of accurate and valid conclusions.

The ERGO program, including the complete mathematical basis, will be made available to all interested parties studying retroreflectivity. Requests should be sent to the author.

ACKNOWLEDGMENTS

The author wishes to acknowledge the support of the Stimsonite Corporation in the development of the ERGO program. He particularly wants to acknowledge the work of Dennis Couzin, an independent consultant, who was solely responsible for the mathematical basis of ERGO and was invaluable in helping to determine the effects of the various defined geometric parameters. He is also grateful to Craig Heinze, who converted the mathematical formulas into a workable program.

REFERENCES

1. *Instrumental Photometric Measurements of Retroreflective Material and Retroreflective Devices.* Federal Test Method 370. General Services Administration, March 1977.
2. Black, K. L., et al. *NCHRP Report 346: Implementation Strategies for Sign Reflectivity Standards.* TRB, National Research Council, Washington, D.C., 1992.
3. McGee, H. W., and D. L. Mace. *Retroreflectivity of Roadway Signs for Adequate Visibility: A Guide.* Report FHWA/DF-88/001. FHWA, U.S. Department of Transportation, 1987.
4. Johnson, N. L. Measuring the Appearance of Retroreflectors by Application-Oriented Goniophotometry. In *Review and Evaluation of Appearance: Methods and Techniques,* ASTM STP-914, Philadelphia, Pa., pp. 49–61.
5. Zwahlen, H. T. Stop Ahead Signs and Their Effect on Driver Eye Scanning and Driving Performance. In *Transportation Research Record 1168,* TRB, National Research Council, Washington, D.C., 1988, pp. 16–24.
6. Woltman, H. L., and T. J. Szczech. Sign Luminance as a Methodology for Matching Driver Needs, Roadway Variables, and Signing Materials. In *Transportation Research Record 1213,* TRB, National Research Council, Washington, D.C., 1989, pp. 21–26.
7. Zwahlen, H. T., Q. Li, and J. Yu. Luminance Measurements of Retroreflective Warning Signs Under Lowbeam and Highbeam Illumination at Night Using the CapCalc System. In *Transportation Research Record 1316,* TRB, National Research Council, Washington, D.C., 1991, pp. 31–38.

The contents of this paper represent the views and findings of the author, who is solely responsible for the opinions and conclusions presented here. The contents do not necessarily reflect the official views or policies of the Stimsonite Corporation.

Publication of this paper sponsored by Committee on Visibility.

Decision Support System for Controlling Traffic Signals

AYELET GAL-TZUR, DAVID MAHALEL, AND JOSEPH N. PRASHKER

In many metropolitan areas traffic control is monitored from control centers. The operator of a control center is asked to make a quick decision and to modify the signal programs of the urban signal network in whole or in part. In order to make a proper decision, the operator must consider a wide range of alternatives and evaluate their expected effects on the whole traffic system. The complicated structure of the problem and the routine occurrence of random events demonstrate the complexity of the decision process in traffic control. A procedure will be described that is aimed at supporting this decision making. The procedure is characterized by the systematic scanning of a wide range of alternatives and includes a special algorithm for reducing the size of the problem and concentrating on the most promising strategies. A statistical decision tree is used for spanning all alternatives and expressing the subjective priorities among them and the projection regarding their consequences. An important option given to the controller is the ability to acquire more information to support his decision by using on-line simulations. This option is time consuming and therefore has a cost. The operator is given the tools to decide whether the additional information is worth the price. In addition, the system contains a systematic procedure to "learn" from past experience and to improve its ability to make decisions under uncertainty conditions.

The growth of congestion in urban networks and the consequent constraints imposed on mobility have made it vital to manage and utilize the existing infrastructure more efficiently. One of the most prominent procedures available for managing traffic control is that of monitoring traffic-signal programs. Research has attempted to find the optimal signal-timing program for a group of intersections during peak hours (*1–4*). Attempts (*5*) were also directed at finding a global optimum for a group of coordinated traffic signals. These programs were prepared off-line so that an operator in a control center could choose the most appropriate program off the shelf, as it were. More recently, efforts have aimed at developing responsive methods (*6*), which are primarily designed to respond to fluctuations in traffic volumes without external intervention. The main doubt concerning the responsive methods is over their ability to converge to a "good" system optimum; the question is whether they merely provide a local solution at the cost of finding a strategy that might better improve the whole system.

The complexity of the control problem stems from the following properties of the system:

1. **Objectives**—The problem has several objectives that should be met simultaneously, for example, minimization of delay, queues, number of stops, energy consumption, and environmental impacts; some of these objectives contradict one another.

Civil Engineering Department, Technion-Israel Institute of Technology, Haifa, Israel.

2. **Dependency**—In an urban network the output of one intersection is the input of another, and the queue at one intersection can block another intersection. As a result, an intersection cannot be treated individually and must be coordinated with its environment. Consequently, the number variables that have to be calculated simultaneously (e.g., cycle time, green splits, and offsets) becomes very large.
3. **Parameter values**—The large number of parameters describing the network (e.g., saturation flows, acceleration times, platoon dispersion, arrival distribution) and the uncertainty regarding their values make it necessary to consider a range of values.
4. **Mathematical model**—The relationships among the various parameters, variables, and objectives are of a complicated nature; attempts to formulate them into one mathematical model end in inadequate results.

This complicated structure of the problem and the routine occurrence of random events demonstrate the complexity of the decision process of traffic control. Logically, however, one may believe that in the future this process will still involve some degree of human judgment and that the operator in a control center will still play an important role.

Described in this paper is a procedure aimed at supporting the decision making of an operator in a traffic-control center. The procedure is characterized by the systematic scanning of a wide range of alternatives. It includes a special algorithm for reducing the size of the problem and concentrating on the most promising strategies. An important option given to the controller is the ability to acquire more information to support his decision. This option has a cost, though, and the operator is given the tools to decide whether the additional information is worth the price. In addition, the system contains a systematic procedure to "learn" from past experience and to improve its ability to make decisions under uncertainty conditions.

In the next section, a description of the structure, the nature, and the dimensionality of the control problem will be given. Next, a tool will be presented for examining the various strategies and options available to the operator of the control center. Then a procedure to reduce the size of the decision process is described, and, finally, the machine-learning ability of the procedure, that is, an automatic process for collecting and then transferring data into useful knowledge, will be discussed.

CHARACTERISTICS OF TRAFFIC-CONTROL PROBLEM

A traffic-control program is a set of parameters (e.g., cycle length, green split, offsets) that control the right-of-way and

that assign priorities among the links of the network. These programs determine the level of service of each link and as a result can affect the routes that drivers choose. During congestion, the amount of green light assigned to a link can actually determine the traffic volume on that link. This ability can be used for controlling the number of vehicles allowed to enter a congested zone.

The traditional method of coping with the massive size of the problem is to divide it into several stages and subproblems:

1. **Division into zones:** The network is divided into several subnetworks, and each can be considered separately. Some degree of dependency between zones is allowed.
2. **Type of signal strategy:** On the basis of various arguments, a general strategy is selected for each zone, for instance, green wave (*7*), critical intersection control (*3*), network design (like TRANSYT). Each strategy can have several variants, such as green wave with one band or a multiband design (*8*) or a network design with various weights assigned to different links.
3. **Design parameters:** At this point, the various parameters are computed for the signals.

Usually, the decisions of the controller do not explicitly take into account the uncertainty regarding changes in demand, changes in routes, incidents, and other random events that might affect the value of many parameters (e.g., saturation flows, start-up delay, acceleration, and speeds). All these random events reject the assumption of stationary conditions and promote the need to combine probabilities and stochastic considerations into the decision process.

The wide range of parameter values together with several control strategies and tactic decisions, and the possible evaluation of each combination through simulation, increase dramatically the dimension of decision space and the number of alternatives that should be considered. The resultant huge dimension complicates the decision process and necessitates basing it more on a systematic process and less on an intuitive one. The tool proposed for handling this problem is the statistical decision tree.

STATISTICAL DECISION TREE AS A DECISION-MAKING TOOL

A decision tree is based on Bayes' decision theory, which formulates decision-making processes under uncertainty. The tool is suitable for situations in which one course of action must be selected from several possible acts; their respective outcomes are known with a certain degree of confidence, but not absolutely; and there are ways to increase the level of confidence by gaining information. The question that this theory wishes to answer is, Which course of action should be taken in order to maximize the expected benefit (or to minimize the expected loss)? Is it worthwhile to "pay" for extra information?

In order to use the Bayesian approach, the following data should be known in advance:

1. Possible courses of action, i.e., the signal programs;
2. Possible outcomes of each signal program and the probability of occurrence of each outcome;
3. The utility gained, given the occurrence of a specific outcome;
4. Possible experiments (simulation) that can be conducted in order to gain information about the probability of occurrence of each outcome and the cost associated with each simulation;
5. Possible results of each simulation; and
6. The probability of occurrence of each outcome, given a specific result of a simulation.

Traffic engineers face several objectives that they wish to achieve simultaneously: minimum delays, maximum throughput, minimum queue lengths, prevention of spillbacks, minimum number of stops, minimum fuel consumption, and so on. A possible outcome can be expressed as a function of these variables, for example, a success can be defined as a condition in which all values are in some critical region.

Often, the decision maker is faced with a situation in which it is believed desirable to obtain more knowledge in regard to the likelihood of a possible outcome after implementation of a certain signal program. Gaining such information is possible by running a short-term simulation. The results of such a simulation can then be used to update prior probabilities and to obtain posterior probabilities. The time needed to run the simulation is considered a cost.

All the above components—that is, the possible programs and their results and utilities, the possible simulations and their results and cost, prior probabilities reflecting the level of confidence before the simulation, and posterior probabilities reflecting the level of confidence after the simulation—make up the decision tree. The expected utility of each branch is calculated by a backward search, and the branch with the maximum expected utility is chosen. Thus, the statistical decision tree answers not only the question of the course of action that seems the most beneficial, but also that of the possible simulations that are worthwhile to conduct.

Figure 1 shows how the components of the decision-making process integrate into a complete decision support system (DSS). The DSS should have access to updated data regarding traffic conditions in the network. Expected volumes during the planning horizon can be estimated by an external algorithm that can exchange information with the DSS, or they can be extracted from existing data bases. The operator provides the system with the planning strategies most appropriate for the conditions in the network. On the basis of this information, alternative timing programs can be calculated in one of two ways: (*a*) by attaching external software packages (such as TRANSYT or PASSER) to the DSS through the appropriate interface (such connections are feasible in most software packages) or (*b*) by writing design procedures for signals as part of the DSS. The next two functions of the system, the tree-building stage and the tree-search stage, are the two modules at the heart of the DSS. They should be designed and programmed especially for these purposes. When these two tasks are completed and the simulations to be executed are chosen, they can be executed by using a software (e.g., NETSIM) that can interact with the computer language in which the DSS is programmed. Finally, through the user interface, the DSS should instruct the operator which program to implement.

Figure 2 shows the structure of a decision tree for a case in which two programs are considered (X and Y). The first

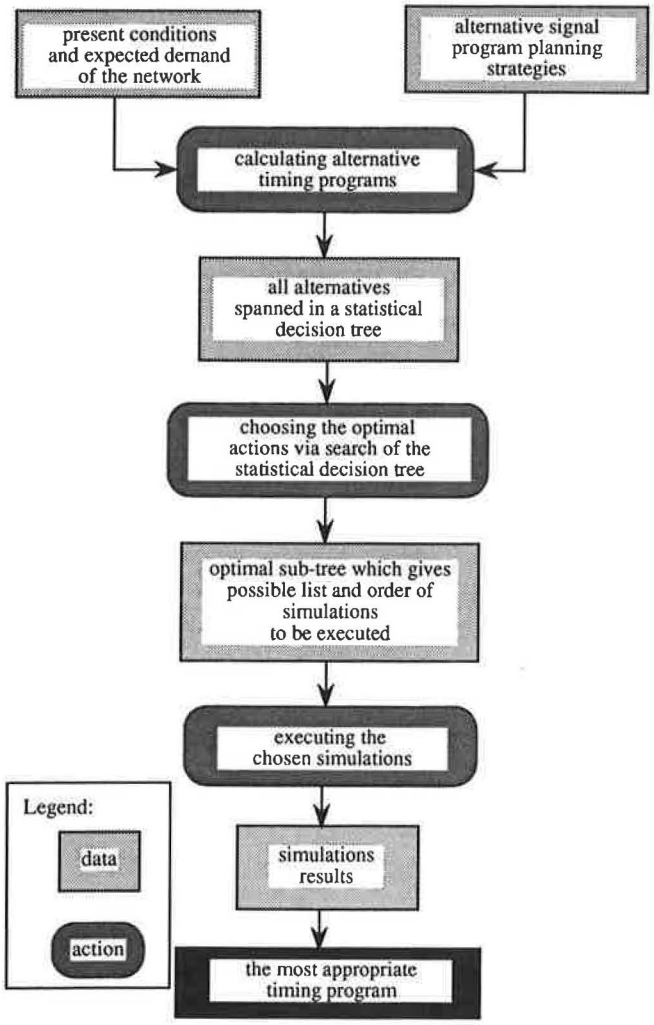


FIGURE 1 Main components of DSS.

decision node has four courses of action (alternatives): implement X, implement Y, simulate X, simulate Y. Each course of action has two possible outcomes (success or failure) with its probabilities. Some outcomes result in a termination node and some in new decision nodes, that is, to implement a signal program or to run another simulation.

The size of the tree is a function of the number of possible alternatives, that is, the number of strategies (R), the number of demand levels (traffic volumes M), and the number of programs taken from the shelf (B). Since all parameters are fully defined for a shelf program and, for strategies, should be estimated together with volumes, the total number of possible signal programs (N) may be denoted

$$N = R \cdot M + B \tag{1}$$

Testing the performance of each timing program under each possible pattern of traffic volume creates $M \cdot N$ possible simulations. Each branch in the decision tree represents the execution of a certain number of simulations (varying from 0 to $M \cdot N$), in a certain order. The number of branches having i simulations is computed as follows:

$$i! \binom{MN}{i} = \frac{(MN)!}{(MN) - i)!} \tag{2}$$

Each simulation, similarly to each timing program, has two possible outcomes (failure or success), so the number of leaves of a branch containing i simulations is

$$2N \cdot 2^i \tag{3}$$

On the basis of Equations 1–3, the total number of termination nodes on a tree becomes

$$2N \cdot (MN)! \sum_{i=0}^{MN} \frac{2^i}{(MN - i)!} \tag{4}$$

To demonstrate, assume a tree with three strategies, two levels of volumes, and one shelf program. The number of possible final outcomes would be $3 \cdot 10^{16}$.

TRUNCATION PROCESS

The dimension of the tree and the time it takes to develop all its branches necessitate that the decision maker decide what parts of the tree to span and what to neglect. Truncating a branch implies that some possible actions will not be considered in detail in later steps. To accomplish a justified truncation, the operator should be provided with a quantitative figure of what might be lost if a certain branch of the tree were ignored. The approach to the problem here was based on developing upper bounds for the expected utilities of branches whose examination involves a cost. With this upper bound, the decision maker can then answer questions such as, Is it worth running a simulation for program j?

The calculation of upper bounds exploits the nature of the control problem and the dependency relationships among the signal programs. The main assumptions are discussed in the following sections.

Assumption 1

The success of a certain program increases the posterior probabilities for the success of all other programs; that is,

$$p(i|\{G,S_k\},\{F\}) - p(i|\{G\},\{F\}) \geq 0 \tag{5}$$

where

G = group of simulations that were executed and succeeded,
F = group of simulations that were executed and failed,
$p\ (i|\{G\},\{F\})$ = conditional probability of program i to obtain successful results, given groups G and F,
$p\ (i|\{G,S_k\},\{F\})$ = conditional probability of program i to obtain successful results, given that group G and simulation k succeeded and group F failed.

This assumption is motivated by the knowledge that the network is not oversaturated and that a signal program can improve conditions.

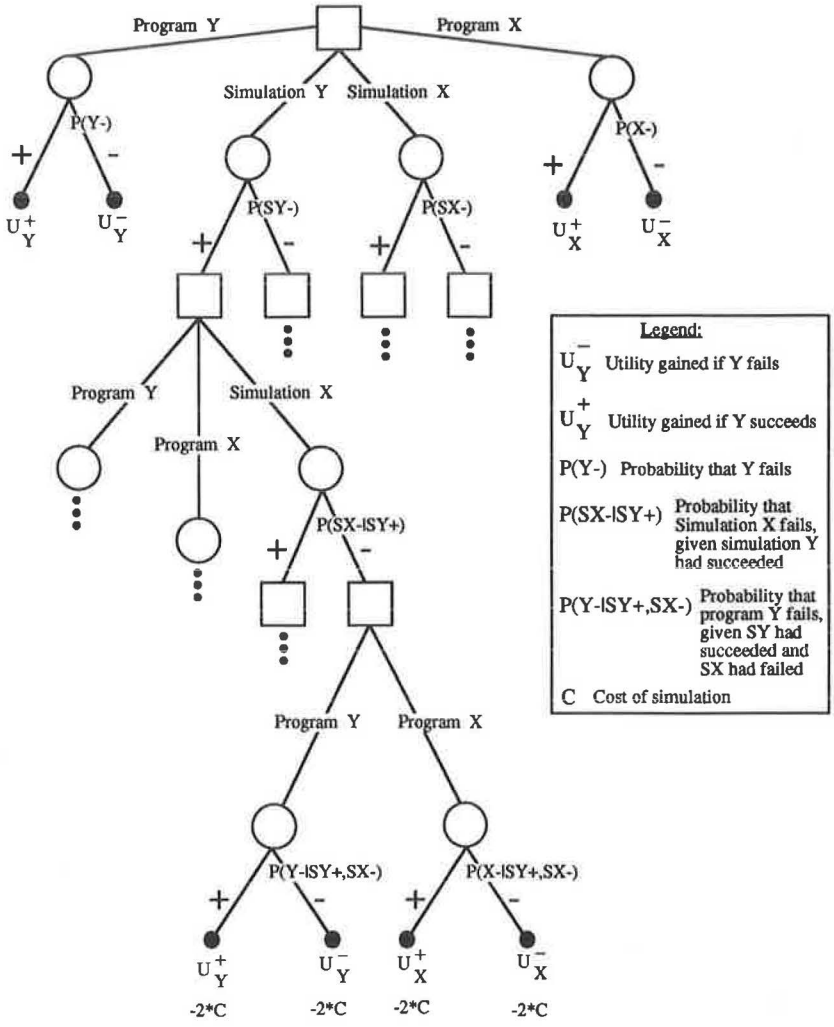


FIGURE 2 Decision tree with two signal programs.

Assumption 2

An additional increase in probabilities following the success of a certain simulation is greatest for the program that was examined, that is,

$$p(i|\{G,S_i\},\{F\}) - p(i|\{G\},\{F\}) \geq p(k|\{G,S_i\},\{F\}) - p(k|\{G\},\{F\}) \quad (6)$$

where i and k are different signal programs.

This assumption simply expresses the obvious fact that the largest contribution of running a simulation accrues to the program examined by the simulation.

Assumption 3

An additional increase in probability following the success of a certain simulation decreases as the number of simulations preceding it increases.

$$p(i|\{G,S_k\},\{F\}) - p(i|\{G\},\{F\}) \geq p(i|\{G,S_m,S_k\},\{F\}) - p(i|\{G,S_m\},\{F\}) \quad (7)$$

This is simply a manifestation of the law of diminishing returns.

After some laborious operations (*9*), it can be shown that the maximum loss of expected utility following the relaxation of a certain simulation is calculated according to

$$p(S_i|\{\emptyset\},\{\emptyset\}) \cdot [EU(i|\{S_i\},\{\emptyset\}) - EU(i|\{\emptyset\},\{\emptyset\})] - C \quad (8)$$

where

$p(S_i|\{G\},\{F\})$ = the probability of success of simulation i given groups G and F,
$EU(i|\{G\},\{F\})$ = the expected utility of program i given groups G and F,
C = the cost of simulation, and
$\emptyset$ = the null set.

This means that the maximum loss of not running simulation is the difference between the expected prior utility and the expected posterior utility of success, provided that the simulation was the first simulation to be run and that it succeeded, multiplied by the probability of success minus the cost of simulation i.

The practical question answered by the mathematical expression above is, Which simulation should be relaxed in the truncation process in order to minimize the expected max-

imum loss of expected utility? Two attributes characterize the simulation to be relaxed:

1. Its contribution to the information about the probability of success of the program tested by it is low.
2. The probability of obtaining successful results when executing it is low.

By multiplying the values of both attributes by each other, the integrated criterion is achieved.

NUMERICAL EXAMPLE

This example demonstrates a decision problem in which the controller has to choose between one of three possible strategies $\{A_1,A_2,A_3\}$. In the case of a successful implementation, the utilities of the three strategies are 370, 340, and 310, respectively. In the case of a failure, they all have a cost of 50 (utility of -50). Figure 3 demonstrates the six possible decisions of the first decision node: to select one of the three strategies or to run a simulation of each one of them. At this stage, the expected utility of the three strategies can be computed and it can be seen that Strategy 1 has the largest expected utility. Instead of selecting a strategy, the controller can run a simulation and can make a decision after obtaining the results. Figure 4 shows part of the tree following the success of simulating A_1. The best process is denoted by a bold line. According to these rules, the best path is as follows: simulate A_1; if it succeeds, simulate A_2; if it succeeds, select A_2; if simulation of A_2 fails, select A_1; if simulation of A_1 fails, select A_3.

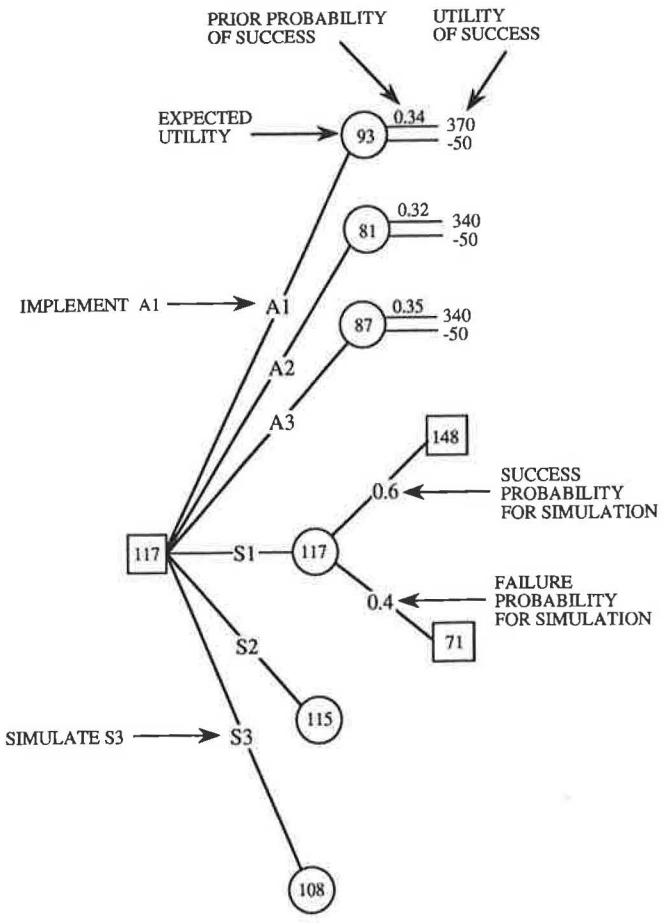


FIGURE 3 First decision node of decision tree.

The value of the simulation is illustrated through several facts:

1. **The change in probabilities:** The success of a simulation increases the posterior probability; for example, the prior probability of success for A_1 is 0.34 and the posterior probability if the simulation succeeds is 0.48. The difference is a result of the extra information.
2. **The change in expected utility:** In this example, without simulation the maximum expected utility is 95, but after running a simulation, this value increases to 117, that is, an additional utility of 24.

If one considers the relaxation of simulation A_1, the upper bound for the utility of this part of the tree should be computed. According to Equation 8 and Figure 4, the calculation is as follows:

$$p(S_i|\{\emptyset\},\{\emptyset\}) = 0.6$$

$$EU(i|\{S_i\},\{\emptyset\}) = 152$$

$$EU(i|\{\emptyset\},\{\emptyset\}) = 93$$

$$C = 10$$

Thus,

$$p(S_i|\{\emptyset\},\{\emptyset\}) \cdot [EU(i|\{S_i\},\{\emptyset\}) - EU(i|\{\emptyset\},\{\emptyset\})] - C = 25$$

The upper bounds of relaxing all other simulations can be computed similarly, and the one with the lowest upper bound is chosen not to be spanned. The upper bound of Simulation 2 is 20 and the upper bound of Simulation 3 is -7. It can be observed that Simulation 3 has the lowest upper bound. Moreover, the negative sign indicates that no loss in expected utility would be obtained as a result of relaxing Simulation 3. This phenomenon is due to the low contribution of Simulation 3 to the information in hand compared with its cost.

DISCUSSION OF RESULTS

Using the decision tree as part of a DSS actually divides the decision-making process into two stages. In the first stage, the controller answers the question of what simulations to run or where one needs to improve one's knowledge at a certain cost. In the second stage, the controller searches for the branch that maximizes expected utility. Afterwards, the necessary simulations are performed and a decision is reached. In this way, the controller can adopt the process that is most suitable for existing conditions.

Between these two stages, the truncation option allows the DSS to limit the time dedicated to the search stage by trimming some branches of the tree. This truncation process might result in a loss of expected utility, but the maximum value of this loss is known in advance to the operator. Appropriate criteria are used to decide whether the truncation is worthwhile.

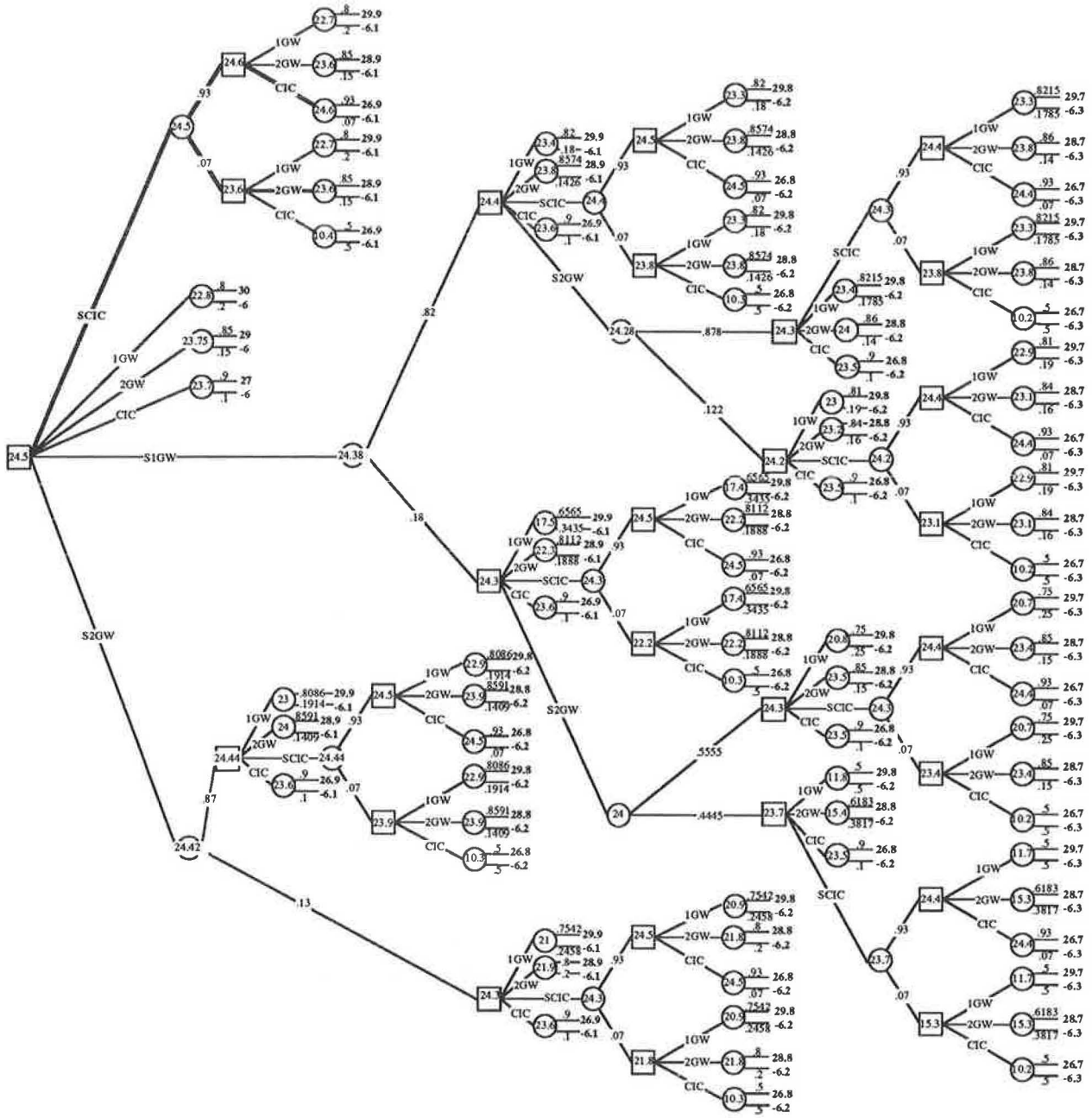


FIGURE 4 Part of a decision tree for case of successful simulation of A_1.

The DSS described in this paper is still under development and has not yet been fully integrated with all necessary components. The DSS should be designed to run on a work station, and the complexity of the tasks it is meant to perform should fall within the range that this type of equipment can handle.

The quality of the decision-making process described in this paper depends heavily on the amount and quality of information and knowledge available to the controller. This need can be satisfied gradually over time, especially if experience is transferred through machine learning into practical expressions like prior probabilities, success probabilities following simulations, and so forth.

Machine learning is an automatic process of collecting data and transferring it into useful knowledge. This process is done efficiently if, from the limited knowledge that is collected every day, a large number of parameters can be updated. This broad inference should be based on a deep understanding of all dependencies and relationships existing in the system.

ACKNOWLEDGMENT

This work was supported by the Office of the Traffic Commissioner, Israeli Ministry of Transportation.

REFERENCES

1. Singh, M. G., and H. Tamura. Modelling and Hierarchical Optimization for Oversaturated Urban Road Traffic Network. *International Journal of Control*, Vol. 20, No. 6, 1974, pp. 913–934.
2. Michalopoulos, P. G., and G. Stephanopoulos. Optimal Control of Oversaturated Intersections: Theoretical and Practical Considerations. *Traffic Engineering and Control*, Vol. 19, No. 5, 1978, pp. 216–222.
3. Michalopoulos, P.G., and G. Stephanopoulos. An Algorithm for Real-Time Control of Critical Intersections. *Traffic Engineering and Control*, Vol. 20, No. 1, 1979, pp. 9–15.
4. Lieberman, E. B. Concepts of Control for Oversaturated Networks. Presented at meeting of TRB Committee on Traffic Signal Systems, Minneapolis, Minn., 1990.
5. Robertson, D. I. *TRANSYT : A Traffic Network Study Tool.* TRRL LR 253. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1969.
6. Robertson, D. I., and P. B. Hunt. A Method of Estimating the Benefit of Co-ordinating Signals by TRANSYT and SCOOT. *Traffic Engineering and Control*, Vol. 23, No. 11, 1982, pp. 527–531.
7. Little, J. D. C., M. D. Kelson, and N. H. Gartner. MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks. In *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981, pp. 40–46.
8. Gartner, N. H., S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND—A Variable-Bandwidth Arterial Progression Scheme. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 212–222.
9. Gal-Tzur, A., and D. Mahalel. Evaluating the Option to Truncate Decision Trees. *TRI Research Report* (in Hebrew). Transportation Research Institute, Technion, Haifa, Israel, 1992.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Methodology for Evaluating Traffic Detector Designs

JAMES A. BONNESON AND PATRICK T. MCCOY

The design of the traffic detection scheme at an intersection can have a considerable impact on traffic safety and efficiency. A detector design (i.e., the detector layout and controller timing) that is not "tuned" to the geometrics of the intersection and its traffic demands can result in higher motorist delays than would be obtained with pretimed control. The detector design can also have an effect on safety. Designs that continually present the yellow signal to drivers when they are in the zone of indecision are likely to be associated with more accidents than designs that detect these drivers and extend the green signal for them until they are clear of the intersection. The safety and efficiency of a traffic detector design can be determined from the probability of max-out [a max-out occurs when the green is extended by a continuous stream of arrivals until the maximum green duration is reached (and a conflicting call is continuously held on one or more phases)] and the amount of time spent waiting time for gap-out and subsequent phase change. A detector design that minimizes these measures of effectiveness should provide safe and efficient operation. Achieving the optimal combination of these measures can be difficult because of complex interactions among detector design elements (i.e., detector location, detector length, vehicle speed, passage time setting, and call-extension setting). The methodology described in this paper will allow the designer to determine the optimal combination of design elements in terms of safety (via infrequent max-out) and operations (via a short waiting time for phase change).

Traffic-actuated control can be used to improve both traffic efficiency and safety at an intersection. The extent to which it improves the efficiency of operations is dependent on the detector layout and its associated controller settings (henceforth referred to as the "detector design"). Designs that are not "tuned" to the geometrics of the intersection and its traffic demands can result in higher motorist delays than would be obtained with pretimed control.

The detector design can also have an effect on safety. Studies of driver behavior at intersections indicate that there is a zone on the approach wherein driver response to the yellow presentation is unpredictable and uncertain; some drivers may decide to stop, whereas others may determine that it is safer to proceed through the intersection. As a result, there is an increased potential for rear-end accidents at the end of the phase when two or more drivers are simultaneously in this zone of indecision. Some agencies use advance detection in this zone to monitor traffic flow and extend the green to any vehicles in the indecision zone (thereby preventing the presentation of a yellow signal). These designs are believed to be safer than designs without advance detection because they effectively reduce the number of potential accident events (e.g., rear-end collisions) occurring at the intersection.

A methodology for evaluating existing traffic detector designs on intersection approaches is described. This methodology is applicable to either presence or pulse-mode detection on low- or high-speed approaches. It is based on a constant passage time (or vehicle extension) setting on the controller and thus is not directly applicable to volume-density controllers using a gap-reduction feature. Evaluation criteria include the frequency of phase max-out [a max-out occurs when the green is extended by a continuous stream of arrivals until the maximum green duration is reached (and a conflicting call is continuously held on one or more phases)] and the time waiting for gap-out after queue service. Designs that minimize these criteria should provide both safe and efficient operations.

TRAFFIC DETECTOR DESIGN PHILOSOPHY

Traffic detector designs are generally formulated to achieve both safe and efficient traffic operations. The degree to which each goal is achieved is based primarily on intersection approach speeds—efficiency receives most of the attention on low-speed approaches, whereas safety may receive greater attention on high-speed approaches. Most detector designs are based on the principle that the stop line detector will be used for traffic queue service (i.e., minimize delay). Designs with advance detection are based on the principle that the advance loop will be used to minimize the number of times that drivers are caught in the indecision zone at the end of the phase (i.e., maximize safety).

Recent research by Lin (*1*) on actuated intersection operation indicates that the length of the stop line detection zone and its detector unit settings have a significant impact on motorist delay. The stop line detector should be designed to minimize the frequency of premature phase gap-out and the frequency of calls to empty approaches. Detector length and vehicle extension combinations that minimize delay have been reported by Lin (*1*).

Advance detector design is based on the location of detectors at one or more locations to provide indecision-zone protection to vehicles traveling within the design speed range. This design speed range typically bounds the range of speeds commonly found on the approach. Detectors are then located throughout the indecision zone on the basis of the design speed range. In operation, these detector loops are positioned such that a vehicle traveling at a speed within the design speed range will be able to maintain a continuous call for green

Civil Engineering Department, University of Nebraska-Lincoln, Lincoln, Neb. 68588-0531.

(assuming that it initially enters the detection zone during the green) until it clears the intersection. Vehicles at speeds above or below the design speed range will still place calls and extend the green; however, they will not be provided protection for the full length of their indecision zone.

EVALUATION METHODOLOGY

Maximum Allowable Headway

The methodology described here is based on the use of a detector design's maximum allowable headway (MAH) to evaluate the safety and efficiency of its operation. MAH represents the maximum time separation between successive calls for continued green (note that MAH is not necessarily the minimum vehicular headway). The relationship between MAH and the passage time setting is shown in Figure 1 for the simple case of one phase serving one traffic lane. The MAH for a phase serving several lanes can be much more complicated and may, in fact, not truly be a constant value.

The detector design evaluation is made on a phase-specific basis and requires that all lane groups served during the phase be identified. The lane group definition used here is consistent with that provided in the *Highway Capacity Manual* (2). Any turn movement made from an exclusive lane (or lanes) would be designated as a lane group. The approach lanes allocated to the through movement and any turn movements not provided an exclusive lane would also be designated as a lane group. Shared lanes with one high-volume movement should be examined to determine if one lane operates as a de facto exclusive lane and thus a separate lane group.

Once the lane groups have been determined, the MAH for each group must be determined. The procedure for combining the lane-group MAHs into an equivalent MAH for the phase is described in a later section. In general, the MAH for a detector design in any one lane group represents the maximum allowable headway between successive calls from vehicles in that group. In order to gap-out a phase, each lane group would have to experience successive call headways that exceed its respective MAH.

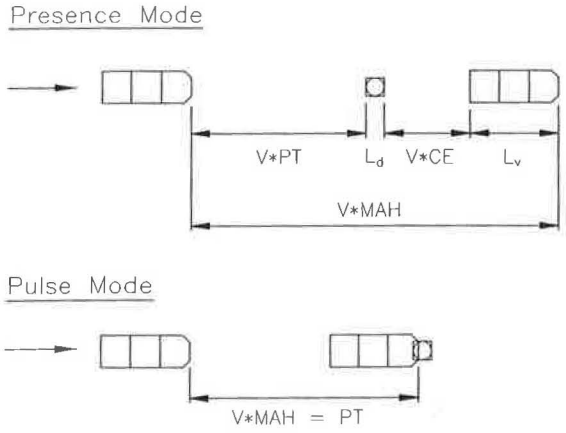


FIGURE 1 Relationship between maximum allowable headway and passage time, where *MAH* = maximum allowable headway; *PT* = passage time setting on the controller; *CE* = call-extension setting on the detector unit; *V* = free flow speed; L_d = length of detector in the direction of travel; and L_v = detected length of vehicle.

The lane group MAH is dependent on a number of design parameters, including the number of loops serving the lane group, the length of these loops, and the distribution of vehicle speeds for the lane group. Because of the wide range of design parameters, selected design types are treated individually instead of one generalized procedure being developed. Because the use of advance detection represents the most fundamental difference among designs, it will be used as the primary point of departure in describing the MAH calculation.

Engineers responsible for advance detector design typically adopt one of two general design goals. Some engineers prefer to carry the clearing vehicle just through the indecision zone. In this paper, this design is referred to as a Goal 1 design. Other engineers prefer to carry the vehicle to the stop line. This is referred to as a Goal 2 design. Thus, the desired design goal represents a secondary point of departure in the MAH calculation.

The procedure for calculating the lane-group MAH is based on several assumptions. One is that single detector units will be used to monitor all adjacent lanes at any one point on the approach for a given lane group. This type of detection could be achieved by having a single, wide loop at the given point or by having one loop in each lane at this point and wiring them together. A second assumption is that the design speed range for the advance detectors (if any) will include at least 70 percent of all vehicles in the lane group. A third assumption is that all advance detectors will operate together such that a vehicle moving at a speed within the design speed range will maintain a continuous call for green as it traverses these advance loops. A final assumption is that the time headway between successive calls is exponentially distributed.

Lane Groups with Only a Stop Line Detector

For this type of design, the MAH is equal to the MAH for the stop line detection zone (MAH_s). This quantity can be calculated using the following equation:

$$MAH_s = PT + CE_s + \frac{L_{ds} + L_v}{V_a} \quad (1)$$

where

PT = passage time (PT) (or vehicle extension) setting (sec),
CE_s = call-extension (CE) setting on the stop line detector unit (sec),
L_{ds} = length of the stop line detection zone (m),
L_v = detected length of vehicle (m), and
V_a = average running speed on the intersection approach of the subject lane group as measured during the unqueued portion of the green (mps).

Equation 1 is based on the assumption that the detector unit is operating in the presence mode. If it is operating in the pulse mode, then MAH_s would equal the PT setting.

Lane Groups with One or More Advance Loop Detectors

For those lane groups with one or more advance loop detectors, the MAH is a function of the average vehicle's travel time from the first advance loop detector to the end of the indecision zone. The location of the end of this zone depends on the design philosophy adopted. If a Goal 1 philosophy is taken, the end of the indecision zone is defined to be about 1 or 2 sec of travel time upstream of the stop line detector. If a Goal 2 philosophy is taken, the end of the zone is defined to be at the stop line.

Regardless of which philosophy is adopted, the settings on the stop line detector unit must be considered in determining the MAH. If the stop line detector is active (i.e., not operating with call delay) during the green, then the MAH will be increased by the magnitude of the CE setting on this unit and the PT setting on the controller. If the call-delay feature of the stop line detector unit is invoked during the green [such as with an EC-DC (extended-call/delay-call) detector unit], the detector is essentially inactive during the green and a clearing vehicle will not place a call at the stop line.

MAH for Goal 1 Designs If the detector design for a lane group reflects a Goal 1 philosophy, its MAH can be calculated using the following equations:

$$MAH = MAH_a + MAH_s \tag{2}$$

$$MAH_a = PT + CE_n + \frac{D_1 - D_n + L_d + L_v}{V_a} \tag{3}$$

where

MAH_a = maximum allowable headway for the advance loop (or group of advance loops) (sec);
MAH_s = maximum allowable headway for the stop line detection zone (from Equation 1) (sec);
n = number of advance detectors, $n = 1, 2, 3, \ldots$;
D_1 = distance to the leading edge of the advance detector furthest from the stop line, as measured from the stop line (m);
D_n = distance to the leading edge of the advance detector nearest to the stop line, as measured from the stop line (m); and
L_d = length of an advance loop detector (all advance loops are assumed to have the same length) (m).

The additive property of the MAHs for the two detection zones stems from the independence assumption made when using the exponential distribution. This assumption is reasonable for lane groups with two or more lanes and should yield conservative (i.e., slightly higher) values for the single-lane lane group.

As with Equation 1, Equation 3 is based on the assumption that the detector unit is operating in the presence mode. If it is operating in the pulse mode, MAH_s would equal PT and MAH_a would be calculated with CE_n, L_d, and L_v equal to zero.

If the stop line detector unit is inactive (i.e., call delay in operation) during the green, MAH_s is zero and the MAH for a Goal 1 design would equal MAH_a only.

MAH for Goal 2 Designs If the detector design in a lane group reflects a Goal 2 philosophy, its MAH can be calculated using the following equations:

$$MAH = \text{larger of} \begin{cases} MAH_t \\ MAH_a \end{cases} \tag{4}$$

$$MAH_t = PT + CE_s + \frac{D_1 + L_v + SL - SB}{V_a} \tag{5}$$

where

SB = distance between the trailing edge of the stop line detection zone and the nearest edge of the crossing travel path (m); and
SL = distance between the stop line and the nearest edge of the crossing travel path (m).

Equation 4 is based on the assumption that the advance and stop line detector units are operating in the presence mode. If they operate in the pulse mode, the values for CE_s, CE_n, SL, SB, L_d, and L_v should be set to zero for Equations 3 and 5.

If the stop line detector unit is inactive during the green, the MAH for a Goal 2 design would equal MAH_a, as calculated from Equation 3.

Max-Out Probability

One measure of intersection performance is the frequency of phase termination by max-out. As stated earlier, a max-out occurs when the green is extended by a continuous stream of arrivals until the maximum green duration is reached (and a conflicting call is continuously held on one or more phases). When a max-out occurs, the yellow indication is presented regardless of whether a vehicle is in the indecision zone. Of course, the more frequently drivers are caught in the indecision zone during the yellow, the more frequent will be situations where one driver decides to stop and a following driver decides to go. Thus, it is likely that the frequency of max-outs is positively correlated with the frequency of rear-end accidents.

One model for predicting the frequency with which a phase maxes out can be formulated by assuming that all calls extending the green emanate from a randomly arriving traffic stream. The distribution of these calls is assumed to be exponentially distributed with a mean flow rate (q) equal to the sum of the flow rates in each lane group served during the phase. This assumption is most valid for phases serving more than one lane because the frequency of small headways (i.e., those less than 2 sec) measured across multiple lanes is more consistent with that predicted by the exponential distribution.

The probability of a max-out can be equated to the joint probability of there being a sequence of calls to the phase in service, each with a headway less than the MAH. This probability can be stated mathematically as follows:

$$P(\text{max-out}) = (1 - e^{-qMAH})^n \quad (6)$$

where

$q = q_1 + q_2 + \ldots + q_m$;
m = number of lane groups served during the phase;
q_i = flow rate in lane group i ($i = 1, 2, \ldots, m$) [in vehicles per second (vps)];
$MAH = (q_1MAH_1 + q_2MAH_2 + \ldots + q_mMAH_m)/q$ (sec);
MAH_i = maximum allowable headway for lane group i (sec); and
n = number of arrivals necessary to extend the green to max out.

The MAH calculated above represents the *equivalent* MAH for the general case where one or more lane groups served by a phase have differing MAHs. If all m lane groups have the same MAH_i, the equation for calculating the equivalent MAH simplifies to this common MAH_i. This equivalent MAH must be used in all subsequent evaluation equations.

The flow rate for various lane groups should be based on the demand traversing the group's detected area. For example, the flow rate for a left-turn lane group would equal the left-turn flow rate on that approach. When the length of an exclusive turn lane is less than the length of the detection zone for its adjacent through-movement lane group, the flow in the exclusive lane will also contribute to the flow in the through lane group. In these situations, it is recommended that the flow rate for the through-movement lane group equal the flow rate *entering* its detection zone. This approach is exact when the stop line detection zone is inactive during the green and is conservative when the stop line zone is active.

Equation 6 requires an estimate of the number of arrivals needed to max out the green. This estimate can be obtained by dividing the maximum green duration by the average headway of all vehicles with headways less than the equivalent MAH. The equations for estimating the number of arrivals and the average headway are

$$n = \frac{G_{\max} - MAH - R}{h} \quad (7)$$

$$h = \frac{\int_0^{MAH} tqe^{-qt}dt}{\int_0^{MAH} qe^{-qt}dt} = \frac{\frac{1}{q} - \left(MAH + \frac{1}{q}\right)e^{-qMAH}}{1 - e^{-qMAH}} \quad (8)$$

$$R = (G_q - h_c)(1 - e^{-q_cG_q}) \quad (9)$$

where

$G_{\max}$ = maximum green duration of the subject phase (sec);
h = average headway for all vehicles with headways less than MAH (sec/veh);
G_q = queue clearance time of subject phase ($G_q \geq G_{\min}$) (sec);
R = time between first call on a conflicting phase and queue clearance (sec);
$G_{\min}$ = minimum green duration of subject phase (sec);
q_c = total flow rate in all conflicting phases (vps); and
h_c = average headway between calls from conflicting phases considering only those headways less than G_q *(sec)*.

The value of h_c can be calculated using Equation 8 with G_q substituted for MAH and q_c substituted for q.

Figure 2 illustrates the relationship between the probability of max-out, total traffic demand for the subject phase, equiv-

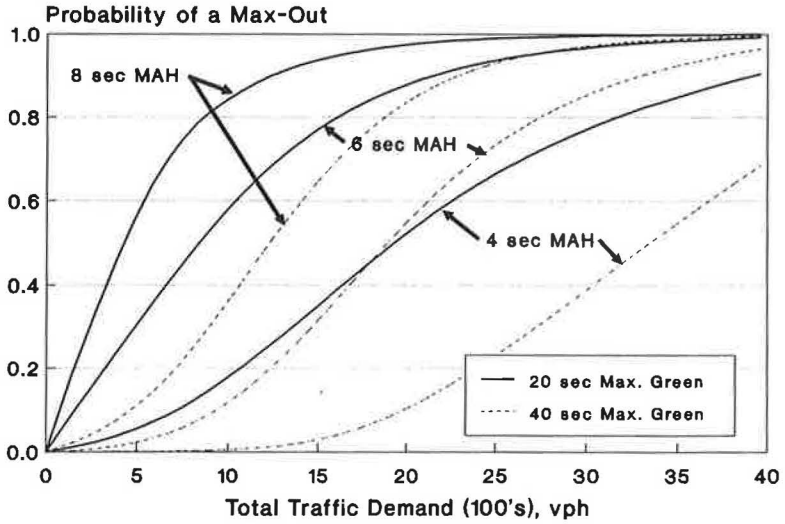


FIGURE 2 Probability of a max-out as a function of traffic demand and maximum green duration.

alent MAH, and maximum green duration. Figure 2 is based on a q_c of 0.14 vps (500 vph) and a G_q of 15 sec. In general, the probability of max-out increases sharply with MAH and traffic demand. Examination of the effect of maximum green indicates a decrease in max-out probability with increasing maximum green duration.

To illustrate the use of Figure 2, assume that the subject phase has a total traffic demand of 1,100 vph and a maximum green duration of 20 sec. If the analyst can formulate a design that yields a 4-sec equivalent MAH, the probability of max-out will be less than 0.2. In other words, 8 of 10 signal cycles will terminate by gap-out and thus indecision-zone protection will be provided about 80 percent of the time.

If the analyst finds that the resulting design yields a MAH of 8 sec, the probability of max-out increases to almost 0.9. This implies that nine of ten cycles will end by max-out and suggests that there may be little to gain by installing an in-decision-zone detection design if it yields an 8-sec MAH.

Waiting Time

The average wait by a traffic queue for a gap-out to occur in a conflicting phase can be determined by using a theoretical approach based on random arrivals to the phase being served. This time can be determined from the following equations:

$$W = (h * N + MAH)p + R \tag{10}$$

$$p = 1 - e^{-qMAH} \tag{11}$$

where

W = average wait by conflicting traffic for a gap-out to occur in the phase being served (sec);
h = average headway for all vehicles with headways less than MAH (sec/veh);
N = average number of extensions of green (i.e., headways < MAH);
p = probability of a headway's being less than the equivalent MAH); and
MAH = maximum allowable headway that will maintain a call for service (sec).

The average number of green extensions before the phase terminates is dependent on the traffic demand in the phase being served, the equivalent MAH, and the maximum green duration for this phase. The average number of extensions can be calculated as

$$N = \frac{\sum_{i=0}^{n-1} ip^i(1 - p) + np^n}{\sum_{i=0}^{n-1} p^i(1 - p) + p^n}$$

$$= \frac{p}{(1 - p)}(1 - p^n) \tag{12}$$

Figure 3 illustrates the relationship among the average waiting time, total traffic demand for the subject phase, equivalent MAH, and maximum green duration. This graph is based on a q_c of 0.14 vps (500 vph) and a G_q of 15 sec. In general, the waiting time increases with MAH and traffic demand. Examination of the effect of maximum green indicates an increase in waiting time with increasing maximum green duration. This trend is the opposite of that for max-out probability, wherein it was noted that larger maximum greens reduced the probability of max-out. In summary, larger maximum greens may improve safety (via less frequent max-outs) but degrade operations (via longer delays).

To illustrate the use of Figure 3, assume that the subject phase has a total traffic demand of 1,100 vph and a maximum green duration of 20 sec. If the analyst can formulate a design that yields a 4-sec equivalent MAH, the average wait for gap-out will be about 14 sec. If the analyst finds that the resulting design yields a MAH of 8 sec, the average wait increases to about 19 sec. Moreover, if the analyst chooses to increase the

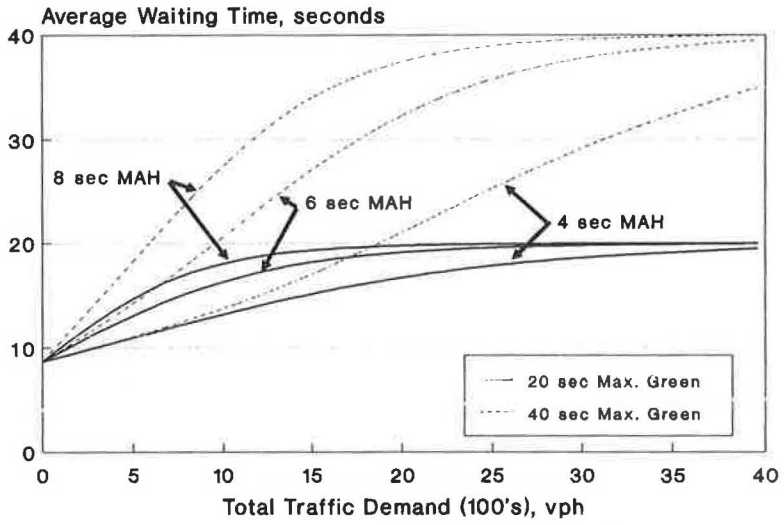


FIGURE 3 Waiting time as a function of traffic demand and maximum green duration.

maximum green duration to 40 sec (to reduce the max-out probability), the corresponding delay will stay at about 14 sec for a 4-sec MAH; however, it will increase to 29 sec for an 8-sec MAH.

Techniques To Reduce MAH

A detector design should offer a balance between safety and efficiency. This balance can generally be achieved by properly locating the detection zones and tuning the PT and CE settings. Designs with advance detectors tend to add additional complexity to the selection of the optimal PT and CE settings. In fact, detection can extend so far back on an approach that it may be impossible to find PT and CE settings that yield both safe and efficient operation. If this situation occurs, three techniques are offered that can help reduce the overall MAH and still provide safe and efficient operation.

One technique for reducing the MAH is to narrow the design speed range. Application of this technique requires a trade-off between the width of the design speed range and the length of the MAH for a detector design. A wider speed range will provide more safety by providing indecision-zone protection for a larger percentage of vehicles; however, it also tends to increase the MAH. Longer MAHs decrease safety because they increase the frequency of max-out and the delay to waiting conflicting traffic. As a minimum, the adjusted design speed range should always include 70 percent or more of the traffic stream (i.e., at least the 15th- to 85th-percentile speed range). The MAH will increase about 20 percent for every 10-mph increase in the design speed range.

A second technique is to adopt a design goal of carrying the last clearing vehicle only through the indecision zone (rather than into the intersection) upon gap-out. This technique was previously described as the Goal 1 design. To achieve maximum efficiency, the stop line detector unit should operate in an EC-DC mode during the green. The combination of a Goal 1 design and an EC-DC stop line detection unit can yield MAHs that are about 30 percent shorter than those from a Goal 2 design.

A third technique for reducing the MAH is to increase the number of advance detectors. In general, the MAH decreases with the number of advance loops provided on the approach. However, the return diminishes rapidly such that there is negligible reduction in MAH for designs with more than three advance loops. A two-advance-loop-detector design will increase the MAH about 1 percent over the three-loop design. A one-advance-loop design will increase the MAH about 8 percent.

There is an added benefit, beyond MAH reduction, from using two or more advance loops in the detector design. Multiple advance loops can provide advance screening of vehicles traveling slower than the design speed range. These vehicles will not be able to extend the green between the first and subsequent loops, and yet a safe gap-out will be possible because these vehicles have not entered the indecision zone.

CONCLUSIONS

The safety and efficiency of a traffic detector design can be determined from the probability of max-out and the time spent waiting for gap-out and subsequent phase change. A detector design that minimizes these measures of effectiveness should provide safe and efficient operation. The performance of a design can be assessed by determining the maximum time separation that it will allow between vehicle calls to the controller (i.e., the maximum allowable headway). For situations where the design serves only a portion of the traffic lanes served by a phase, the analysis must proceed on a phase-specific basis and the performance evaluation would relate to the overall phase operation.

Achieving a detector design with optimal performance characteristics can be difficult because of complex interactions between the design elements (i.e., detector location, detector length, vehicle speed, passage time setting, and call-extension setting). In general, a large MAH will have an adverse effect on performance by increasing the max-out probability and the length of wait for phase change. The methodology described in this paper will allow the designer to determine the optimal combination of design elements in terms of safety (via infrequent max-out) and operations (via a short waiting time for phase change).

REFERENCES

1. Lin, F. B. Optimal Timing Settings and Detector Lengths of Presence Mode Full-Actuated Control. In *Transportation Research Record 1010,* TRB, National Research Council, Washington, D.C., 1985, pp. 37–45.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985, Chapter 9.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

PASSER IV: A Program for Optimizing Signal Timing in Grid Networks

NADEEM A. CHAUDHARY AND CARROLL J. MESSER

The development of PASSER IV, a practical state-of-the-art program for simultaneously optimizing progression bandwidth in multiarterial traffic signal networks, is described. PASSER IV is efficient and is being developed for use on personal computers. A user-friendly mouse-driven graphic interface that provides data entry and file management functions makes the program extremely easy to use. However, the main core of the program, written in FORTRAN 77 using structured programming techniques, is usable on virtually any type of computer. The existing version of PASSER IV determines all four signal timing parameters: cycle length, green split, offset, and phasing sequence. The program optimizes cycle lengths, offsets, and phasing sequences to maximize progression bandwidth. The green splits, however, are determined in a preprocessor using Webster's method. In addition, PASSER IV is capable of minimizing cycle length and can report signal timings for several alternative optimal solutions. Also described is ongoing research to enhance the capabilities of PASSER IV. This research includes optimization of green splits, optimization of two additional main-cross split (circular) phasing sequences, delay calculation procedure, and the capability to generate data files for TRANSYT 7F to facilitate fine-tuning of bandwidth solutions through bandwidth-constrained delay optimization. The final version of PASSER IV will be available in mid-1994.

Optimal traffic signal timing in urban and suburban networks is essential for the full utilization of existing roadways. The objective of signal timing optimization in undersaturated networks is to determine four signal timing parameters, namely, signal cycle lengths, offsets, phasing sequences, and green splits, that optimize (*a*) progression bandwidth, (*b*) a combination of delay and stops, or (*c*) a compromised function based on bandwidth and delay. The existing technology, however, has limitations that do not allow the full achievement of desired objectives.

MAXBAND 86 (*1*,*2*), the only program now generally available for progression bandwidth maximization in multiarterial networks, does not optimize green splits, has a very simplistic traffic model, and is extremely inefficient for practical computations. In addition, it has no capability for reporting traffic measures such as delay, stops, and level of service. TRANSYT 7F (*3*), a program for delay minimization in traffic networks, is the most widely used network signal timing optimization program. TRANSYT 7F, however, is incapable of phasing sequence optimization. Further, its final solution is dependent on the quality of the starting solution, which is not always available.

Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843-3135.

Recent research has shown that concurrent use of MAXBAND 86 and TRANSYT 7F produces signal timings better than those produced by either program alone (*4*,*5*). This approach suggests that the initial starting solution for TRANSYT 7F should be obtained using a bandwidth maximization program and fine-tuned using the bandwidth-constrained delay minimization capability in TRANSYT 7F. However, unlike Arterial Analysis Package (AAP) (*6*) for arterial problems, no program currently exists that provides traffic engineers an automated capability for employing this coordinated approach to multiarterial network optimization problems.

PASSER IV is being developed to overcome many of the above limitations in the existing programs for optimizing signal timing in traffic networks. The focus of this paper is the undersaturated traffic control problem. PASSER IV has evolved from MAXBAND 86 over a period of several years, and all the basic features of MAXBAND 86 have been retained in PASSER IV. However, several enhancements and additional features make the new program easier and more practical to use by traffic engineers. In the following sections, key features of PASSER IV and the current developmental work are described. To begin, the network data sets used for illustrating computational results in the remainder of the paper are described.

DESCRIPTION OF TEST DATA

Thirteen network data sets are used for illustrating computational results described in this paper. Table 1 describes these network problems. The information includes network name and location and the number of arterials, signals, links, and closed loops in the network. More detail and extensive computational experience with these specific problems are described by Chaudhary et al. (*7*).

DESCRIPTION OF PASSER IV PROGRAM

PASSER IV is an advanced network signal timing optimization program. It currently is the only practical personal computer (PC)-based computer program that can optimize signal timings for large multiarterial networks based on maximizing platoon progression. PASSER IV simultaneously maximizes progression bandwidth on all arterials (one-way and two-way) in closed networks such as that shown in Figure 1. PASSER IV explicitly handles one-way streets. It calculates green splits

TABLE 1 Description of Network Problems

NO.	NETWORK NAME	NETWORK GEOMETRY			
		ARTERIALS	SIGNALS	LINKS	LOOPS
1.	University/ Canyon/ 12th/ Street	3	11	11	1
2.	Wisconsin/ Massachusetts/ Garfield	3	15	15	1
3.	Pennsylvania/ Connecticut/ K Street	3	17	17	1
4.	Hawthorne Blvd. mini network, California	5	9	10	2
5.	Walnut Creek Network, California	6	13	15	3
6.	Daytona Beach Network, Florida	7	12	17	6
7.	Post Oak Network, Houston, Texas	8	13	18	6
8.	Ogden Network, Utah	8	13	18	6
9.	Ann Arbor Michigan	8	14	20	7
10.	Los Angeles, California	8	15	21	7
11.	Owosso, Michigan	8	16	18	3
12.	Bay City, Michigan	8	16	20	5
13.	Downtown Memphis Network, Tennessee	8	17	22	6

(from volume and saturation flow data) using Webster's method (*8*) and then optimizes cycle length, offsets, and National Electrical Manufacturers Association (NEMA) phasing sequences with overlap. In addition, PASSER IV allows link-to-link speed variations together with arterial and directional priority options.

PASSER IV is the result of several years of research at Texas Transportation Institute on methods to improve the mathematical model for optimizing progression bandwidth in networks together with the computational efficiency of the underlying mixed-integer linear programs for simultaneously maximizing progression bands on all arterials in the network. The program is being developed with a focus on PC users; however, the core of PASSER IV is adaptable for use on any computer with a FORTRAN compiler. Two PC versions of PASSER IV have been developed and are being enhanced. The standard PC version can be used with any IBM-compatible PC with 640K of random access memory (RAM) and can handle networks having up to 20 arterials and 35 intersections. The advanced PC version is designed for use on 80486 and 80386 (with math coprocessor) based PCs with at least 8 megabytes of RAM. This version can handle larger networks with up to 50 intersections. The advanced version is also twice as fast as the standard PC version. In the following sections, key features of PASSER IV, additional options currently being implemented in PASSER IV, and future plans for enhancing the program are described. The final version of PASSER IV with all these features, PASSER IV-94, will be ready for distribution in mid-1994.

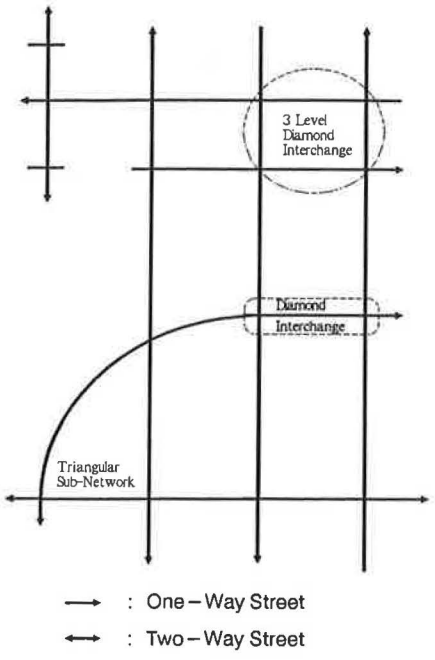


FIGURE 1 Example network with possible subcomponents.

KEY FEATURES OF PASSER IV

Graphic User Interface

PASSER IV's menu-driven graphic user interface (GUI), with pull-down menus and mouse support, makes the program extremely easy to use. Data are entered arterial by arterial until the total network is described. Arterial data can be entered in any order. However, data for intersections on an arterial must be entered in sequential order. This format is slightly restrictive as compared with other programs, but it reduces the linkage data coding requirements to only link distances and travel speeds. The other linkage information is automatically obtained by the program. The program requires that each intersection be assigned a unique (node) identification number. This allows the program to determine the network structure. In addition, this scheme permits the data for a signal (which falls on two intersecting arteries) to be entered only once. Figures 2 and 3 shown two video screens of the GUI.

Computational Efficiency

In the past, some researchers have speculated that MAXBAND's optimization routine MPCODE (*9*) was inef-

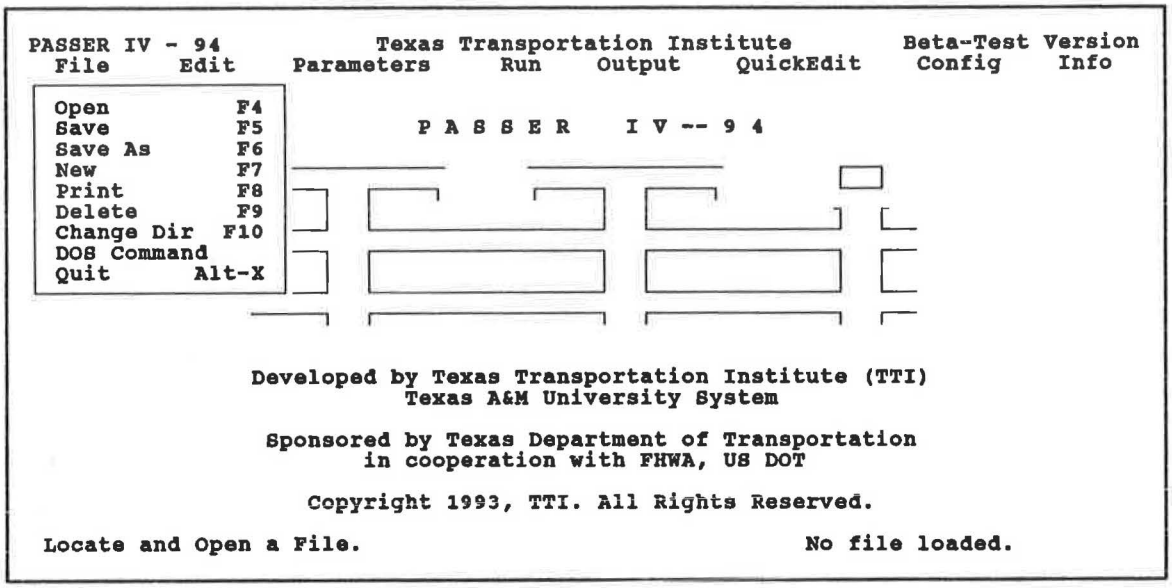


FIGURE 2 Pull-down menus.

ficient and recommended that it be replaced by a more efficient routine (*10,11*). Experimental work by Chaudhary et al. and comparison of MPCODE with LINDO (*12*) (a fairly recent efficient optimization package), however, demonstrated that MPCODE is as efficient as LINDO for signal timing optimization problems (*13*). Chaudhary et al. further concluded that the underlying mixed-integer linear programming (MILP) problem formulations for signal synchronization in networks are inherently difficult, requiring the need to develop efficient heuristic optimization procedures. Therefore, MPCODE is retained in PASSER IV and the optimization efficiency of PASSER IV is increased by implementing the following techniques:

1. The simultaneous optimization method of MAXBAND 86 has been retained in PASSER IV. In addition, two heuristic optimization techniques developed by Chaudhary et al. (*7,14*) have been implemented in PASSER IV. The two-step heuristic method is 10 times faster than the simultaneous optimization (SO) of all variables and produces the same results as the SO method. The three-step method is up to 99 percent faster than the SO approach, but it does not guarantee the absolute maximum bandwidth, although it produces the best possible solution in many cases. For large network problems, however, the three-step method seems to be the only feasible approach from a practical point of view.

2. In PASSER IV, one-way arterials are explicitly modeled as compared with the approach used in MAXBAND 86. This reduces the MILP size and computational complexity of network problems with one-way arterials. As a consequence, the central processing unit (CPU) time is reduced and wider bands are produced (*7*).

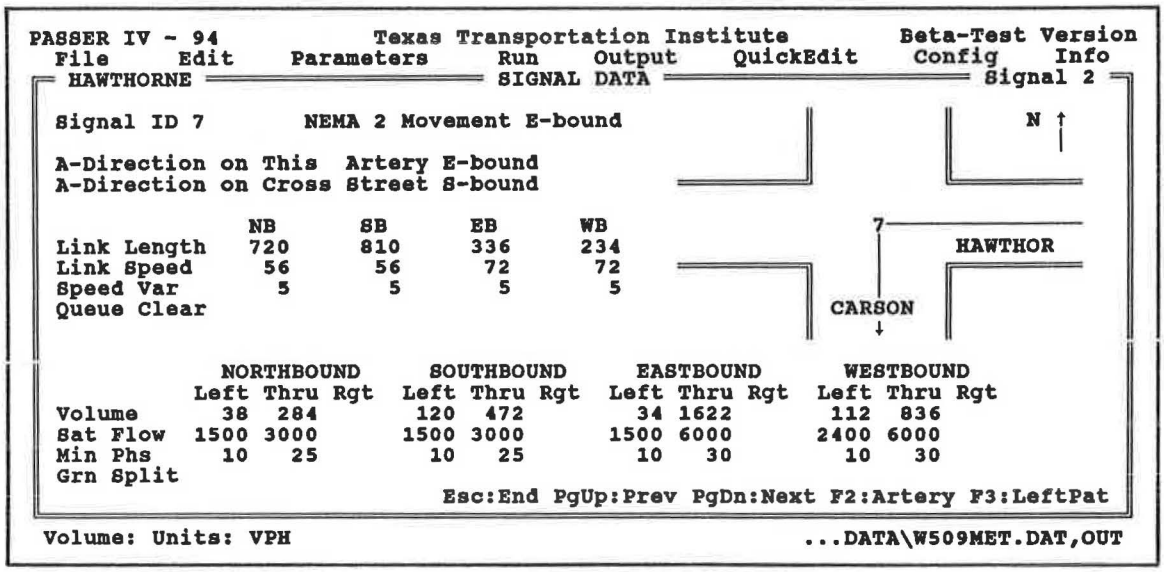


FIGURE 3 Signal data entry screen.

3. Chaudhary et al. (*7*) demonstrated that the use of tighter bounds for link synchronization variables significantly enhanced the computational efficiency of MILPs for progression bandwidth optimization. However, this approach is not practical since the results were based on the usage of bounds obtained through observation of only a few test problems. Recently, a formal, data-specific scheme for calculating tighter bounds for these variables was developed by Chaudhary et al. (*14*). This scheme has been implemented in PASSER IV. The use of tighter bounds reduces the search region. In addition, integer variables having the same lower and upper bounds are eliminated from the MILP. As a consequence, PASSER IV produces solutions much faster than MAXBAND 86.

Table 2 provides a summary of optimization results for the test problems using the advanced version of PASSER IV on a 80486-based PC. The three-step heuristic method was selected for optimizing all the test network problems. Information given in Table 2 includes total bandwidth as a fraction of cycle length (entries in parentheses give the best possible total bandwidth as a fraction of cycle length using simultaneous optimization), average total arterial bandwidth obtained by dividing numbers in the previous column by the number of arterials in the network, and the CPU run time in seconds required on the PC. Observations of the results are summarized as follows:

1. Except for the second problem, all total bandwidths obtained were within 95 percent of the best possible bandwidths. Further, total bandwidths for eight problems were within 99 percent of the best possible bandwidths. These results demonstrate that the three-step method provides good (sometimes the best) solutions for network problems.
2. None of the problems required more than 8 min of CPU time for optimization. In contrast, the same problems required several (sometimes up to 10) hours of CPU time when optimized using MAXBAND 86 (*7*).

In summary, the three-step optimization capability in PASSER IV makes the program feasible for use even on a PC. Given the fact that the traffic data used in the optimization program are never 100 percent accurate, this heuristic strategy is more than sufficient for practical purposes. However, for those users who wish to obtain absolutely the best solutions, PASSER IV is equipped with two-step and simultaneous optimization capabilities.

Minimization of Cycle Length

Often a signal timing optimization problem has multiple optimal solutions with the same bandwidth efficiency but different cycle lengths. PASSER IV has an optional capability to select the solution with the lowest cycle length. The user can activate this capability by setting the cycle length optimization switch and specifying the weight to be given to cycle length optimization. The higher the weight, the better the chance of finding a solution having a lower cycle length. However, care should be taken because too high a weight may result in a nonoptimal bandwidth solution.

To illustrate the fact that same best bandwidth (as a fraction of cycle length) may result at various cycle lengths, an actual arterial (12th Street) is used (Figure 4). Two volume conditions, a.m. peak and off peak, were examined. Multiple bandwidth solutions using the simultaneous optimization method for these problems were obtained and analyzed using TRANSYT 7F. Tables 3 and 4 summarize the results. The following is a discussion of the results:

1. For the a.m. peak case, four alternative optimal solutions having the best bandwidth of 0.37564 (fraction of cycle length) were found. For the off-peak volume conditions, two solutions with the best total bandwidth equal to 0.4223 (fraction of cycle length) were found.
2. For the a.m. peak condition, neither the lowest nor the highest cycle length resulted in the least delay. In fact, the

TABLE 2 Summary of PASSER IV Runs

NO.	NETWORK NAME	Total Bandwidth Efficiency	Average Arterial Bandwidth Efficiency	CPU Run Time (seconds)
1.	University/ Canyon/ 12th/ Street	1.281(1.304)	.427	10
2.	Wisconsin/ Massachusetts/ Garfield	1.182(1.371)	.394	27
3.	Pennsylvania/ Connecticut/ K Street	1.051(1.051)	.350	36
4.	Hawthorne Blvd. mini network, California	3.996(3.996)	.799	14
5.	Walnut Creek Network, California	2.770(2.771)	.462	84
6.	Daytona Beach Network, Florida	2.910(2.911)	.416	403
7.	Post Oak Network, Houston, Texas	2.715(2.816)	.339	161
8.	Ogden Network, Utah	3.099(3.156)	.387	186
9.	Ann Arbor Michigan	3.868(3.869)	.484	202
10.	Los Angeles, California	3.609(3.609)	.451	479
11.	Owosso, Michigan	4.196(4.196)	.525	137
12.	Bay City, Michigan	3.576(3.732)	.447	147
13.	Downtown Memphis Network, Tennessee	3.408(3.418)	.426	131

Note: Entries in parentheses give the best possible total bandwidth as a fraction of cycle length using simultaneous optimization.

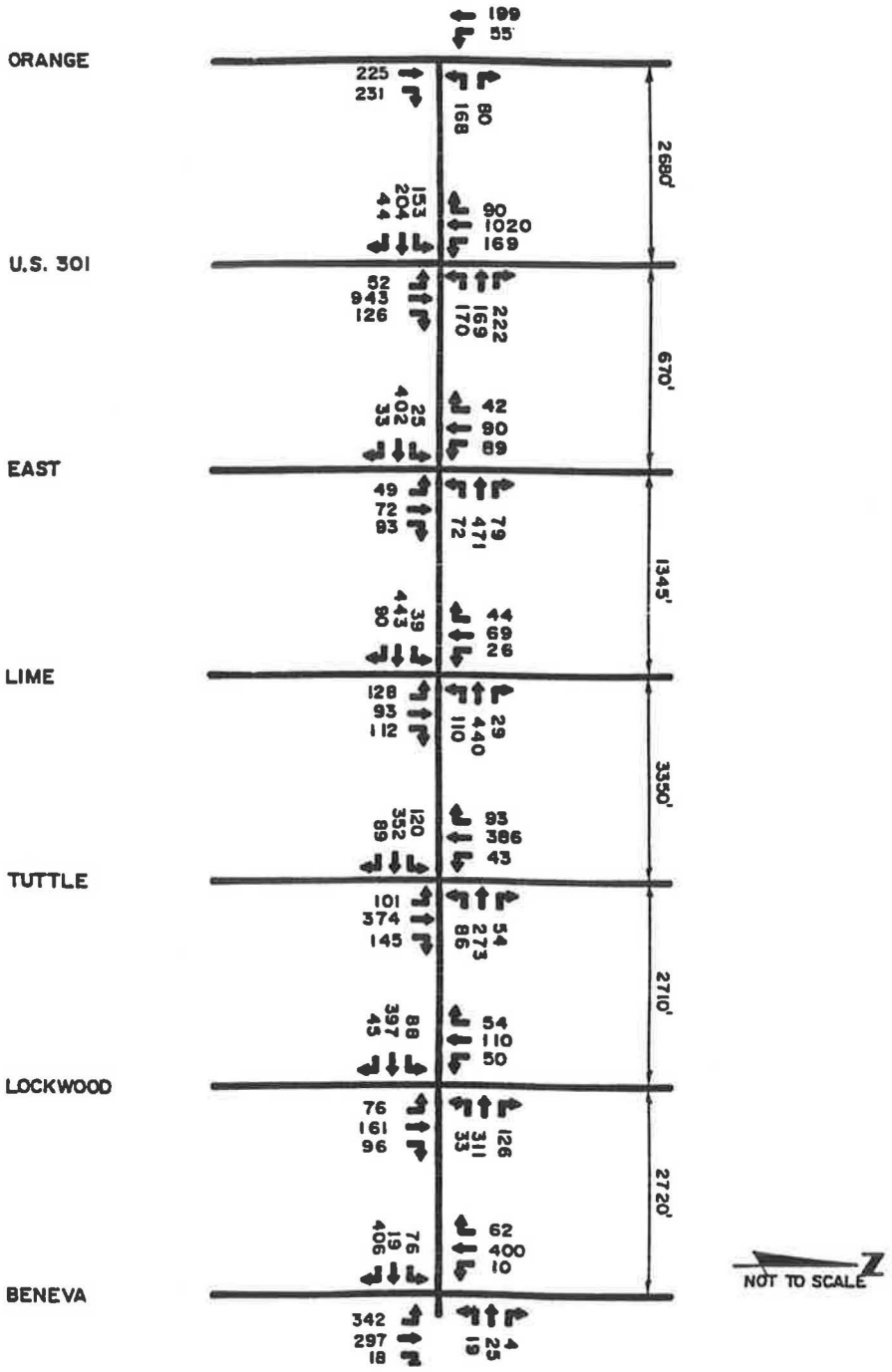


FIGURE 4 Off-peak turning movements for 12th Street.

solution with a cycle length of 79 sec (19 sec more than the lowest cycle length) had the lowest TRANSYT performance index (PI). Similarly, for the off-peak period, the solution with the lowest cycle length had a higher (delay and stops) TRANSYT PI.

In summary, it may be seen that the lowest cycle length solution is not necessarily the minimum delay solution. The cycle length range chosen for given traffic conditions may be the biggest factor affecting delay. However, it is not the only factor, since the authors' experience with arterial problems has shown that even two alternative solutions with the same cycle length may exhibit significantly different delay measures. More research is needed to understand the effects of other signal timing variables on delay. Despite the need for more research, the cycle length minimization capability in PASSER IV is a useful tool that can enable the generation of alternative solutions for the same problem.

Output Reports for Multiple Solutions

As mentioned earlier (and demonstrated above for the arterial cases), multiple optimal or very good suboptimal solutions may exist for a network problem. These solutions may have

TABLE 3 TRANSYT Delay Comparison of Alternative Solutions for 12th Street: A.M. Peak Case

	Total Delay (veh-hr /hr)	Average Delay (sec/ veh)	Number of Stops Per Trip	Average Speed (mph)	Transyt. P.I.	Cycle Length (Sec)
Solu 1	222	53.9	12125	14.6	229.2	60
Solu 2	181	44.0	12102	17.2	199.3	71
Solu 3	168	40.7	12050	16.6	189.1	79
Solu 4	207	50.4	11915	14.9	217.5	88

Note: All solutions have the same optimum bandwidth, 0.37564.

significantly different estimates of delay and stops. In maximizing progression bandwidth, a traffic engineer would probably want to select an alternative solution (if more than one solution is available) that results in the lowest delay and stops. PASSER IV has been equipped to allow printing of signal timing reports for a specified number of multiple solutions. The maximum number of solutions that can be printed is 5 and 10 for the standard and advanced PC versions, respectively.

A delay analysis was performed of the six best bandwidth solutions for each of the test networks. Since PASSER IV does not currently have the capability to estimate traffic performance measures such as delay, stops, and fuel consumption, TRANSYT 7F was used to evaluate each solution on the basis of PI (a linear combination of stops and delay). In addition, for each alternative solution, TRANSYT 7F was used to perform a bandwidth-constrained delay minimization with the option to minimize fuel consumption. TRANSYT results showed that bandwidth-constrained delay optimization (BCDO) further reduces delay. In order to perform an unbiased comparison, five replications of microscopic simulation using TRAF NETSIM (*15*) were performed to analyze each PASSER IV and TRANSYT 7F solution. It was discovered that NETSIM results do not always match those of TRANSYT. For many cases, TRANSYT BCDO results were worse than those of PASSER IV. A surprising finding was that for many cases, PASSER IV solutions had lower fuel consumption than TRANSYT BCDO solutions, even when TRANSYT solutions had lower delay and stop estimates. More research is needed to pinpoint and correct this discrepancy between TRANSYT and NETSIM. Also, the amount of work involved to code data and to perform all TRANSYT and NETSIM computer runs warrants that an automated capability be developed for such work. This issue is discussed later.

PASSER IV Output

PASSER IV prints an extensive signal timing solution report. The solution report includes a section summarizing the input data and calculations done in the preprocessor, a section showing the performance at each step during optimization, signal timings for each arterial in the network, time-space diagrams, and a section giving timings for each signal in the network. PASSER IV GUI allows the user to view or print the complete output or selected portions of the solution report. Figure 5 shows a network solution being viewed through the GUI. Figure 6 shows a time-space diagram and a signal setting table.

Data Requirements

PASSER IV requires the following network data: a unique identification number for each signal, link lengths, saturation flow rates, traffic volumes, average travel speeds on links, and cycle length range. The green splits are calculated internally from volume and saturation flow data using Webster's equation. Optionally, the user can directly specify green splits, which are used by the program without modification. However, the data input is relatively simple as compared with other network optimization/simulation programs.

TABLE 4 TRANSYT Delay Comparison of Alternative Solutions for 12th Street: Off-Peak Case

	Total Delay (veh-hr /hr)	Average Delay (sec/ veh)	Number of Stops Per Trip	Average Speed (mph)	Transyt. P.I.	Cycle Length (Sec)
Solu 1	97	28.1	9780	19.3	124.6	60
Solu 2	90	26.2	9667	19.6	119.2	70

Note: Both solutions have optimum bandwidth, 0.4223.

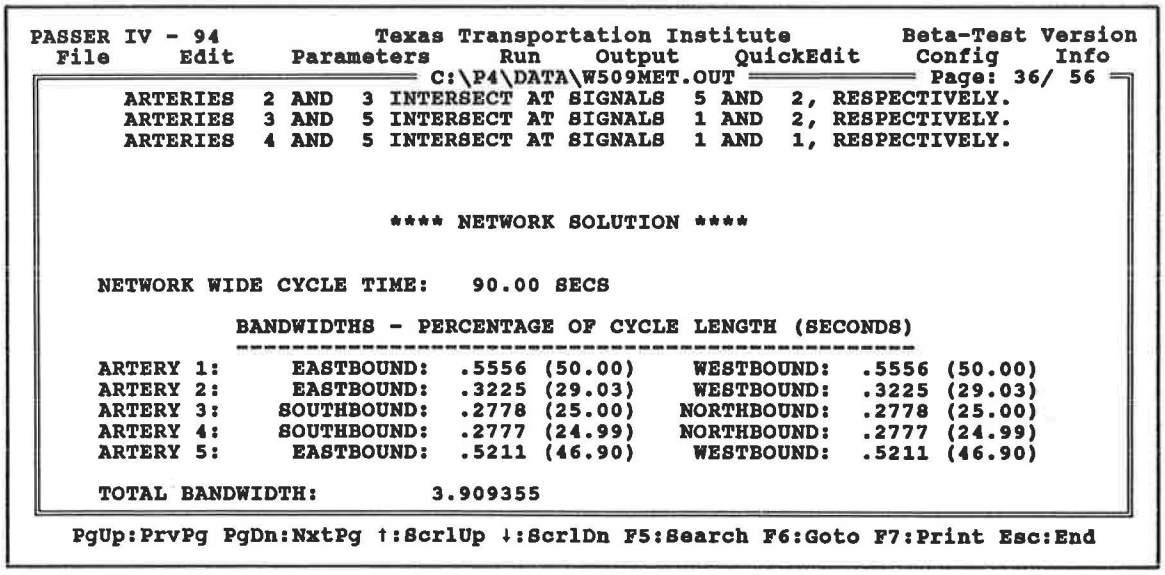


FIGURE 5 Section of output from PASSER IV GUI.

Computer Hardware Requirements

PASSER IV is developed specifically for IBM PCs and compatible computers. However, the core of the program, its signal timing optimization program, can be compiled and used on virtually any machine that has a FORTRAN compiler available. This capability will permit the program to be used for large urban networks controlled by existing traffic management systems and future intelligent vehicle-highway (IVHS) systems.

ADDITIONAL FEATURES TO BE IMPLEMENTED

Research is currently under way to add more options to PASSER IV. The new options will further enhance the program and provide better signal timing solutions. Some of these enhancements are described in detail in the following sections.

Green Time Optimization

Recently, Chaudhary et al. developed the necessary mathematics to allow the simultaneous optimization of cycle length, offsets, signal phasing sequences, and green splits for arterial problems (*16*). The enhanced arterial formulation produced wider progression bands as compared with those produced by all the existing programs (including PASSER IV) for arterial bandwidth optimization. These enhancements were applied to some network optimization problems. The results of the enhanced network formulation showed significant improvement in total bandwidth. However, from a practical viewpoint, the increase in MILP formulation size and an exponential increase in the CPU time make this formulation impractical at the present time. Therefore, it has been decided to implement the following alternative approach in PASSER IV:

Step 1. Obtain optimal signal timing solutions using PASSER IV as before.

Step 2. Fix all integer variables from Step 1 in the enhanced formulation. Optimize progression bandwidth. Since this results in a simple linear program, it can be easily optimized in a few seconds.

This approach was used on a subset of five network test problems. Step 1 problems were optimized using the three-step heuristic method. In these optimization runs, all the minor cross streets had volume-to-capacity ratios less than 0.95. Table 5 shows a summary of computational results and provides a comparison of these results with those given in Table 2. The results show significant improvement in total bandwidth for each problem. These improvements range from 18.00 percent to 125.29 percent with an average improvement of 73.79 percent. It should be noted that these improvements were achieved with an insignificant number of additional calculations requiring only a few seconds.

Concurrent Bandwidth and Delay Optimization

A considerable amount of research by Cohen and Liu (*4,5,17*) has demonstrated that concurrent use of bandwidth and delay programs can produce better signal timings than either program alone. However, unlike the AAP for arterial problems, no computer package exists that provides for the automated use of this methodology for networks. This gap is being filled by adding a NAP (Network Analysis Package) option in PASSER IV. This option allows utilization of PASSER IV input data and optimal PASSER IV signal timing solution to generate a TRANSYT 7F input data file for bandwidth-constrained optimization. The implementation of this option in PASSER IV is complete and is being tested.

Delay Estimation Routine

The existing PASSER IV package has no capability to estimate delay at intersections. Such a capability is needed to

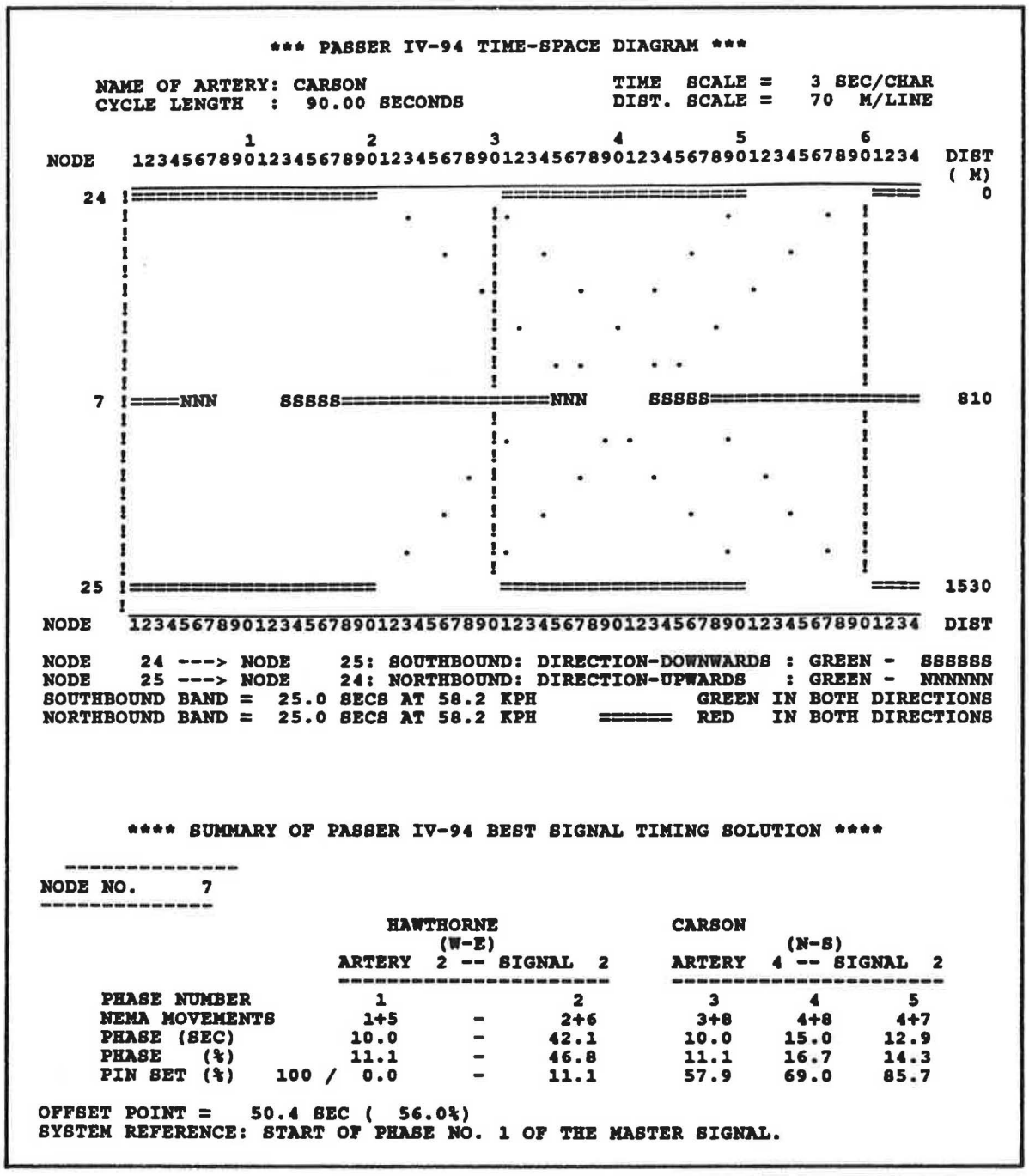


FIGURE 6 PASSER IV time-space diagram and signal timing table.

select the alternative optimal solution for implementation or further analysis. A delay estimation routine is currently being added in the program. The delay estimation approach described by Malakapalli and Messer in another paper in this Record is being used.

FUTURE PASSER IV ENHANCEMENTS

The following enhancements are scheduled to be implemented before the release of PASSER IV. Major research related to each of the following enhancements has been conducted; however, the implementation phase for these options is not as advanced as that for the options described in the previous sections.

Combined NEMA and Circular Phasing Optimization

Chaudhary et al. developed a scheme to simultaneously optimize NEMA and circular phasing sequences and produced an arterial optimization program called MAXBAND 89T that provided for an automated use of this capability (*18*). Circular phasing is a subset of main-cross split phasing with four phases

TABLE 5 Results from Green Split Optimization

NO.	NETWORK NAME	PASSER IV Bandwidth	Enhanced Bandwidth	% increase in Bandwidth	Additional LP iteration
1.	University/ Canyon/ 12th/ Street	1.281	2.886	125.29	207
4.	Hawthorne Blvd. mini network, California	3.996	4.728	18	281
6.	Daytona Beach Network, Florida	2.910	4.360	49.83	541
11.	Owosso, Michigan	4.196	6.609	57.51	459
13.	Downtown Memphis Network, Tennessee	3.408	7.441	118.34	689

(main lead, cross lead, main lag, and cross lag) at an intersection. This phasing is applicable in some special cases. Combined NEMA and circular phasing has been shown to provide larger progression bands. Once extensive testing is completed and successful, the circular phasing sequence optimization capability will be incorporated into PASSER IV.

Multiband Maximization

Traditional bandwidth optimization programs maximize uniform progression bands along the arterials. Gartner et al. (*19*) recently developed an arterial signal timing program, MULTIBAND, that maximizes volume-weighted bands for each link. In MULTIBAND, the center of the progression band on each link coincides with a line that goes through all intersections on the arterial. Gartner's research showed that MULTIBAND produces solutions with less delay then did its parent, MAXBAND. The authors have applied combined multiband and green-split optimization to a subset of multiarterial test networks. Although this feature results in a significant increase in the optimization problem size, the initial results look promising. All problems were formulated manually and solved using the two-step and three-step heuristic methods. The resulting sum of bands (objective function) was much larger than that for the uniform bandwidth cases. A direct comparison of the two cases for each network problem is not possible without further analysis of traffic performance measures. This research, however, remains to be done and will be facilitated by the NAP option in PASSER IV.

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

PASSER IV is now a practical program for optimizing progression bandwidth-based signal timings for arterial as well as multiarterial closed-loop networks. To the authors' knowledge, it is the only network program of its type that is available to traffic engineers for use on a PC. It is envisioned that the program will be extremely useful for solving many problems that are experienced by cities in the United States and in many other parts of the world. Nevertheless, PASSER IV has room for further enhancements, some of which are listed in the next section.

Recommendations

The following are possible enhancements that can increase PASSER IV utility:

1. Bus route optimization as a secondary objective.
2. Explicit optimization of NEMA phasing sequences without overlap.
3. Explicit protected/permitted phasing optimization.
4. Special phasing sequence optimization, such as starting a phase twice within a cycle (sometimes called conditional service).
5. Double cycling some intersections.
6. Capability to assess advantages of removing a traffic signal.
7. Capability to run multiple jobs using the BATCH mode, given only new demand data (this would make the program more suitable for use as a submodule in traffic management systems for urban networks).
8. Fine-tuning existing signal timings for changed traffic conditions without resolving a new MILP from scratch.
9. Capability to integrate signal optimization programs such that one can fix offsets and signal timings for a subnetwork and optimize the remaining network [i.e., diamond interchanges within a network could be optimized using PASSER III (*20*)].
10. Currently, PASSER IV is only applicable to undersaturated networks; however, situations where some subnetworks are oversaturated are not uncommon. Extension of PASSER IV mathematical formulation to address such networks would greatly enhance the program's utility. One such approach could be the combination of internal metering principles (*21*) and bandwidth optimization.

ACKNOWLEDGMENTS

The research on which this paper is based was funded by the Texas Higher Education Coordinating Board under their Energy Research in Applications Programs and by the Texas Department of Transportation under the Highway Planning and Research Program. Their support is greatly appreciated. Herman E. Haenel of Advanced Traffic Engineering, Inc., before retiring from Texas DOT, provided initial support for development of PASSER IV and supplied useful ideas. B. Ray Derr, the current Texas DOT representative on this research, continued to provide excellent technical support.

Stephen L. Cohen of FHWA provided the network data sets. To all, the authors are extremely grateful.

The authors would also like to acknowledge contributions made to PASSER IV by the following graduate students: Anulark Pinnoi implemented and tested the two-step and three-step heuristic methods and main-cross split phasing optimization option. Tai-Hsi Wu examined the nonuniform bandwidth models and green time optimization models and implemented the NAP option in PASSER IV. Rajivendra Nath developed the data structures and the first version of the PASSER IV GUI. Bogju Lee is performing ongoing enhancements to the GUI.

REFERENCES

1. Chang, E. C., S. L. Cohen, C. Liu, C. J. Messer, and N. A. Chaudhary. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequences in Multiarterial Closed Networks. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1989, pp. 61–67.
2. Messer, C. J., G. L. Hogg, N. A. Chaudhary, and E. C. P. Chang. *Optimization of Left-Turn Phase Sequence in Signalized Networks Using MAXBAND 86, Volume 1: Summary Report*. Report FHWA/RD-84/082. FHWA, U.S. Department of Transportation, Jan. 1986.
3. Wallace, C. E., K. G. Courage, D. P. Reaves, G. W. Schoene, G. W. Euler, and A. Wilbur. *TRANSYT 7F Users' Manual*. FHWA, U.S. Department of Transportation, 1988.
4. Cohen, S. L., and C. C. Liu. The Bandwidth-Constrained TRANSYT Signal-Optimization Program. In *Transportation Research Record 1057*, TRB, National Research Council, Washington, D.C., 1986, pp. 1–7.
5. Liu, C. C. Bandwidth-Constrained Delay Optimization for Signal Systems. *ITE Journal*, Dec. 1988, pp. 21–26.
6. Courage, K. G., and C. E. Wallace. *Arterial Analysis Package User's Guide*. FHWA, U.S. Department of Transportation, Dec. 1990.
7. Chaudhary, N. A., A. Pinnoi, and C. J. Messer. Proposed Enhancements to MAXBAND 86 Program. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991, pp. 98–104.
8. Webster, F. V., and B. M. Cobbe. *Traffic Signals*. Technical Paper 56. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1966.
9. Land, A., and S. Powell. *Fortran Codes for Mathematical Programming*. John Wiley & Sons, Ltd., London, 1973.
10. Chaudhary, N. A. *A Mixed Integer Linear Programming Approach for Obtaining an Optimal Signal Timing Plan in General Traffic Networks*. Ph.D. dissertation. Texas A&M University, College Station, Aug. 1987.
11. Cohen, S. L. *Optimization of Left Turn Phase Sequence in Signalized Closed Networks*. Report FHWA/RD-89/157. FHWA, U.S. Department of Transportation, May 1988.
12. Schrage, L. *Users' Manual for LINDO*, 3rd ed. The Scientific Press, Redwood City, Calif., 1987.
13. Chaudhary, N. A., C. J. Messer, and A. Pinnoi. Efficiency of Mixed Integer Linear Programs for Traffic Signal Synchronization Problems. *Proc., 25th Annual SE TIMS Meeting*, Oct. 1989, pp. 155–257.
14. Chaudhary, N. A., and A. Pinnoi. Mixed Integer Linear Programs for Optimizing Signal Settings in Traffic Networks—Characteristics and Computational Efficiency. Presented at TIMS/ORSA meeting, Orlando, Fla., April 1992 (to appear in *Applications of Management Science: Network Optimization Applications*, Vol. 9).
15. *TRAF NETSIM Users' Manual*. FHWA, U.S. Department of Transportation, 1988.
16. Chaudhary, N. A., and T. Wu. Mixed Integer Programs for Traffic Signal Coordination: Linear Transformation of Non-linear Constraints. Presented at TIMS/ORSA meeting, Orlando, Fla., April 1992 (to appear in *Applications of Management Science: Network Optimization Applications*, Vol. 9).
17. Cohen, S. L. Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 81–84.
18. Chaudhary, N. A., A. Pinnoi, and C. J. Messer. Scheme to Optimize Circular Phasing Sequences. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991, pp. 72–82.
19. Gartner, N. H., S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND—A Variable-Bandwidth Arterial Progression Scheme. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
20. Fambro, D. B., N. A. Chaudhary, J. Bonneson, C. J. Messer, and L. Arabie. *PASSER III-90: Users' Manual and Applications Guide*. Texas Transportation Institute, Texas A&M University, College Station, March 1991.
21. *NCHRP Report 3-38(4): Internal Metering Policy for Oversaturated Networks*, Vols. 1 and 2. TRB, National Research Council, Washington, D.C., June 1992.

DISCUSSION

NATHAN H. GARTNER AND JOHN D. C. LITTLE
Department of Civil Engineering, University of Massachusetts, Lowell, Mass. 01854; Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

PASSER IV is a new name given by Chaudhary and Messer to a bandwidth optimization model that has existed for a long time under the name MAXBAND. This raises a number of important questions beyond the technical details of their paper.

The name MAXBAND was coined by Little et al. (*1*) in 1981 for a mixed-integer linear programming (MILP) formulation and computer program for the bandwidth maximization problem in arteries and triangular networks. The development of this program was supported by FHWA and was based on earlier work by Little (*2*) and subsequent refinements. Chang et al. (*3*) then extended the MAXBAND model, also under FHWA sponsorship, to grid network optimization. The new version was dubbed MAXBAND-86. In a 1991 paper, Chaudhary et al. (*4*) showed that the application of various heuristic techniques to the MILP optimization process in MAXBAND can lead to substantial reductions in execution time and can make it feasible to run this model on personal computers. Other researchers have also proposed a variety of enhancements to the MAXBAND optimization process, and this continues to be an area of intensive research (*5,6*). We are pleased to see the development of improved solution strategies, which are likely to lead to significant reductions in running times for the larger network bandwidth optimization problems and to make the program more accessible and usable for practicing traffic engineers.

In their 1991 paper (*4*), Chaudhary et al. state:

> MAXBAND-86 is the only operational traffic signal program that allows progression bandwidth optimization in multiarterial closed-loop networks. The program formulates the problem as a mixed integer linear program and is capable of optimizing network-wide cycle length, signal offsets, and signal phasing sequences.

Now PASSER IV claims identical capabilities. We believe that renaming the enhanced version of MAXBAND-86 as

PASSER IV is inappropriate on several grounds. First, it implies that this program is a continuation of the well-known PASSER family, which it is not—other programs in this family do not use the MILP model and optimization. Second, the introduction of a well-known model (MAXBAND) under a different guise is likely to lead to misunderstanding and general confusion in the traffic profession. Third, the introduction of a new name for a model that is well known by a different name obfuscates the origins and the intellectual ownership of said model.

We believe that it would be desirable for the authors to find another name for their program that more adequately reflects the source of the model and the share of their contribution to its development.

REFERENCES

1. Little, J. D. C., M. D. Kelson, and N. H. Gartner. MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks. In *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981.
2. Little, J. D. C. Synchronization of Traffic Signals by Mixed-Integer Linear Programming. *Operations Research*, Vol. 14, 1966.
3. Chang, E. C., S. L. Cohen, C. Liu, C. J. Messer, and N. A. Chaudhary. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequences in Multiarterial Closed Networks. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1989.
4. Chaudhary, N. A., A. Pinnoi, and C. J. Messer. Proposed Enhancements to MAXBAND 86 Program. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991.
5. Mireault, P. Solving the Single Artery Traffic Signal Synchronization with Benders Decomposition. Presented at Joint National Meeting of CORS/ORSA/TIMS, Vancouver, Canada, May 1990.
6. Solanki, R., R. S. Pillai, and A. K. Rathi. *A Fast Heuristic for Maximizing Bandwidth in Traffic Networks*. Oak Ridge National Laboratory, Oak Ridge, Tenn., May 1993.

AUTHORS' CLOSURE

In the discussion of our paper describing PASSER IV, Gartner and Little question the use of the name PASSER IV for our computer software. In support of their logic, they have cited only selected references. We present here more representative citations that negate their claim.

Little developed mixed-integer linear programming (MILP) formulation for arterial bandwidth optimization in 1966 (*1*). In 1980, under an FHWA contract, Little and Kelson extended the basic formulation and developed MAXBAND, a program for optimizing arterial and triangular network problems (*2*). The discussants state that the name MAXBAND was coined for both the mathematical program and the computer software; however, the MAXBAND Summary Report (*2*) clearly indicates that this name was given only to the computer program, which is the property of FHWA. In addition, MAXBAND was hard-wired to handle restricted networks composed of only three arterials in a triangular configuration. The MILP optimization module used in MAXBAND is composed of a set of routines developed by others and available to the general public (*3*).

In a subsequent FHWA contract, Messer et al. at Texas Transportation Institute (TTI) developed MAXBAND-86 by enhancing MAXBAND to handle general grid networks with up to 20 arterials and 50 signals (*4*). Chang et al. (of TTI) later described MAXBAND-86 (*5*). Some of the specific enhancements included a revision of the data structures, modifications to the input data stream, incorporation of a general loop generation algorithm, additional output to provide phase interval setting for each signal, and more. Although the program retained the basic arterial mathematical formulation developed by Little (*1*), the computer program was substantially upgraded by TTI to formulate the MILP for general multiarterial closed-loop network problems. However, to our knowledge, the MAXBAND-86 program never became widely accepted among the traffic engineering community because of its computational inefficiency and dependence on mainframe computers. It has, however, been used by selected groups of researchers.

A number of researchers have developed other modifications to the basic arterial mathematical formulation as well as to the MAXBAND program (*6–11*). However, by referring only to the work of Mireault and Solanki (*6,7*), the discussants give the impression that none of the other researchers but us have given new names to the programs they produced as a result of their enhancements to the basic arterial formulation. Some of the known cases that indicate otherwise are as follows:

- Gartner et al. (*8*) developed an enhancement to MAXBAND and called the resulting program MULTIBAND,
- Tsay and Lin (*9*) modified MAXBAND and gave it the name BANDTOP, and
- Khatib (*10*) called his modified version the ZMODEL.

Thus, using a new name for the resulting software product based on mathematical programming is not unusual. In fact, one of the discussants has himself coined a new name recently.

As for the use of the name PASSER IV, we would like to point out that PASSER IV is not a new name. In fact, a TTI research team, originally led by Messer, began in September 1979 to develop a software package by this name and with the same applications in mind. This research was funded by the Texas State Department of Highways and Public Transportation in cooperation with FHWA. The first version of PASSER IV was released in 1984 (*12*). The PASSER series of programs is extremely popular among the traffic engineering community because of its computational efficiency and ease of use. Our continued use of the name PASSER IV reflects our continued commitment to enhancing our progression optimization programs.

The PASSER IV-94 version, which is about to be released, draws on all the work that we have performed since 1979, including the original PASSER IV research program. It has several features not available in MAXBAND-86, which TTI developed for FHWA. These include

- A user-friendly interface for PCs,
- Enhanced green split calculation routine,
- Ability to explicitly model one-way arterials,
- Efficient heuristic optimization procedures,
- Ability to run on PCs with 640K of RAM,
- Ability to minimize cycle length,
- Ability to estimate traffic delay and other measures of effectiveness,

- Ability to generate multiple solutions, and
- Ability to generate TRANSYT-7F input data files and to run TRANSYT-7F from the main menu.

Last, we would like to point out that we have consistently acknowledged the fact that the basic arterial formulation for optimizing arterial problems used in PASSER IV-94 is due to Little. His work has been fully acknowledged in all of our related technical papers. It is, however, only a small portion of the overall program's operation. We certainly do not believe users will be confused between FHWA's program (using the title MAXBAND) and our software (using the trademark PASSER).

REFERENCES

1. Little, J. D. C. The Synchronization of Traffic Signals by Mixed-Integer Linear Programming. *Operations Research*, Vol. 14, 1966.
2. Little, J. D. C., and M. D. Kelson. *Optimal Signal Timing for Arterial Signal Systems*, Vol. 1. Summary Report. Report FHWA/RD-80/082. Dec. 1980.
3. Land, A. H., and S. Powell. *FORTRAN Codes for Mathematical Programming*. John Wiley & Sons, Ltd., London, 1973.
4. Messer, C. J., G. L. Hogg, E. C. Chang, and N. A. Chaudhary. *Optimization of Left Turn Phase Sequences in Signalized Networks Using MAXBAND 86*. Final Report, Vols. 1–3. Report FHWA/RD-84. Jan. 1986.
5. Chang, E. C., S. L. Cohen, C. Liu, C. J. Messer, and N. A. Chaudhary. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequences in Multiarterial Closed Networks. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1989.
6. Mireault, P. Solving the Single Artery Traffic Signal Synchronization with Benders Decomposition. Presented at CORS/ORSA/TIMS Joint National Meeting, Vancouver, Canada, May 1990.
7. Solanki, R., R. S. Pillai, and A. K. Rathi. *A Fast Heuristic for Maximizing Bandwidth in Traffic Networks*. Oak Ridge National Laboratory, May 1993.
8. Gartner, N. H., S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND—A Variable-Bandwidth Arterial Progression Scheme. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
9. Tsay, H., and L. Lin. New Algorithm for Solving the Maximum Progression Bandwidth. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988.
10. Khatib, Z. *Bandwidth Optimization with Consideration for Left Turn Treatments*. Ph.D. dissertation. University of Illinois at Chicago, 1990.
11. Chaudhary, N. A., A. Pinnoi, and C. J. Messer. Scheme to Optimize Circular Phasing Sequences. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991.
12. Cunagin, W. D., D. Borchardt, and E. C. P. Chang. *Development of a Freeway Corridor Evaluation System—PASSER IV*. Report FHWA/TX-85 281-2F. Texas Transportation Institute, College Station, June 1985.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Enhancements to the PASSER II-90 Delay Estimation Procedures

MEHER P. MALAKAPALLI AND CARROLL J. MESSER

An enhanced delay estimation model for the popular traffic signal optimization model PASSER II-90 is described. Although the results from this model focus on enhancements to PASSER II-90, the findings presented should be useful to the future formulation of the *Highway Capacity Manual* (HCM) methodology for arterial streets. Development of the enhanced delay model primarily involved a four-step arrival rate model instead of the current two-step arrival rate model. Total delay was calculated on the basis of whether the traffic arrivals were early or late. Specifically, delay was estimated using the length and the time of arrival of the traffic platoon at the downstream intersection. TRANSYT-7F was used to investigate the effectiveness of the current PASSER II model and the enhanced PASSER II model. The enhanced PASSER II delay model resulted in large reductions in deviations of the delay values from TRANSYT-7F. Delay-offset trends in enhanced PASSER II-90 now closely follow the TRANSYT-7F delay-offset curves. Delays were also observed to closely follow the NETSIM curves in some regions. It was also observed that in the optimization mode, there was no significant difference in the calculated delay values between the old and the new estimation models. The new delay estimation model in PASSER II-90 also demonstrated that the platoon dispersion modeling in PASSER II compares favorably with TRANSYT's platoon dispersion factor of 0.30 to 0.35. Conclusively, the new model in PASSER II-90 has substantially improved delay estimation over all possible offsets for through traffic.

Delay analysis of signalized coordinated intersections is a very intricate process that requires a thorough understanding of the complex interactions among traffic demand, signal timing parameters, and traffic behavior. Chapter 9 of the 1985 *Highway Capacity Manual* (HCM) (*1*) devotes considerable space to the analysis of signalized intersections. The HCM uses average stopped delay per vehicle as the sole criterion for defining the levels of service provided at signalized intersections. One of the more important operational factors in determining the level of service at signalized intersections is the quality of traffic progression. Of all the variables affecting delay, the quality of traffic signal progression has the largest potential impact as shown by the wide range of progression adjustment factors (PFs), 0.4 to 1.85, in Table 9-13 of the HCM. Of concern to traffic engineers, however, is the fact that the PFs are based on limited field data. Hence, selection from a reasonable range of PFs in the table may often result in changes in the level-of-service designation for the approach (*2*).

Because of these concerns and because of the complexity involved in estimating and optimizing several signal-timing parameters, several computer simulation models have been developed by researchers for optimizing signal timing for signalized coordinated arterial streets and for networks. Familiar models to traffic engineers among these are TRANSYT-7F (*3*), MAXBAND (*4*), and PASSER II (*5*).

Despite the fact that TRANSYT-7F and its traffic model are realistic, it produces signal timing parameters that attempt to minimize disutility functions such as delay, stops, fuel consumption, and so on. But in reality, a major consideration in designing traffic signal timings for arterials (i.e., a series of intersections) is to achieve a reasonable amount of progression so that drivers who are traveling in the progression band are not required to stop at subsequent intersections once they have cleared the first intersection in green. TRANSYT-7F, thus, may not the best model for the traffic engineer to use where progression is the main consideration.

MAXBAND produces signal settings that achieve good progression but cannot guarantee delay minimization. Hence, results generated by MAXBAND may be efficient at providing large bands, but at the same time may cause undesirable systemwide delays. This deficiency in MAXBAND narrows its range of applications.

A model that overcomes these deficiencies to some extent is the PASSER II-90 program. PASSER II-90 is a macroscopic, deterministic model designed to optimize signal timing parameters to provide good progression along arterial streets. When the model was first developed in 1974, the sole purpose was to provide progression for the arterial through traffic. In 1978, delay evaluation for the progression solution was incorporated into the program. The model was further enhanced in 1987 by building simulation output into the program. Although the delay model in PASSER II-87 was better than that adopted by the 1985 HCM, it still had some inherent deficiencies in that the model did not take into account early or late traffic arrivals. Hence, the present paper focuses on more appropriate calculations of delay in the PASSER II program.

The main objectives of this paper are (*a*) to analyze the effectiveness of the traffic model and delay estimation for early or late traffic arrivals in PASSER II-90; (*b*) to demonstrate enhancements to the platoon dispersion or delay estimation models, or both, in PASSER II-90; and (*c*) to recommend analytical equations that are useful to the future HCM methodology for calculating delay to progressed movements at signalized intersections.

BACKGROUND

In the following sections, techniques adopted by the 1985 HCM and other models in estimating delay are elucidated.

M. P. Malakapalli, Bell-Walker Engineers, Inc., 914 140th Avenue NE, Suite 100, Bellevue, Wash. 98007. C. J. Messer, Texas Transportation Institute, Texas A&M University, College Station, Tex. 77840.

HCM Methodology (1985)

The HCM uses average stopped delay per vehicle for defining levels of service at signalized intersections. Stopped delay is estimated in the HCM using the following equations:

$$d = d_1 + d_2 \quad (1)$$

$$d_1 = \frac{0.38 * C * (1 - g/C)^2}{[1 - (g/C) * X]} \quad (2)$$

$$d_2 = 173X^2 * [(X - 1) + \sqrt{(X - 1)^2 + 16X/c}] \quad (3)$$

where

d = average stopped delay per vehicle (sec/veh),
d_1 = first-term delay for uniform arrivals (sec/veh),
d_2 = second-term delay for incremental random and overflow effects (sec/veh),
C = cycle length (sec),
g = effective green time (sec),
c = signal capacity (veh/hr), and
X = ratio of demand volume to signal capacity (v/c).

The HCM accounts for the effects of progression (platoon and dispersion effects) through the use of some adjustment factors called progression adjustment factors (PFs). The delay term d in Equation 1 is multiplied by the appropriate PF to obtain the actual average stopped delay. This PF is obtained from Table 9-13 of the HCM and depends on the v/c ratio and arrival type of the approach traffic. Five arrival types are used based on a variable called the platoon ratio, which is defined as the ratio of the percent vehicles arriving on green (PVG) to the green ratio of the movement. Platoon ratios (R_p) may range from a minimum value of 0 to a value greater than or equal to 1.5. Qualitatively, increasing platoon ratios or increasing arrival type numbers signify increasing progression.

Proposed Enhancements to HCM Methodology

A recent study (*2*) has suggested that applying the PF to the incremental delay term (Equation 3) is not appropriate. This argument seems logical because progression effects become negligible when oversaturated conditions exist, and hence the second term of the delay equation should not contain any external adjustment factors for platoon traffic.

Further, as has been pointed out by several researchers (*2,6*), the R_p and PF used by HCM are dependent on the g/C ratio of the approach. Since the quality of progression is a function of several variables such as signal offset, spacing between the intersections, dispersion, and traffic volume, the platoon ratio may not be an accurate descriptor of the quality of signal progression. The PFs thus derived from the corresponding R_p-values may not best determine the delay values. In addition, no consideration is given to the early or late traffic arrivals by the PFs.

Fambro et al. (*2*) developed a new set of PFs to be used as replacements for Table 9-13 of the HCM. The existing delay equation was modified to include a term for the quality of signal progression. A set of empirical factors was also proposed to take into account early and late traffic arrivals at successive intersections. Those factors were derived on the basis of whether the front of the platoon arrived during the first, middle, or last third of the green or red periods. The equation and the adjustment factors may eliminate some of the existing discrepancies in the current method of HCM delay estimation. In PASSER II, however, more detailed knowledge of dispersion, offset, and other variables may be used to more accurately predict the delay without using empirical adjustment factors.

Other Delay Formulations

Rouphail (*7*) derived several delay formulations for mixed platoon and secondary flows. One model assumes two arrival flow rates, one within the progression band and another outside the progression band. Though the model seems better than most existing models, it effectively disregards the early or late traffic arrivals at the downstream intersection. When traffic arrivals vary or straddle the green, there are, in effect, three arrival rates. The method also requires bandwidth as an input.

In addition to the delay equation proposed by Fambro et al. in National Cooperative Highway Research Program (NCHRP) Report 339, which includes a term for the quality of signal progression, Staniewicz and Levinson (*8*) developed several equations for various arrival types. These equations, however, may not be used for secondary flow conditions. They are also more microscopic in nature and thus are not practical to incorporate into PASSER II-90.

TRANSYT Methodology

TRANSYT (*9*) has become one of the most widely used tools for traffic flow analysis and traffic signal timing optimization in the world. The effectiveness of the signal timings developed by the program depends heavily on the delay calculated by the model. The delay calculation in TRANSYT for coordinated signalized intersections is estimated by integrating the arrival and departure profiles of traffic at the downstream intersection. The accuracy of the arrival flow profile at the downstream intersection in turn depends on the platoon dispersion algorithm utilized by TRANSYT. Thus, the fundamental principle of traffic representation in TRANSYT is the platoon dispersion behavior. TRANSYT uses a recursive formula to predict the platoon dispersion behavior of the traffic. Further discussion on platoon dispersion modeling in TRANSYT may be found in the User's Manual (*3*).

MODEL DEVELOPMENT

Three major stages were involved in developing the model together with analyzing and evaluating the methods of traffic delay modeling in PASSER II. First, an arbitrary arterial street system was established with all the traffic and signal timing variables affecting the delay estimation well defined.

Second, a factual method of examining the accuracy of the traffic modeling or delay estimation procedures was adopted. This analysis was intended to determine if the delay estimation procedures indeed showed some inconsistencies and modifications were in order. This determination was achieved by observing delay-offset relationships in PASSER II-90 against those in TRANSYT-7F.

The third stage involved developing enhancements to the existing modeling procedures in PASSER II. On the basis of results from the second stage, new or enhanced modeling techniques that would have significant impact on the output were devised.

Stage 1: Establishment of Arterial Street System

A two-intersection arterial was defined for the purposes of this research. The traffic modeling or delay estimation methods (hereafter referred to as the TRAMDE methods) and the trends of results in PASSER II would be the same irrespective of the number of intersections in the arterial. Though the study on the response of PASSER II to different traffic and signal settings was made with varying spacings, the principal focus was on a spacing of 403 m, which was deemed to be a reasonable and ideal representation of platoon dispersion along an arterial system.

The signal timing parameters that were needed for the simulation were cycle lengths, phase splits, offsets, and so forth. A cycle length of 100 sec was chosen for convenience, with green splits of 40 and 60 percent for progressed and nonprogressed traffic flow, respectively. The choices were based on the fact that main street green splits for multiphase signals are generally not more than 50 percent of the cycle length. Offsets were varied from zero to the cycle length in multiples of 5 sec.

Traffic volumes were such that the intersections always remained undersaturated, since quality of progression has an insignificant effect on the uniform delay component for oversaturated conditions. A *v/c* ratio of 0.5 or 0.8 was considered to be reasonable to represent moderate and high-volume conditions, respectively.

Two cases of volume variations were examined. The first variation excluded any secondary flow component from the upstream intersection to the downstream intersection. The second variation was to assign 20 percent of the through volume at the downstream intersection to the nonprogressed traffic at the upstream intersection. This variation was done to examine the appropriateness on the part of the program in modeling the secondary flow.

Stage 2: Identification of Procedures for Investigating TRAMDE Methods in PASSER II

As has been pointed out in earlier sections, a widely accepted program for traffic model and delay estimation is TRANSYT-7F. Though there are some conflicting views on an appropriate platoon dispersion factor for TRANSYT, the recommended value of 0.35 seems a plausible value that represents fairly good traffic conditions in the field in most cases. The delay estimates of TRANSYT-7F have proved to be reliable throughout the world. Hence, TRANSYT-7F was chosen for examining the proposed delay estimation enhancements in PASSER II.

Delay-Offset Relationships

A logical way to study the delay estimation method was to examine whether the delay-offset trends in PASSER II and TRANSYT were similar for various intersection spacings and traffic volumes. In order to verify accurate trends, an established microscopic simulation program, NETSIM (*10*), was used to corroborate the findings. Various combinations of inputs that were tested for this analysis were (*a*) two main street volume variations with *v/c* ratios of 0.48 and 0.8, respectively; (*b*) a spacing of 403 and 805 m; (*c*) offset variations ranging from 0 to the cycle length in 5-sec increments; and (*d*) a platoon dispersion factor (α) of 0.35. Several values of α would make the analysis too complicated for the anticipated benefits. Hence, an α-value of 0.35 was used.

Figure 1 shows the delay-offset curves for 0 percent nonprogressed volume and a *v/c* ratio of 0.48 for progressed traffic at a spacing of 403 m for PASSER II-90, TRANSYT-7F, and NETSIM. Figure 2 shows a similar curve for the same parameters with a *v/c* ratio of 0.8. The plots clearly show an inconsistency of shapes on the part of PASSER II-90 in estimating delay in some offset regions.

In both of the above cases, it can be clearly seen that the delay in PASSER II was either overestimated or underestimated in two or more regions. Figures 1 and 2 reveal that PASSER II consistently overestimates the delay on the right side of the ideal offset and underestimates the delay on the left side of the curves. The portion of the curves on the right side of the ideal offset signifies early traffic arrivals, where the front of the traffic platoon arrives in the later part of the red period. Increasing offsets to the right of the ideal offset indicate that the green time to the platoon traffic is being displayed late, and hence traffic arrivals automatically become early. On the other hand, the portion of the curves to the left of the ideal offset indicates late arrivals, where the rear of the platoon arrives in the early portion of the red. This inconsistency was largely due to the delay estimation in the red period for early and late traffic arrivals made by PASSER II.

This flaw in the delay estimation necessitates a thorough understanding of the traffic and delay modeling techniques currently used in PASSER II-90. These techniques of PASSER II will be detailed in the following section.

Traffic and Delay Modeling in PASSER II

A major component of traffic representation in any macroscopic model for signalized intersections is the platoon dispersion model. The model in PASSER II (*11*) uses platoon length at the upstream intersection to estimate platoon length at the downstream intersection. The length of the platoon at the upstream intersection i, LP_i, is given by

$$LP_i = g_o[PVR + (PVG * g_o)/g] + PVG(g - g_o) \quad (4)$$

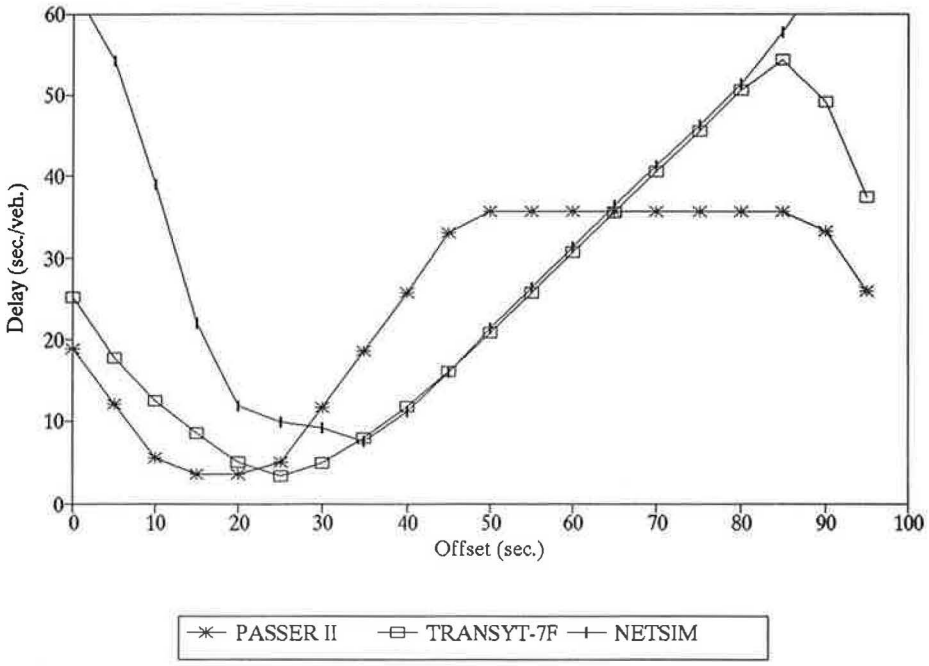


FIGURE 1 Delay-offset relationships for v/c = 0.48 and spacing = 403 m (100 percent platoon traffic; alpha = 0.35, ϕ = 26.4).

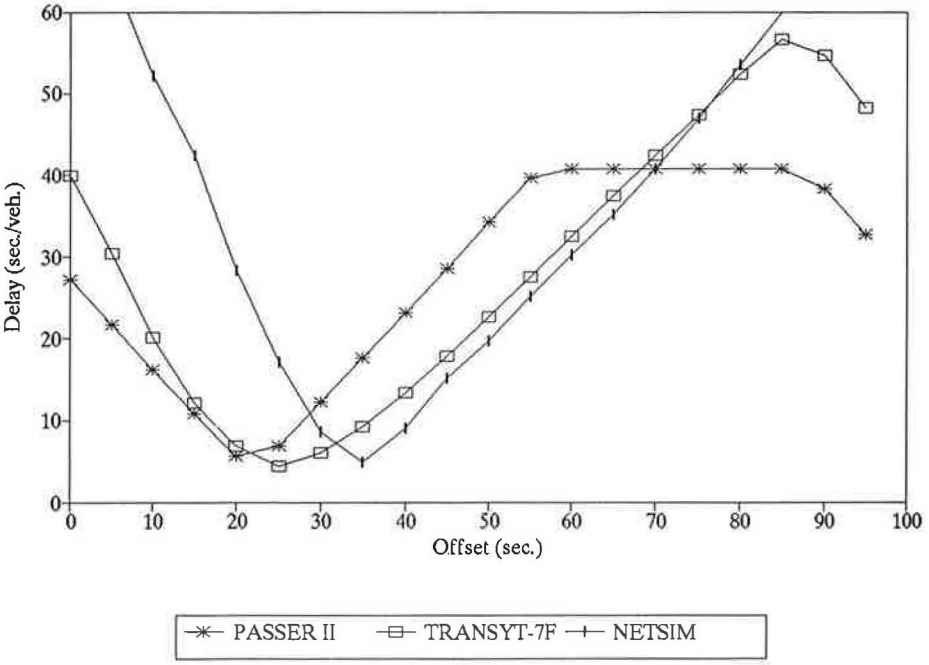


FIGURE 2 Delay-offset relationships for v/c = 0.8 and spacing = 403 m (100 percent platoon traffic; alpha = 0.35, ϕ = 26.4).

where

g_o = time required for queued vehicles to clear the intersection at i (sec),
PVR = percent vehicles arriving on red at i,
PVG = percent vehicles arriving on green at i, and
g = effective green time for the main street at i (sec).

The platoon length at the downstream intersection (j), LP_j (see Figure 3) is now estimated as

$$LP_j = LP_i * PD_{ij} + 0.8(0.9 + 0.056t_{ij}) \quad (5)$$

where

PD_{ij} = platoon dispersion factor written as in the report by Messer et al. (5),
= $1.0 + (0.026 - 0.0014 * NP)\, t_{ij}$, in which t_{ij} = travel time between i and j in seconds and NP = number of vehicles in platoon at i.

The percent vehicles arriving on green (PVG) is a critical factor in the delay calculation. PASSER II-90 estimates PVG using the following formula:

$$PVG = PTT_j * GO_j/LP_j + (1 - PTT_j)RO_j/(C - LP_j) \quad (6)$$

where

PTT_j = percent of total through traffic arriving from i at j,
= (through traffic at i/through traffic at j),
GO_j = green overlap for the through traffic from i at j as shown in Figure 3 (sec), and
RO_j = green overlap for the secondary flow component from i at j (sec).

The flow rate in the green period (q_g) is calculated by the relation

$$q_g = PVG * q * C/g \quad (7)$$

The percent vehicles and the flow rate in the red period at j are calculated using the following relation:

$$PVR = 1 - PVG$$

$$q_r = PVR * q * C/r \quad (8)$$

where

q = average flow rate of through traffic at j (veh/sec),
C = cycle length (sec),
g = effective green (sec),
PVR = percent vehicles in red at j, and
r = effective red (sec).

Figure 4 shows how PASSER II defines these two flow rates, q_g and q_r, in the cycle at j. This definition was also proposed by Olszewski (6). These are the two flow rates that PASSER II uses to calculate the uniform delay component of the average delay. The uniform delay is now computed using a stepwise integration of the queue lengths in the red and green periods. An approximation of the uniform delay (UD) calculation, in seconds per vehicle, can be written in the following form:

$$UD = q_r * r^2 / (2 * q * C)[1 + q_r/(s - q_g)] \quad (9)$$

where s is the saturation flow rate in vehicles per second per green per lane and all other terms are as explained before.

A deeper look at Equation 6 would suggest that for a given C, g, r, and platoon volumes, PVG and hence PVR would always yield the same value if GO_j and RO_j are constant. Under these conditions, q_r would always be the same irrespective of the time at which the platoon arrives in the red period. Consequently, the obtained UD would be the same and the delays experienced by traffic arrivals in the early part of the red (late arrivals) and later part of the red (early arrivals) are also the same when in reality they are considerably different. In the former case (late arrivals), delay is much

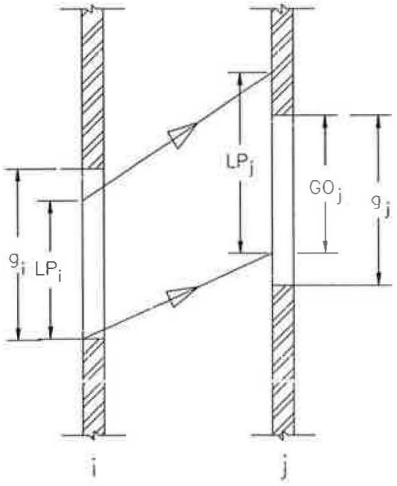


FIGURE 3 Model of progression platoon movement from intersection i to j used by PASSER II-90.

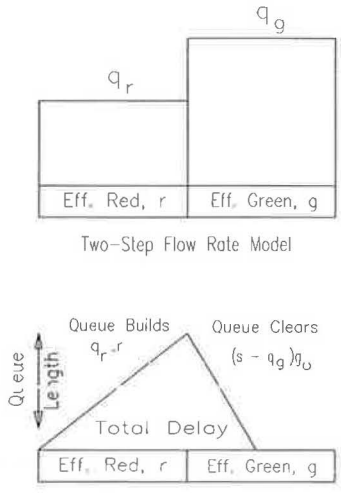


FIGURE 4 Flow-rate definition and delay calculation in PASSER II (11).

higher than in the latter case (early arrivals). It can also be observed that the flow rate for the platoon and secondary flows is combined into one flow rate in both green and red periods, which may not invariably be true. A major deficiency of PASSER II lies in these flow rate definitions and subsequent delay computation methods. Enhancements in these two techniques will be dealt with in the next section.

Stage 3: Enhancements to Existing Model

Two enhancements were made to the existing model; the major enhancement was the delay estimation technique in PASSER II.

First Enhancement

The first modification was concerned with the platoon dispersion aspect of PASSER II and was developed for easy comprehension. An equivalent form of Equation 4 can be written as

$$LP_i = g_o + PVG(g - g_o)^2/g \tag{10}$$

where all the terms are as previously defined.

Though Equation 10 was developed analytically, the same equation can also be derived by mathematical manipulation of Equation 4. Equation 10 is simplistic, easy to understand, and also easy to incorporate into the program. Further, the boundary conditions of Equation 10 are easily discernible, unlike those of Equation 4. For example, when $PVG = 0$, LP_i is equal to g_o, and when $g_o = g$, $LP_i = g$. From a glance at Equation 10, one can easily determine these boundary conditions, whereas Equation 4 requires some computation to arrive at the same boundary conditions.

Second Enhancement

Major modification in PASSER II-90 involved the delay calculation made by PASSER II for the vehicles arriving in the red period. Figure 5 shows the proposed modification made for PASSER II for the estimation of q_r. The modification involves defining three arrival rates in the red period at the downstream intersection: a flow rate for the early traffic arrivals, which are part of the main street platoon traffic; a flow rate for the late arrivals, also part of the main street platoon traffic; and a flow rate for the nonprogressed traffic during the red. The flow rates were calculated using the following equations:

$$q_{re} = PTT_j * r_e/LP_j$$

$$q_{rl} = PTT_j * r_l/LP_j$$

$$q_{\text{rel}} = (1 - PTT_j) * [r - (r_e + r_l)]/(C - LP_j)$$

where all the variables are as defined earlier except

q_{re} = flow rate for the early arrivals of the platoon traffic,
q_{rl} = flow rate for the late arrivals of the platoon traffic,

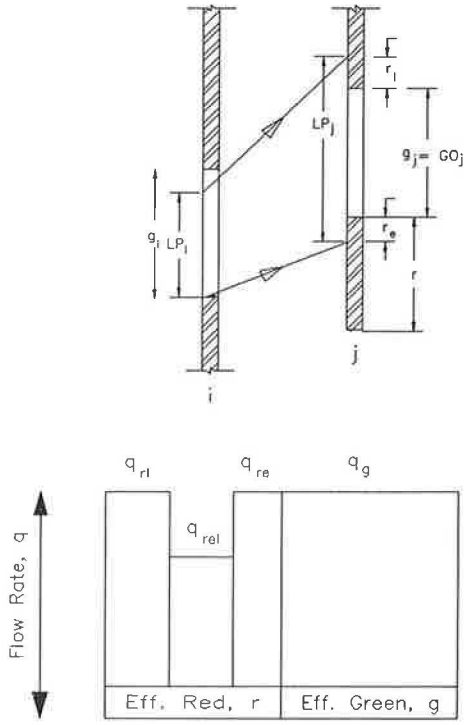


FIGURE 5 Modified flow-rate definitions in PASSER II.

q_{rel} = flow rate (early/late) for the nonplatoon traffic,
r_e = red overlap for the early platoon traffic (Figure 5), and
r_l = red overlap for the late platoon traffic (Figure 5).

The flow rates q_{re} and q_{rl} will be equal because of the assumption of a constant flow rate in the platoon length LP_j as shown by Equation 6. The nonplatoon flow rate will be late whenever the platoon flow rate is early or straddles the red. Similarly, q_{rel} will be early whenever the platoon traffic is late. When the front and rear of the platoon traffic arrive in the red period, q_{rel} will be both early and late with equal flow rates as for the platoon traffic. The uniform delay (UD) estimation was then made in the program using the same stepwise demand integration with only minor modifications.

An approximate equation similar to the HCM equation for the foregoing estimation in most cases was derived in two parts. The first part (UD1) was meant for the platoon traffic delay in the red, and the second part (UD2) was meant for the nonplatoon traffic. The first part is given below:

$$UD1 = [q_r * r^2/(2 * q * C)] * \text{FEAL} \tag{11}$$

where

q_r = platoon flow rate in the red, q_{rp};
= $(PTT_j - PVG_{jp}) * q * C/r$ (veh/sec);
PVG_{jp} = percent vehicles in green for the platoon traffic;
= $(PTT_j * GO_j/LP_j)$; and
FEAL = factor for early and/or late arrivals as given by $[(r_e - r_l)/r] + [2 * r_l/(r_l + r_e)]$.

All other variables have been defined previously. Note that Equation 11 is similar to the uniform delay equation in the

red part as proposed by Fambro et al. in NCHRP Report 339 with just one additional analytical factor to explicitly take early and/or late traffic arrivals into account.

The second part of the uniform delay term for the nonplatoon traffic (UD2) is the same as Equation 11 except that r_l and r_e have different values for the secondary flow. Also, q_r, nonplatoon flow rate in the red, will be different. It is estimated in vehicles per second as

$$q_r = (1 - PTT_j) - PVG_{np} * q * C/r$$

where PVG_{np}, percent vehicles in the green for the nonplatoon traffic, is $(1 - PTT_j) * RO_j/(C - LP_j)$.

The final approximate uniform delay term is

$$UD = UD1 + UD2 + q_r^2 * r^2/(2 * q * C)[1/(s - q_g)] \quad (12)$$

where q_r is the value obtained from Equation 8.

It can be noted that the delay during the queue clearance time at the downstream intersection is not affected by the early and/or late platoon or nonplatoon arrivals, which is logically true. Equation 12 is similar to the equation in NCHRP Report 339 with arrival rates for platoon and nonplatoon traffic distinctly computed. The NCHRP Report 339 equation for uniform delay is given as

$$UD = [q_r * r^2/(2 * q * C)] * [1 + q_r/(s - q_g)] \quad (13)$$

MODEL RESULTS

Delay-Offset Relationships

The delay-offset relationships were further examined with respect to the modified equations in PASSER II-90. Plots of the results are shown in Figures 6–8.

Figure 6 shows the delay-offset relationships for 0 percent nonprogressed traffic and a *v/c* ratio of 0.48. Figure 7 presents a delay-offset plot for a *v/c* ratio split of 0.67:0.17 between the platoon and the nonplatoon traffic, respectively, at the upstream intersection. This split represents a nonprogressed traffic *v/c* ratio at the upstream intersection that is 20 percent of the total approach *v/c* ratio (approximately 16 percent nonplatoon flow). The spacing between the intersections in both cases was 403 m. To add generality, Figure 8 shows the delay-offset curve for a 30 percent nonplatoon flow at a spacing of 201 m. Note that all the TRANSYT-7F plots were calculated with a platoon dispersion factor of 0.35 ($\alpha = 0.35$).

The graphs clearly show a significant improvement in delay estimation by the PASSER II-90 model. The delay-offset trend clearly traces the TRANSYT curve in virtually all cases. Table 1 summarizes the average and the maximum percent deviation of the old PASSER II-90 delay estimation (similar to Equation 13) and the new PASSER II-90 delay estimation (similar to Equation 12).

The average percent deviation in Table 1 is the percent deviation of all the delay values averaged from a 0-sec offset to a 95-sec offset. The maximum percent deviation of delay between the old and new PASSER II from TRANSYT is also given in the table.

Simulation Versus Optimization Results

The modified delay estimation technique was also tested with two optimization runs for an arterial consisting of four intersections. The maximum percent deviation obtained for the through traffic movement was 47 percent with the optimization run. There was no significant difference between the old and the new delay values as far as the absolute differences were concerned. The maximum absolute difference between

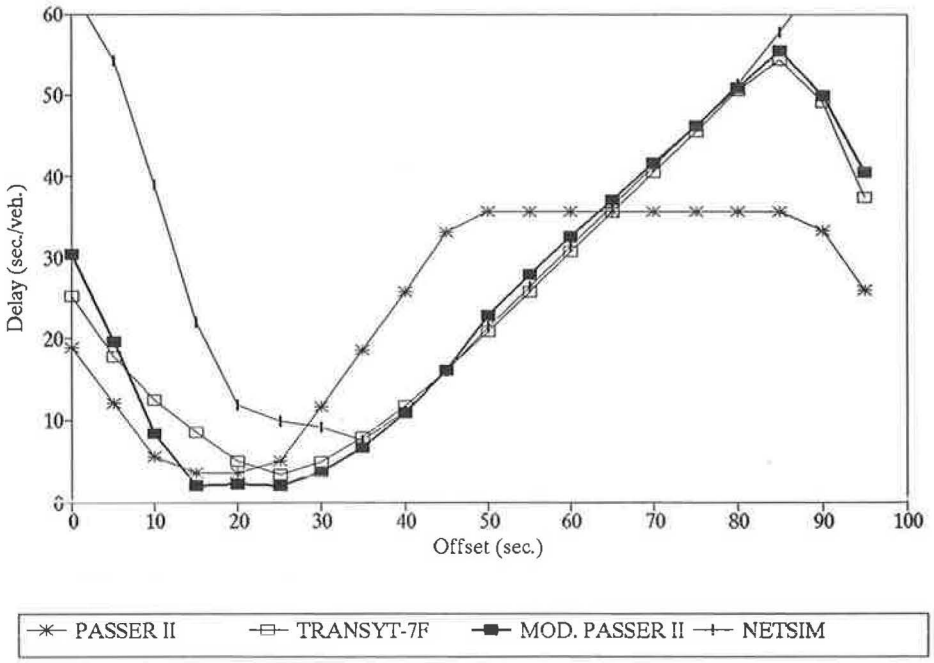


FIGURE 6 Modified delay-offset relationships for *v/c* = 0.48 and spacing = 403 m (100 percent platoon traffic; alpha = 0.35, ϕ = 26.4).

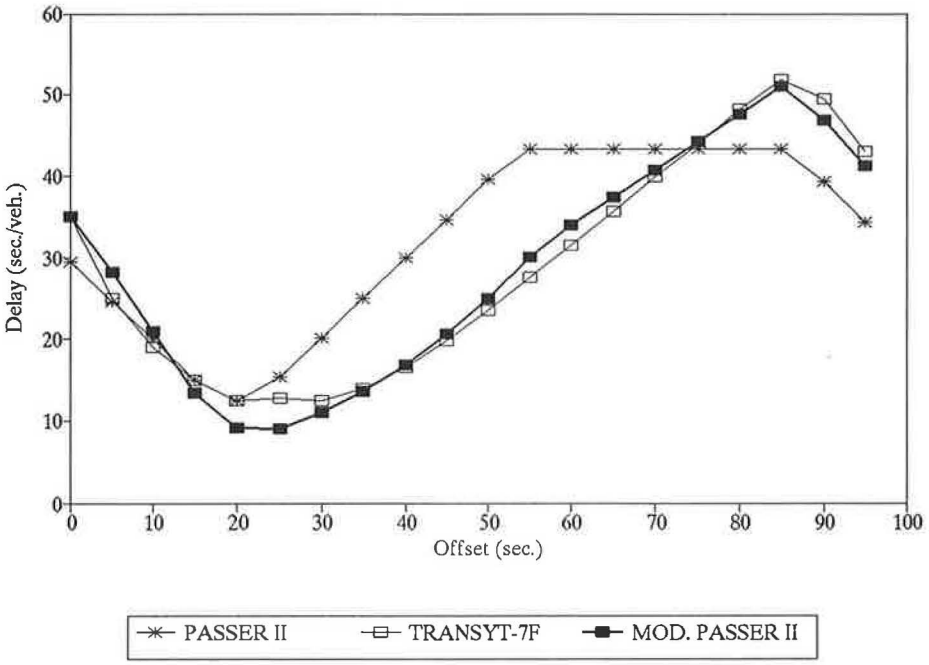


FIGURE 7 Modified delay-offset relationships for v/c = 0.8 and spacing = 403 m (16 percent nonplatoon traffic; alpha = 0.35, ϕ = 26.4).

the old and new PASSER II delay estimates was observed as 4 sec in the optimization runs. This negligible difference between the old and the new delay estimates can be attributed to the fact that with optimization, the model attempts to maximize through traffic arrivals and departures in the green time, thus minimizing early or late traffic arrivals in the red period. In addition, observations of Figure 1 show that small differences in delay between PASSER II-90 and TRANSYT-7F occur near optimal progression delays.

CONCLUSIONS AND RECOMMENDATIONS

Summary

An enhanced delay estimation model for the popular traffic signal optimization model PASSER II-90 has been provided. The enhanced delay estimation model primarily involved development of a four-step arrival rate model instead of the current two-step arrival rate model. Total delay was calculated

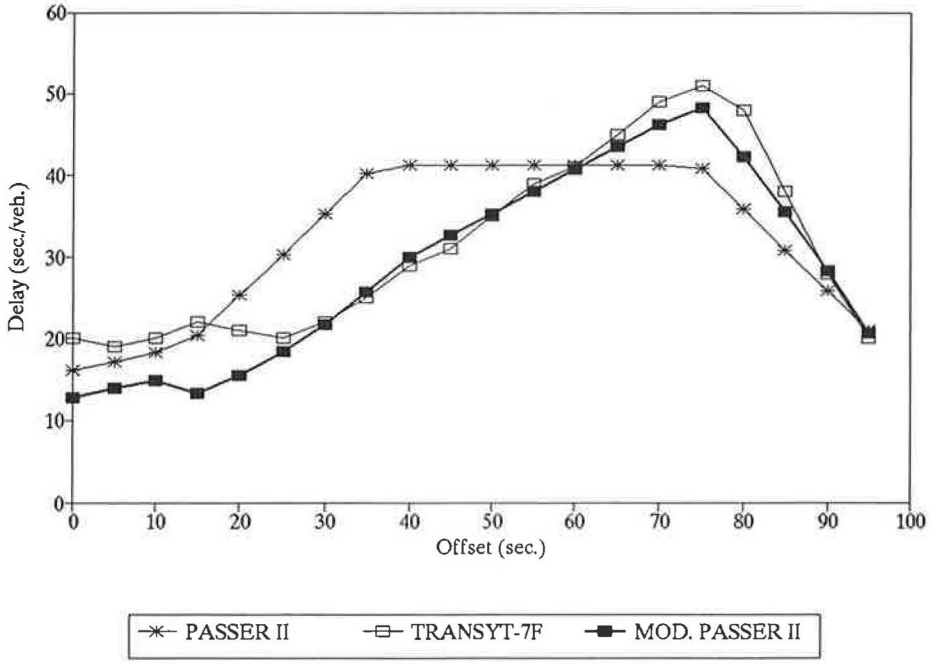


FIGURE 8 Modified delay-offset relationships for v/c = 0.8 and spacing = 201 m (30 percent nonplatoon traffic; alpha = 0.35, ϕ = 13.2).

TABLE 1 Deviations of Old and New PASSER II-90 from TRANSYT-7F

	Deviations (%)			
	Average		Maximum	
Case	Old	New	Old	New
v/c = 0.48 100% platoon s = 403 meters	50.2	16.4	140	76
v/c = 0.8 100% platoon s = 403 meters	37.6	8.2	103	32
v/c= 0.8 16% non-platoon s = 403 meters	25	11.8	53	42
v/c = 0.8 30% non-platoon s = 201 meters	21.9	10.9	61	40

on the basis of early or late traffic arrivals. In other words, delay was estimated using the length and the time of arrival of the traffic platoon at the downstream intersection.

TRANSYT-7F was assumed as a model that could predict accurate delay values and was used to investigate the effectiveness of the current PASSER II model and the enhanced PASSER II model. In some cases, NETSIM was also used as a check for consistency in delay estimation.

Conclusions

The enhanced PASSER II-90 delay model resulted in large reductions in percent deviations of the delay values from TRANSYT-7F. Delay-offset trends in PASSER II-90 were observed to closely follow the TRANSYT-7F delay-offset curves. Delays were also observed to closely follow the NETSIM curves in the portion to the right of the target offset. NETSIM predicted much higher delay values in the region where the offset was lower than the target offset in the delay-offset curves. It is possible that this disparity occurred because NETSIM was estimating too many late traffic arrivals.

The enhanced delay estimation technique was also examined with respect to signal optimization in PASSER II-90. It was observed that there was no significant difference in the calculated delay values between the old and the new estimation models. The maximum absolute difference was observed to be about 4 sec, and in terms of deviation from TRANSYT-7F, it was 47 percent. This negligible difference between the old PASSER II-90 delay modeling and the new delay modeling in PASSER II-90 was mainly because most of the through traffic platoon was arriving and leaving in the green time, and hence there were few early or late traffic arrivals.

The new delay estimation model in PASSER II-90 also demonstrated that the platoon dispersion modeling in PASSER II compares with TRANSYT's platoon dispersion factor of 0.30 to 0.35 (as indicated by the delay-offset curves). Hence, it can be stated that PASSER II's platoon dispersion model may not need refinements or enhancements. Conclusively, the new model in PASSER II-90 has substantially improved the delay estimation for the through traffic.

Recommendations to NCHRP Report 339 Equation

As discussed earlier, Fambro et al. (*2*) proposed modifications to the HCM methodology for calculating uniform delay. Progression adjustment factors were proposed to take early or late traffic arrivals into account. As a modification to these empirical factors, an analytical factor (FEAL) is recommended as given in Equation 11. Uniform delay is divided into three parts as given in Equation 12. The first two parts take into account the delay in the red period for primary and secondary flows and the third part calculates delay during the clearance time.

Estimation of all variables in Equation 12 was described in earlier sections. A discussion on measuring delay in the field may be found in NCHRP Report 339 (Chapter 1, p. 15). Note that to apply Equation 12, two additional steps are required:

1. Isolate measurement of the proportion of volume arriving on the green for the primary traffic from the total proportion of volume arriving on the green, and
2. Measure the time of arrival of the front and/or rear of the primary platoon traffic with respect to the start of green.

Once these steps have been completed, Equation 12 can be applied. It is hoped that Equation 12 in conjunction with the NCHRP Report 339 equation will yield reliable uniform delay estimates for application in Chapters 9 and 11 of the HCM.

ACKNOWLEDGMENT

This research was performed when the first author was a research assistant at Texas Transportation Institute, Texas A&M University.

REFERENCES

1. Signalized Intersections. In *Special Report 209: Highway Capacity Manual*, TRB, National Research Council, Washington, D.C., 1985.
2. Fambro, D. B., E. C. P.Chang, and C. J. Messer. *NCHRP Report 339: Effects of the Quality of Traffic Signal Progression on Delay*. TRB, National Research Council, Washington, D.C., Sept. 1991.
3. *TRANSYT-7F User's Manual.* Transportation Research Center, University of Florida, Gainsville, Feb. 1983.
4. Little, J. D. C., and M. D. Kelson. *Optimal Signal Timing for Arterial Signal Systems, MAXBAND.* Summary Report 1. Operations Research Center, Massachusetts Institute of Technology, Cambridge, 1980.
5. Messer, C. J., et al. *A Report on the User's Manual for Progression Analysis Signal System Evaluation Routine—PASSER II.* Research Report 165-14. Texas Transportation Institute, Texas A&M University, College Station, 1987.
6. Olszewski, P. S. Traffic Signal Delay Model for Non-Uniform

Arrivals. Presented at 69th Annual Meeting of the Transportation Research Board, Washington, D.C., 1990.
7. Rouphail, N. M. Delay Models for Mixed Platoon and Secondary Flows. *Journal of Transportation Engineering*, Vol. 114, No. 2, March 1988.
8. Staniewicz, J. M., and H. S. Levinson. Signal Delay with Platoon Arrivals. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985.
9. Robertson, D. I. *TRANSYT—A Traffic Network Study Tool.* Laboratory Report 253. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1969.
10. *Traffic Network Analysis with NETSIM—A User's Guide.* FHWA Implementation Package FHWA-IP-80-3. FHWA, U.S. Department of Transportation, 1980.
11. Messer, C. J., D. B. Fambro, and D. A. Anderson. *A Study of the Effects of Design and Operational Performance of Signal Systems.* Research Report 203-2F. Texas Transportation Institute, Texas A&M University, College Station, 1975.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Hybrid Genetic Algorithm To Optimize Signal Phasing and Timing

MOHAMMED A. HADI AND CHARLES E. WALLACE

Signal timing optimization involves the selection of four basic design elements: phase sequence, cycle length, green split, and offset. None of the available signal timing models is considered adequate to optimize all four design elements, particularly in two-dimensional networks. Among the current models, TRANSYT-7F is most effective for timing, but it does not optimize phasing. Researchers have considered several methods for enhancing TRANSYT-7F to include phasing optimization but thus far no method has proven practical. An exhaustive search of possible phasing combinations is computationally prohibitive; thus a new approach is needed. Genetic algorithms (GAs) are heuristic probalistic search procedures that have been applied to a wide range of engineering problems. The use is investigated of a GA in combination with the TRANSYT-7F optimization routine to select all signal timing design elements. The main purpose of the GA in the proposed scheme is to optimize phase sequences. Two implementations of the GA model are presented. In the first, the GA and TRANSYT-7F optimization routines are executed concurrently to achieve an optimal solution. In the second, the GA is allowed to optimize cycle length, phase sequences, and offsets. Then TRANSYT-7F is used to adjust the resultant signal timing. The results suggest that both implementations have potential for optimizing signal phasing and timing. However, the first method produces more consistent results. It also requires longer execution time.

Several computer models have been developed for off-line optimization of signal timing. Each of these models has its own area of application and its own particular strengths and weaknesses. For coordinated systems, the optimization models currently used include PASSER II (*1,2*), MAXBAND (*3*), and TRANSYT-7F (*4*). None of these models is all inclusive, and a combination of them is often used to achieve a desired design policy (*5–9*).

PASSER II and MAXBAND select a signal timing plan by maximizing bandwidth efficiency (the ratio of total bandwidth to the cycle length) and have been used primarily for arterial streets. Although MAXBAND has been extended for application to multiple-arterial networks (*10*), this version is not widely used, primarily because of excessive computer time and its current limitation to mainframe computers.

Both MAXBAND and PASSER II can optimize cycle length, phase sequences, and offsets. The main advantage of these programs is their ability to optimize phase sequences. One important disadvantage is that maximizing bandwidth does not necessarily result in optimal system performance in terms of stops, delay, and fuel consumption. This is because the maximal bandwidth design strategy does not explicitly recognize traffic demand as a function of time on individual links. In addition, these programs do not optimize green splits, although they do calculate splits based on manipulating the degrees of saturation.

TRANSYT-7F has been used for arterial streets as well as two-dimensional networks. Traditionally, TRANSYT-7F optimizes cycle length, splits, and offsets by minimizing a disutility index (DI), which is a function of delay, stops, fuel consumption, and, optionally, queue spillover. In the latest release of the program, a progression-based optimization model that uses the progression opportunities (PROS) concept has been implemented, overcoming an earlier shortcoming of the model (*11,12*). This concept expands upon the maximal bandwidth approach by considering short-term progression opportunities within the system.

In spite of this major enhancement, the most significant deficiency in TRANSYT-7F remains—its inability to optimize phase sequences. Phasing is an input to the model and cannot be changed in a given run. To examine alternative sequences, they have to be explicitly coded in multiple runs. Cohen (*9*) noted that adding a phase sequence optimization capability to TRANSYT-7F would involve combining a linear gradient search technique with a combinatorial problem. He recognized that this appears to be computationally infeasible since there are 4^n possible phase sequence combinations at n intersections, assuming four possible phase sequences, namely, leading left, lagging left, leading in one direction and lagging in the other, and vice versa.

To deal with this problem, other computer programs such as PASSER II and MAXBAND have been used as a "preprocessor" to determine the phase sequence before TRANSYT-7F is run to optimize signal timing. It has been shown that this method has the potential for improving the signal timing plan produced by TRANSYT-7F (*8,9*). This strategy has been particularly successful for designing arterial signal timing.

For this purpose, model integration programs like the Arterial Analysis Package (AAP) have been used to run PASSER II and TRANSYT-7F from a common data base, thus reducing the effort required to implement this strategy (*13*). PASSER II can be run from the AAP, which can import the phasing and timing; then these can be automatically mapped as the initial timing for a TRANSYT-7F optimization run. This also has the advantage of providing TRANSYT-7F with a better "starting point," which has also been shown to lead to potentially better solutions (*5,6*), although with the new PROS options this factor is less important (*11*).

For networks, the development of integration models similar to the AAP has been delayed because of the unavailability of a phase sequence optimization program that can deal with networks and run efficiently on microcomputers.

Transportation Research Center, University of Florida, 512 Weil Hall, Gainesville, Fla. 32611.

(Note that FHWA will be developing such a package in the near future.)

The dilemma facing software developers is how to overcome these functional gaps, particularly for networks:

- Merging the functions of PASSER II or its network counterpart, PASSER IV, which is under development [reported functionally by Chaudhary et al. elsewhere (*14*)], with TRANSYT-7F. This does not seem to be practical for both functional and programmatic reasons, not to mention the problem of software "ownership."
- Improving upon the integrated application of a bandwidth model (PASSER IV or MAXBAND) and TRANSYT-7F to be more automatic. This is certainly feasible, but it will still require user intervention.
- Putting phase sequence optimization explicitly into TRANSYT-7F. This is technically possible, but the massive restructuring of the model makes the chances for a timely and successful undertaking unlikely.
- Applying a new approach to using TRANSYT-7F in an iterative fashion to optimize phasing by trial and error. As mentioned earlier, the possible combinations seem to make this approach prohibitive unless a more effective method than exhaustive search can be found.

The last possibility is the aim of this paper. A new approach is applied to an iterative, heuristic algorithm for phase sequence optimization using TRANSYT-7F.

BACKGROUND OF PROPOSED METHOD

Heuristic search strategies called genetic algorithms (GAs) are finding increasing application in a variety of problems in science, engineering, business, and the social sciences (*15*). The GA strategies are based on the mechanics of natural selection and natural genetics. A number of analytical and empirical studies have demonstrated their capability in function optimization and artificial intelligence applications.

The number of solutions evaluated using GAs to locate a satisfactory solution to a given problem is small in comparison with the size of the search space. Nevertheless, the computational requirements for GAs can be severe. A hybrid scheme that switches from the genetic search to a conventional nonlinear programming approach has been suggested in the literature to deal with this problem (*15*).

Foy et al. (*16*) investigated the implementation of a GA to produce optimal, or near optimal, signal timing. In that study, the GA was used to optimize cycle length, splits, and course offsets (actually simply flipping the two phases with no explicit offsets) for two-phase operations for a four-intersection network. In the optimization process, a simple microscopic simulation model was used to evaluate alternative solutions based on minimizing delay.

The results of Foy et al. show an improvement in the system performance when this GA strategy is used and suggest that GAs have potential in optimizing signal timing. The results obtained, however, were not compared with those that could be achieved using existing optimization models. In addition, the GA model was applied to a very simple system with two-phase operation and no explicit offsets between intersections. More typical real-world applications of the GA model are needed to prove its effectiveness.

In this study, a GA was used in conjunction with the TRANSYT-7F hill-climbing routine to optimize all four signal timing parameters (offset, split, cycle length, and phase sequence). This hybrid scheme works with arteries as well as two-dimensional networks and can optimize timing based on system performance (TRANSYT-7F DI), PROS, bandwidth, or a combination of these.

TRANSYT-7F MODELS

TRANSYT-7F is one of the most powerful computer programs for traffic signal timing and traffic flow analysis. Traditionally, TRANSYT has consisted of two main parts: a traffic flow model and an optimization model.

The traffic flow model in TRANSYT-7F is a deterministic, macroscopic, time-scan simulation. It simulates traffic flow in a street network of signalized intersections to compute a DI for a given signal timing and phasing plan. This DI is a function of stops, delay, fuel consumption, and, optionally, queue spillover.

The TRANSYT-7F optimization procedure is based on an iterative, gradient search technique known as hill-climbing. It makes changes to the signal timing to determine whether a performance index (PI) has improved. By adopting only those changes that improve the PI, the optimizer tries to find a set of timing that optimizes the PI subject to any constraints.

In the latest release of the program (*4*), an improvement was added that allows it to calculate the PROS value for multiarterial networks. PROS is a measure of progression that considers not only through bands but also short-term progression opportunities within the system (*11,12,17*).

In the past, the PI used in the optimization routine of TRANSYT-7F was always the DI calculated by the traffic flow model. However, the latest release of the program has the ability to optimize signal timing on the basis of PROS only, PROS and DI, PROS or DI, and DI only. It has been shown that the PROS-related strategies produce significant improvements in progression compared with DI-only optimizations for single arteries as well as for multiarterial networks (*11*).

ELEMENTS OF GENETIC ALGORITHMS

GAs are a family of adaptive search procedures that are loosely based on models of genetic changes in a population of individuals. The main advantage of GAs is their ability to use accumulating information about initially unknown search space in order to bias subsequent searches into useful subspaces.

GAs differ from conventional nonlinear optimization techniques in that they search by maintaining a population (or data base) of solutions from which better solutions are created rather than making incremental changes to a single solution to the problem. Thus, they are less susceptible to the local optimum problem experienced by traditional nonlinear programming methods, in particular hill-climbing algorithms like those of TRANSYT. In addition, many optimization techniques are calculus-based and depend on the restrictive re-

quirements of continuity and function derivatives. GAs use only objective function information; thus, they do not need derivatives of the objective function and can work with noisy and discontinuous functions.

GAs are iterative procedures that search by allowing a population (data base) of alternative solutions to reproduce and cross among themselves with biases allocated to the most fit members of the population. Possible solutions within each population are coded as binary strings (chromosomes). For example, if a solution to a certain problem takes a maximum value of, say, 31, it can be coded as a five-bit string (e.g., 11011 for 27).

A solution to a signal optimization problem consists of a cycle length, offsets, phase sequences, and phase splits; thus it requires a multiparameter solution. To code a multiparameter solution as a binary string, each parameter (cycle length, offset, etc.) is first coded as a binary variable as described above. Then these single parameters are concatenated to obtain the required binary string. For example, if six-bit, three-bit, and six-bit binary variables are selected to represent cycle length, phase sequence, and offset, respectively, the binary string representing the solution would be as follows (assuming that all bits in the solution are zeros):

000		000	000000		000000	000000
*n*th	. . .	1st	*m*th	. . .	1st	Cycle
sequence		sequence	offset		offset	length

where m and n are the number of offsets and sequences to be optimized, respectively.

A set of solutions is selected at random in the initial population using "successive coin flips" (heads = 1, tails = 0). Thereafter, during each iteration step, called a generation, the solutions of the current population are evaluated using an objective function, and on the basis of this evaluation, a new population of candidate solutions is formed (see Figure 1). The objective function used in the process is model specific. For example, in signal timing optimization this could be fuel consumption, bandwidth, PROS, or a combination of the foregoing.

The result is a new set of solutions ("offspring"), and the new solutions are more fit (that is, have a better objective function value) than the parent solutions from the previous generation. The GA continues to generate successive populations until some solution criterion is met. This could be reaching a fixed maximum number of generations or being unable to achieve further improvement in the objective function.

Three basic components are required for GA implementation:

1. A scheme to allow for a binary string (a string of 0's and 1's) representation of alternative solutions to the problem;
2. An objective function to evaluate the fitness of each solution; and
3. Genetic operators that mimic the biological evolution process (these operators are used in the formation of successive populations).

A simple GA that produces good results in many problems consists of three genetic operators: reproduction, crossover, and mutation. Further improvements to GAs have been suggested in the literature and more advanced operators and techniques have been implemented (*15*). These include dominance, inversion, intrachromosomal duplication, deletion, translocation, segregation, niche exploitation and speciation, migration, marriage restriction, and sharing functions.

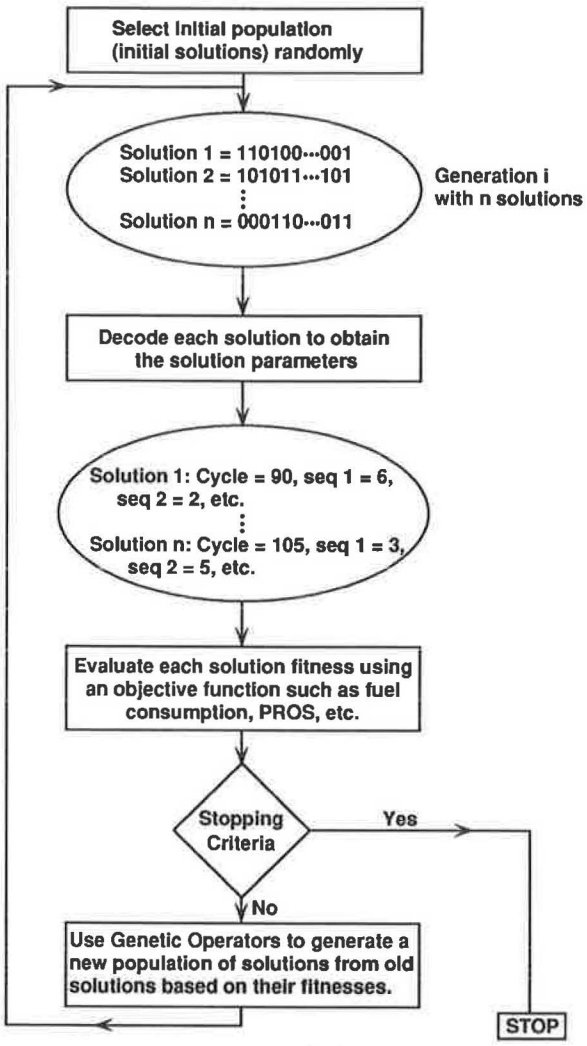


FIGURE 1 Diagram explaining the possible use of a GA in signal timing optimization.

A simple GA that uses the three basic operators (reproduction, crossover, and mutation) is employed in this study. The following is a brief description of the GA operators and techniques used.

Reproduction

Reproduction is the process of selecting those solutions (individual binary strings) within a given generation that will be allowed to contribute to the next generation. The procedure randomly selects these individuals on the basis of their objective function values (fitness). Strings with higher fitness have a higher probability of contributing one or more offspring in the next generation. Thus, the effect of reproduction is to bias the population to contain more-fit members. The

reproduction process is repeated until the number of the selected strings equals the specified population size. The population size is an input to the model.

Crossover

Crossover randomly chooses two individuals from the population selected using the reproduction operator explained in the previous section. These two individuals (parents) are then "mated" to produce two offspring. This is done by choosing a random number K between 1 and the string length (L) less 1 as a "crossover point." The first K bits of Parent 1 and the last L-K bits of Parent 2 are joined to form one offspring. Similarly, the first K bits of Parent 2 and the last L-K bits of Parent 1 are joined to form the other offspring. This procedure is repeated until there is a new population that has the same number of individuals as the old population (the population size).

For example, assume that the string length in a given problem is 8 and the two selected parents are

Parent 1 = 01101011
Parent 2 = 11010100

If the random number selected between 1 and L-1 is 3, then the resulting crossover yields the following offspring, as illustrated by the underlined substrings above for Child 1 and not underlined for Child 2:

Child 1 = 01110100
Child 2 = 11001011

Crossover probability (CP) is a GA parameter that controls the frequency with which the crossover operator is applied. In each new population $CP \cdot N$ individuals undergo crossover, where N is the population size. If CP is too high, good solutions are discarded faster than selection can produce improvements. If the crossover rate is too low, the search may stagnate (*18*).

Mutation

This operator, with a probability equal to the mutation probability (MP), randomly alters the value of a string position (0 to 1 and vice versa). The function of this operator is to prevent any given bit position from remaining forever fixed to a single value over the entire set of solution alternatives.

MP is a parameter that controls the probability with which a given string position alters its value. Small values of MP are always used because a high level of mutation results in an essentially random search (*18*).

Scaling

At the start of a GA exercise, it is common to have a few solutions with much higher fitness than other solutions. This causes the reproduction process to produce a population with a high proportion of the high-fitness individuals, causing premature convergence.

Late in the exercise, a different problem arises. At this near-optimization stage, the population average fitness may become close to the maximum fitness, although there may still be significant diversity within the population. This leads the reproduction process to select the same number of average and best individuals, leading the process to a random walk.

To deal with these two problems, scaling of the objective function has been suggested to keep appropriate levels of competition throughout the process. Three linear scaling mechanisms have been used. These are linear scaling, sigma truncation, and power law scaling (*15*).

MODEL IMPLEMENTATION

A simple GA similar to the one described in the previous section was used in conjunction with the TRANSYT-7F hill-climbing procedure to optimize signal timing. In this scheme, the main objective of the genetic search was to identify the optimal phase sequences, cycle length, and (possibly) the initial offsets. TRANSYT-7F was then used normally to adjust cycle length within a narrow cycle length range and to optimize splits and offsets.

The objective function used by the GA in selecting the optimal phase sequences was the PROS value calculated by TRANSYT-7F release 7. Other objective functions such as the TRANSYT-7F DI either alone or in combination with PROS could be used instead; however, DI calculation requires executing the TRANSYT-7F simulation model, which takes a much longer time than the PROS calculation.

In this study, a separate computer program was written to exercise the GA and call TRANSYT-7F whenever the hill-climbing procedure, the PROS calculation, or both were needed. Two different implementations of the hybrid GA model were tested in this study, as described below.

Method 1: Concurrent Use of GA and TRANSYT-7F Hill-Climbing

In Method 1, each alternative solution within a GA generation consisted of a phase subsequence for each arterial intersection and a cycle length. Offsets and splits for these solutions were calculated using the TRANSYT-7F hill-climbing procedure and the resulting PROS were used as the solution fitness values. The splits were calculated by the TRANSYT-7F internal initial timing routine based on equalizing the degrees of saturation on the critical conflicting links and were fixed.

The first step in using a GA is to choose a scheme for coding each possible solution as a finite-length binary string. In this implementation, each binary string consisted of a six-digit binary number representing a cycle length and a series of three-digit binary numbers representing arterial phase subsequences on all arterial intersections (defined below). An arterial phase subsequence is defined as the order of all phases serving arterial movements. If all the bits in a coded string were 0's, this string would be as follows:

0 0 0		0 0 0	0 0 0	0 0 0 0 0 0
*n*th	. . .	2nd	1st	Cycle
sequence		sequence	sequence	length

The resultant string was of length $6 + (3 \cdot n)$, where n is the total number of subsequences to be optimized. A fixed number of these strings had to be selected for each GA generation. It has been reported (*15*) that even for very large and complex search spaces, GA can rapidly locate structures with high fitness ratings using a generation population of 50 to 100 alternative solutions (binary strings). The population size (the number of alterative solutions within each generation) used in this study was 50.

The selection of these strings was random in the initial generation. In every generation thereafter, the three basic genetic operators (reproduction, crossover, and mutation) were used to create new strings based on the fitness values. As described above, the fitness values used in this implementation were the PROS values obtained from TRANSYT-7F optimization of offsets for each alternative solution within a GA generation. To perform these runs, a procedure was developed to decode each binary string as a cycle length and phase subsequences, which were then used as inputs to TRANSYT-7F.

The six bits representing cycle length were first converted to an integer (N) between 0 and 63. This number was then mapped into an integer between the minimum and maximum cycle lengths as follows:

$$\text{Cycle} = \text{Min} + N \cdot (\text{Max} - \text{Min})/63 \quad (1)$$

where Min and Max are the minimum and maximum allowable cycle lengths, respectively. This number was then rounded into a multiple of 5 sec.

The three-bit binary numbers were first converted to integers between 0 and 7. Each number was then mapped into a phase subsequence using Table 1. (Note that more exotic phasings such as split phasing were not considered in this study, but can be easily added.)

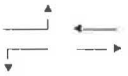
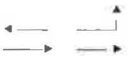
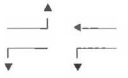
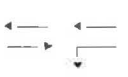
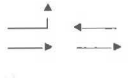
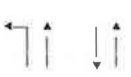
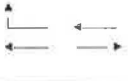
After each TRANSYT-7F run, the GA reads the resultant PROS value from the graphic data file (GDF) produced by TRANSYT-7F. Before these PROS values were used by the GA to evaluate different alternatives, they were scaled using a linear scaling function. This scaling procedure was suggested to improve the GA performance as described above.

In this study, each GA run was executed for a maximum of 50 generations. After the GA had been exercised for the maximum number of generations, the set of phase sequences that produced the highest PROS value was selected as the best phasing.

Method 2: Use of GA Followed by TRANSYT-7F Hill-Climbing

Method 2 differed from Method 1 in that the GA was extended to optimize offsets in addition to cycle length and

TABLE 1 Look-Up Table To Transform Three-Bit Binary Numbers to Phase Subsequences

Binary Number	Integer	Phase Subsequence E-W Artery	Phase Subsequence N-S Artery
000	0		
001	1		
010	2		
011	3		
100	4		
101	5		
110	6		
111	7		

phase sequence. This means that each alternative solution within a GA generation consisted of a cycle length, phase sequences, and offsets.

TRANSYT-7F evaluation runs were performed for each of these solutions to determine their fitness (PROS). The splits in these runs were calculated using the TRANSYT-7F initial timing routine. Offset optimization was performed by the GA, not TRANSYT-7F; thus the computational requirement of this method was much lower than that of Method 1.

In this implementation, each binary string consisted of a six-digit number representing a cycle length, a series of three-digit binary numbers representing phase subsequences, as in Method 1, and a series of six-digit numbers representing offsets for all arterial intersections. For example, if all the bits in a string were 0's, this string would be as follows:

000		000	000000		000000	000000
*n*th	. . .	1st	*m*th	. . .	1st	Cycle
sequence		sequence	offset		offset	length

Each string was of length $(3 \cdot n) + (6 \cdot m) + 6$, where n and m are the number of subsequences and offsets to be optimized, respectively.

The cycle length and phase sequences were decoded from these strings as described in Method 1. To decode offsets, each six-bit binary number representing an offset was first transformed to an integer (M) between 0 and 63. Then it was mapped into an integer between 0 and 100 representing the offset as a percentage of cycle length as follows:

$$\text{Offset} = M \cdot 100/63 \tag{2}$$

This value was then converted into seconds and used together with cycle length and sequences as inputs to TRANSYT-7F.

After the GA had been executed for the maximum number of generations, the four solutions with the highest fitness (PROS) values were located. The TRANSYT-7F hill-climbing routine was then used to adjust the offsets and cycle length within a small cycle-length range. The solutions from these runs were compared and the phase sequences that produced the highest PROS value were selected as the optimal phase sequences. In effect, this method uses the GA to locate the peaks and the TRANSYT-7F hill-climbing procedure to climb them.

Other than the differences mentioned above, all other aspects of Method 2 were the same as those of Method 1.

MODEL APPLICATIONS

Three real-world traffic systems were used to evaluate the two implementations of the hybrid GA discussed in this paper. These systems were Cape Coral Parkway, a seven-intersection artery in the city of Cape Coral, Florida; Volusia Avenue, a 12-intersection east-west artery in the city of Daytona Beach, Florida; and a 12-intersection grid network in the city of Daytona Beach, Florida. For the Daytona Beach network, the objective was to select the phase sequences for two intersecting arteries within the network. These were Ridgewood Avenue, a four-intersection north-south artery, and Volusia Avenue, a three-intersection east-west artery.

In most cases, the existing phase sequences were leading dual lefts without overlap. For the purpose of this study several permitted-only left turns were changed to protected, even though they were not warranted, to provide multiple phasing.

In this comparative study, the splits used were always those calculated by the TRANSYT-7F internal initial timing routine, and they were held constant. For all systems investigated, the cycle range was 100 to 120 sec.

The population size (the number of alternative solutions within each generation), maximum number of generations, crossover probability, and mutation probability used in these studies were 50, 50, 0.90, and 0.01, respectively, based on previous GA research.

The resultant designs from the two GA implementations were compared with those obtained using TRANSYT-7F optimizations with both the "existing" phase sequences and the phase sequences selected by PASSER II-90. In all cases, the objective function used in the optimization was the PROS value calculated by TRANSYT-7F. The comparison was based on perceived progression as measured by the PROS and bandwidth efficiency.

Figure 2 shows the variation of the population maximum fitness and the population average fitness over the generations of evolution when Method 1 was used. For all three systems investigated, this method was able to locate good solutions after only a few GA generations. The solution with the highest fitness could be achieved after 10 to 24 iterations (generations) depending on the system investigated. This means that 500 to 1,200 TRANSYT-7F PROS-only optimization runs were required. This is very efficient compared with the implementation of enumerative schemes in which every possible sequence combination is tried. In that case, the number of possible combinations for a 12-intersection artery and eight possible phase sequences is $(8^{12} = 6.87 \cdot 10^{10})$.

PROS-only optimization is very quick compared with traditional TRANSYT-7F DI optimization. On a 20-MHz 80386 machine with a math coprocessor, it took less than half a second per intersection. For a 12-intersection artery, each PROS optimization took about 6 sec; thus, 1,200 optimization runs would require less than 2 hr. This time can possibly be reduced by further improvement of the GA model as discussed below and by using a quicker TRANSYT-7F optimization that uses fewer optimization steps.

Figure 3 shows the increase in the population maximum PROS and the population average PROS over the GA generations when using Method 2. In this method, after the GA had been run for 50 generations, TRANSYT-7F was used to adjust the resultant timing. This adjustment produced additional improvements in PROS, which can be observed by comparing the maximum PROS values that could be achieved using the GA alone (from Figure 3) with the best PROS values produced by Method 2, as reported in Table 2. These improvements were 13 percent (43 versus 38), 26 percent (29 versus 23), and 9 percent (38 versus 35) for the three cases.

Method 2 is much more efficient than Method 1. The reason for this is that TRANSYT-7F was used only to calculate the PROS for each alternative solution rather than optimizing the offsets. It is expected that if the PROS values were calculated internally without a need to call TRANSYT-7F externally, the time required to run 50 generations (2,500 PROS calculations) would be less than a minute for a 12-intersection artery.

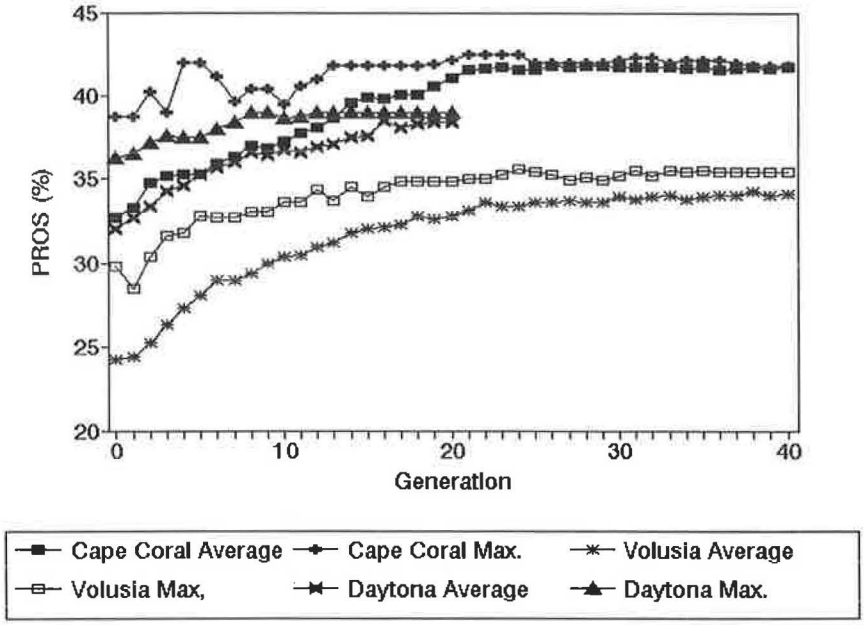


FIGURE 2 Improvement in PROS when optimized using Method 1 with 0.90 crossover probability.

It can be observed from Figure 3, however, that after reaching a certain point in the optimization, the maximum PROS value did not always increase from generation to generation with Method 2. In these cases, the PROS value did not show a clear trend of improvement late in the GA process. This is true for at least two of the three examples and suggests that further improvements might be needed for Method 2.

Table 2 and Figure 4 indicate that using Method 1 or Method 2 to select signal phasing could result in significant improvements in PROS compared with using the existing phase sequences. Method 1 improved PROS by 27 percent (42 versus 33), 44 percent (36 versus 25), and 18 percent (39 versus 33) for the three cases. Method 2 produced 30 percent (43 versus 33), 16 percent (29 versus 25), and 15 percent (38 versus 33) PROS improvements, respectively.

In addition, using Method 1 or Method 2 to optimize phasing produced higher bandwidth efficiency compared with existing phase sequences. For example, Method 1 increased the bandwidth efficiency by 209 percent (34 versus 11) and 138 percent (31 versus 13) for Cape Coral Parkway and Volusia

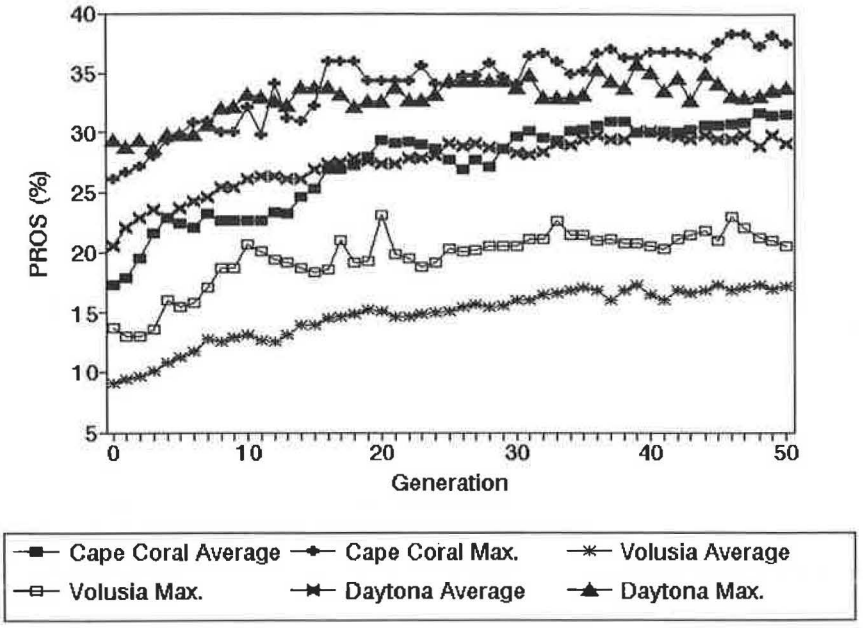


FIGURE 3 Improvement in PROS when optimized using Method 2 with 0.90 crossover probability.

TABLE 2 Comparison of the Results Obtained Using TRANSYT-7F To Optimize PROS with Different Phase Sequences

System	Sequence Source	Artery	Crossover Probability	Effective PROS (%)			Bandwidth Efficiency (%)		
				Right[a]	Left[a]	Average	Right[a]	Left[a]	Average
Cape Coral Parkway	Existing	1	0.90	40	26	33	21	0	11
	PASSER II	1	0.90	31	51	41	22	45	33
	Method 1	1	0.90	38	47	42	27	41	34
	Method 2	1	0.90	35	50	43	27	45	36
	Method 2	1	0.65	42	35	38	38	19	29
Volusia Avenue	Existing	1	0.90	24	25	25	18	7	13
	PASSER II	1	0.90	30	35	32	24	31	28
	Method 1	1	0.90	34	37	36	30	32	31
	Method 2	1	0.90	27	32	29	0	27	14
	Method 2	1	0.65	34	33	34	30	26	28
Daytona Beach Network	Existing	1	0.90	31	37	33	15	28	21
		2		34	28		25	4	15
	PASSER II	1	0.90	45	33	37	45	26	35
		2		31	36		17	33	25
	Method I	1	0.90	42	42	39	33	38	35
		2		32	32		22	18	20
	Method 2	1	0.90	43	42	38	34	38	36
		2		38	22		36	0	18

[a] Right and Left refers to the right-bound and left-bound travel on the artery.

Avenue, respectively. For the two arteries of the Daytona Beach network, Method 1 increased the bandwidth by 66 percent (35 versus 21) and 33 percent (20 versus 15), respectively.

The two GA methods were also compared with using PASSER II for phase sequence optimization on individual arteries to determine whether they could be as effective. The results are shown in Table 2 and Figure 4. Method 1 was able to produce better PROS values in all cases. The improvements achieved were 2 percent (42 versus 41), 12 percent (36 versus 32), and 5 percent (39 versus 37) for the three systems investigated. In addition, Method 1 was able to produce good solutions in terms of bandwidth efficiency relative to PASSER II solutions, as shown in Table 2.

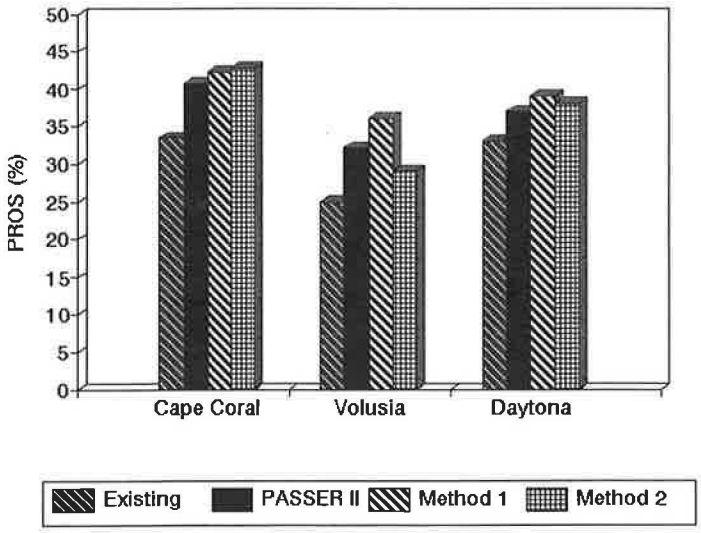


FIGURE 4 Comparison of maximum PROS achieved using different phase sequence sources.

Method 2 was able to select solutions with higher PROS values compared with PASSER II solutions in only two of the three systems investigated (Cape Coral Parkway and Daytona Beach network). In the Volusia Avenue example, Method 2 could not reach as good a solution as that produced by PASSER II, again suggesting that Method 2 needs to be improved.

A sensitivity analysis was performed to examine the effect of varying crossover probability (CP) on Method 2 performance. It was thought that a lower value of this parameter might improve the performance of the model by preventing premature convergence. Table 1 shows that reducing CP from 0.90 to 0.65 increased PROS by 17 percent (34 percent versus 29 percent) for Volusia Avenue; however, it also reduced PROS for Cape Coral Parkway by 10 percent (38 versus 42). This indicated that varying genetic operators and parameters might have important effects on the performance of the model.

CONCLUSIONS AND RECOMMENDATIONS

From the results of this study, it can be concluded that the GA method has potential for use in signal phasing and timing optimization. It appears that the concurrent use of the GA and TRANSYT-7F (that is, Method 1) can optimize the phase sequences for arterial streets as well as multiarterial networks.

Using the GA followed by TRANSYT-7F as in Method 2 can produce as good results as those produced by Method 1 or PASSER II only in some of the cases; however, this method is much more efficient than Method 1 in terms of execution time. Several suggestions for improving the performance of Method 2 might include

1. Varying various GA parameters, including crossover probabilty, mutation probability, population size, string length, and objective function scaling strategy. Lower values of crossover probability might produce better results.
2. Using advanced GA operators.

Implementation of the hybrid GA internally in the TRANSYT-7F program should be investigated further. This will give TRANSYT-7F the ability of simultaneous optimization of all signal timing elements. Further research is needed to reduce the execution time of this strategy, particularly Method 1. Further research is also needed to improve the general optimization process of TRANSYT-7F, because this would also improve its performance with respect to phasing optimization.

REFERENCES

1. Chang, E. C. P., and C. J. Messer. *Arterial Signal Timing Optimization Using Passer II-90—Program User's Manual.* Texas Transportation Institute, Texas A&M University System, College Station, 1991.
2. Wallace, C. E., E. C. P. Chang, C. J. Messer, and K. G. Courage. *PASSER II Users' Guide,* Vol. 3. FHWA, U.S. Department of Transportation, 1991.
3. Wallace, C. E. *MAXBAND Users' Manual.* Transportation Research Center, University of Florida, Gainesville, 1987.
4. Wallace, C. E., K. G. Courage, and M. A. Hadi. *TRANSYT-7F Users' Guide*, Vol. 4. FHWA, U.S. Department of Transportation, 1991.
5. Rogness, R. O., and C. J. Messer. Heuristic Programming Approach to Arterial Signal Timing. In *Transportation Research Record 906,* TRB, National Research Council, Washington, D.C., 1983, pp. 67–75.
6. Cohen, S. L. Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In *Transportation Research Record 906,* TRB, National Research Council, Washington, D.C., 1983, pp. 81–84.
7. Cohen, S. L., and C. C. Liu. The Bandwidth-Constrained TRANSYT Signal-Optimization Program. In *Transportation Research Record 1057,* TRB, National Research Council, Washington, D.C., 1986, pp. 1–8.
8. Skabardonis, A., and A. D. May. Comparative Analysis of Computer Models for Arterial Signal Timing. In *Transportation Research Record 1021,* TRB, National Research Council, Washington, D.C., 1985, pp. 45–52.
9. Cohen, S. L., and J. R. Mekemson. Optimization of Left-Turn Phase Sequence on Signalized Arteries. In *Transportation Research Record 1021,* TRB, National Research Council, Washington, D.C., 1985, pp. 53–58.
10. Chang, E. C. P., S. L. Cohen, C. Liu, N. A. Chaudhary, and C. J. Messer. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequence in Multiarterial Closed Networks. In *Transportation Research Record 1181,* TRB, National Research Council, Washington, D.C., 1988, pp. 61–67.
11. Hadi, M. A., and C. E. Wallace. A Progression-Based Optimization Model in TRANSYT-7F. In *Transportation Research Record 1360,* TRB, National Research Council, Washington, D.C., 1992.
12. Hadi, M. A., and C. E. Wallace. Major Enhancements to TRANSYT-7F. Presented at the Fourth International Conference on Microcomputers in Transportation, American Society of Civil Engineers, Baltimore, Md., 1992.
13. Wallace, C. E., and K. G. Courage. *AAP Users' Guide,* Vol. 2. FHWA, U.S. Department of Transportation, 1991.
14. Chaudhary, N. A., A. Pinnoi, and C. J. Messer. Proposed Enhancements to MAXBAND 86 Program. In *Transportation Research Record 1324,* TRB, National Research Council, Washington, D.C., 1991, pp. 98–104.
15. Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, Mass., 1989.
16. Foy, M., R. F. Benekohal, and D. E. Goldberg. Signal Timing Determination Using Genetic Algorithms. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992.
17. Wallace, C. E., and K. G. Courage. Arterial Progression—New Design Approach. In *Transportation Research Record 881,* TRB, National Research Council, Washington, D.C., 1982, pp. 53–59.
18. Grefenstette, J. J. Optimization of Control Parameters for Genetic Algorithms. *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-16, No. 1, 1986, pp. 122–128.

Publication of this paper sponsored by Committee on Traffic Signal Systems.