Title

Optimizing Comprehension of Changeable Message Signs (CMS)

Author
Greenhouse, Daniel

Publication Date
2007-11-01

CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Optimizing Comprehension of Changeable Message Signs (CMS)

Daniel Greenhouse
University of California, Berkeley

California PATH Research Report
UCB-ITS-PRR-2007-24

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation, and the United States Department of Transportation, Federal Highway Administration.

The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California. This report does not constitute a standard, specification, or regulation.

Final Report for Task Order 5203

November 2007

ISSN 1055-1425

Final Report

Optimizing Comprehension of Changeable Message Signs (CMS)

Prepared for

Caltrans TO#5203

California PATH

By

Visual Detection Laboratory
University of California, Berkeley
360 Minor Hall
Berkeley, CA 94720

November 15, 2007

Final Report

Optimizing Comprehension of Changeable Message Signs (CMS)

Note on Figures

Many of the figures in this report contain important color information which is not reproduced in the printed version of this document. Please see the PDF format document on the attached CD, or available online, to view the figures in color.

Executive Summary

The goal of this research was to assist the California Department of Transportation (DOT) in optimizing the message content and presentation within changeable message signs (CMS). Optimized content will improve information transfer while at the same time minimizing the likelihood of congestion owing to slowing by motorists attempting to read the message. The research was restricted to simulated signs displaying 16 characters in each of three lines, representing permanent CMS displays, or signs containing only 8 characters in each of three lines, as is the case for portable CMS displays. While all information can be contained in a single screen for the permanent signs, multiple screens are often required for the portable CMS displays. This study specifically focused on "early vision" which is distinct from "cognitive processes". Early vision problems are those relating to the limitations of the first several stages of the visual system. An example question of early vision is whether flashing the letters 'NB' on a CMS message is more effective than leaving them steadily on. For these questions of recognition, accuracy is used as a measure of intelligibility. Cognitive process limitations, by contrast, involve events that are due to thinking on the part of the observer. An example of a question involving a cognitive limitation is whether observers interpret the abbreviation 'NB' as meaning 'north bound'.

The project was divided into two phases, a laboratory phase and then a field test.

In the laboratory portion of the project, experiments consisted of a computer-based simulation that tested the ability of licensed drivers to reproduce the information contained in simulated CMS messages while simultaneously preoccupied with a separate, "distracter" task akin to lane keeping. Each sign both expanded and moved across the observer's visual field in a manner similar to that which would occur during approach at highway speeds of 65MPH. The computer controlled CMS display program quantified a driver's ability to stay within a simulated lane while measuring the accuracy of the driver's response (comprehension of message content) as a function of various manipulations of the CMS message display including:

- timing of successive frames (on times and off times) – portable CMS only
- spatial structure of the message display – permanent CMS only
- movement vs. steady presentation of the message display – permanent and portable displays
- changing one element within the message display – permanent CMS only
- compacting the CMS information – permanent and portable CMS

The accuracy of the reproduced message content was quantified by the percentage of words correct (WC) and the percentage of numbers correct (NC). Word percentage correct measures the number of words that the observer got correct divided by the possible number of words contained within the CMS message frames for all trials in a given test. Number percentage correct measures the number of *digits* that the observer got correct divided by the possible number of digits for all CMS messages in a given test. For example, in one trial of a given test RDWK 101 SAN MATEO & 3RD AVE would

count as 4 word elements and 4 number elements. Nine observers (6 young and 3 old) whose ages ranged from 20 – 74 and whose vision was fully corrected were tested in the experiments. The accuracy of lane keeping was measured by the percentage of time that the observer kept within the lane.

The first set of experiments examined the effect of inserting a blank “off screen” between successive frames in a portable CMS. The duration of the off period between frames ranged from 0 milliseconds (ms), which means no blank screen (the current standard advised by the California DOT), to 500 ms. The results indicated the greatest improvements in WC and NC for off screens lasting for a duration of 300 ms. The addition of a 300 ms off screen caused the average WC score for the younger observers to improve from 70% to 82%, as compared no off screen ($p < 0.01$, paired t-test). The WC score for the older observers went up to 59% with the 300 ms off screen from 49% with no off screen ($p < 0.05$). Similarly, the NC score for the young observers improved from 58% to 71% ($p < 0.05$, t-test). The NC scores for the older observers showed little change with the insertion of an off screen and the magnitude of the change was not statistically significant. On balance though, these findings suggest that the insertion of blank screen between frames of a two-frame message enables the observer to better process the information content.

The next set of experiments examined the effect of varying the spatial structure of a permanent CMS. The message content was presented left-justified (LJ), center-justified (CJ), right-justified (RJ), and what we term staircase-justified (SJ). In the SJ case, the top row is LJ, the middle row is CJ and bottom row is RJ. Observers were better able to reproduce the information contained in LJ and SJ geometries than for the current standard of CJ displays. The average WC score of the younger observers improved from 76% to 81% for the LJ geometry and to 86% for the SJ configuration ($p < 0.05$ and $p < 0.01$ respectively, t-test). The LJ geometry did not show a statistically significant improvement for the older observers in the case of their WC score but the SJ geometry did, improving from 59% to 76% ($p < 0.05$). The average NC score for younger observers improved from 69% to 74% for the LJ case and to 79% for the SJ case ($p < 0.05$ for LJ and $p < 0.01$ for SJ, t-test). The average NC score for the older observers did not show any statistically significant improvement. Still, the overall results indicate that both left and staircase justification are likely to improve an observer’s ability to reproduce information content contained within a CMS. It is likely that the geometric alignment of the staircase-justified text fits the natural progression of an observer’s eye movements as he tries to read the content from a message

The third set of experiments tested whether artificially expanding the text contained within a CMS message has an effect on intelligibility. CMS displays with looming text (in which the physical size of the letters within the display is gradually increased) were tested against displays in which the size of the text is held constant (with respect to the overall size of the CMS). For portable CMS, *none* of the WC or NC scores for older or younger observers showed statistically significant improvement with one exception: a looming/300 ms off-time combination showed a WC score for younger observers of 78% improving from 70% for the standard—no looming, no off-time ($p < 0.05$). But since a

300 ms off-time alone caused the WC score to improve even more, to 82% for younger observers, this is not a positive result for looming. The preliminary tests on looming for permanent CMS displays showed no statistically significant improvements in WC or NC scores. The results taken in total indicate that artificial looming in the message does not improve the observer's ability to reproduce the information contained within the message and may possibly degrade the intelligibility of these messages.

The fourth set of experiments tested the effect of varying one specific element like a number or letter on a permanent CMS. These tests were done to see if the change blindness phenomenon occurs on CMS signs when a blank screen is used in between a two-frame message. The change blindness phenomenon refers to an inability to perceive that some elements in a scene have changed, and would normally be expected to be absent when no blank screen is imposed (which is to say that changes would be easy to perceive). The results showed that the six younger observers were able to correctly identify that the sign changed 84% of the time (on average) while the three older observers could correctly identify sign changes 62% of the time; however, intelligibility scores generally declined. The NC score for the older group was the only score showing improvement when change blindness was invoked.

The fifth experiment tested whether abbreviating words on a CMS in order to fit the necessary information into the limited space can affect message intelligibility. CMS displays with compacted words and common abbreviations were tested against CMS displays that had no abbreviation of the words. For permanent and portable CMS displays, both the WC and NC scores did not show statistically significant improvement when comparing the compacted CMS message with the unabbreviated version. In fact, the results indicate that the compacted CMS usually impaired the observers' ability to reproduce the informational content of the CMS message. Taken in total, the abbreviated CMS display lessened the intelligibility of the CMS by causing observers to stop reading the rest of the message because they did not understand the abbreviations.

A final set of laboratory experiments, slightly different in nature, was conducted later than those mentioned above using a set of 10 observers. Although most of this new set was younger, there was no breakdown by age used for analysis. Each trial they underwent had one of two types of messages. The message would either show a truck license plate in a message such as AMBER ALERT CA LIC 9L14317 or it would show a direction of East or West in a message such as DETOUR GO E ON I-580. The only variation between trials in a "directional" message was the direction ('E' or 'W'). The only variation between trials in a "license" message was the license plate itself. Observers were only asked to recall the direction ('E' or 'W') or the license plate; no other parts of the messages were required to be recalled. The goal was to see if the recall of license plates was as good as the recall of the single-character directional messages. It was not, thereby strongly suggesting that license plates have greater "informational content" than directional messages. This experiment was performed in order to clarify the concept of the "informational unit", as presented elsewhere in the literature. Correlation between distracter task performance and license plate recall was also examined.

The second part of this project involved field testing. A duplication of the full set of laboratory tests in the field situation was impractical because (1) the number of test configurations in the laboratory was very large, (2) there was no practical way to query or survey drivers in the field after they had viewed a CMS message, and (3) we could not allow any field test to compromise the safety of the driving public, thus limiting the type of message we could present. Thus we concluded that the optimization of content and presentation of a CMS message would have to be solely based on the in-lab tests, and the field work reserved for testing the goal of minimal disruption to traffic flow under changed message conditions.

Due to limitations imposed by the government transportation agencies involved, and by the University of California Committee for Protection of Human Subjects, we limited testing to a single text message and a single novel message configuration. The message chosen was *"REPORT DRUNK DRIVERS CALL 911"*. The point of the field testing was to see if the novel message configuration had any significant impact (positive or negative) on traffic flow. The field test was conducted along a stretch of westbound Interstate 80 that constitutes part of the Berkeley Highway Lab, a 2.7 mile portion of I-80 that has been outfitted with large quantity of instrumentation (video monitoring and loop detectors) in order to learn about traffic properties and serve as a test bed for research in transportation.

We presented our test message in two different configurations, standard (center-justified) and novel (staircase-justified, which we had judged as optimal on the basis of our laboratory tests). These were presented for multiple ten-minute intervals, always separated by a five-minute period in which no message was presented. Data was analyzed for five lanes, at two different stations relative to the CMS, and at two different times of day (daylight and nighttime). Rigorous pair-wise statistical tests were performed on the data. Only insignificantly small differences in vehicle behavior were discerned between the two message configurations. We concluded that there is no significant disruption to traffic flow when a CMS displays a message in a presentation that we consider optimal on the basis of our laboratory tests.

Keywords

CMS, changeable message sign, message, frame, looming, justification, blank screen, movement, compacting, comprehension, recall, change blindness, informational unit, traffic flow

Laboratory Studies Report

Optimizing Comprehension of Changeable Message Signs (CMS)

Prepared for

**Caltrans TO#5203: Optimizing the CMS
California PATH**

By

**Visual Detection Laboratory
University of California, Berkeley
360 Minor Hall
Berkeley, CA 94720**

**Theodore Cohn
Khoi Nguyen
Kent Christianson
Daniel Greenhouse
Lance Tammero**

Laboratory Tests

Abstract

The goal of this research was to optimize the message content and presentation within changeable message signs (CMS) so as to maximize comprehension of those messages with minimal disruption to traffic flow. Tests were done on nine observers (segregated by age cohort) to determine word and number comprehension, as measured by recall, after those observers had viewed CMS messages with a variety of configurations under simulated driving conditions. Interframe blanking time ("off-time") for portable CMS, spatial structure of the message for permanent CMS, looming for both permanent and portable CMS, change blindness for permanent CMS, and message compaction for permanent and portable CMS were all examined for their effects on intelligibility. Introduction of interframe off-time and certain changes in message spatial structure had a positive effect on intelligibility; message compaction and change blindness had a negative effect on intelligibility. Looming had a neutral or possibly negative effect on intelligibility.

A later series of tests involving ten observers, who were not divided by age, used the same CMS and driving simulation to investigate recall as an inverse measure of information content. These tests took the straightforward viewpoint that *greater* recall corresponds to *lesser* information content. The recall of seven-character truck license plates was compared to the recall of single letter direction indicators ('E' or 'W'). This quantification of "informational units" and the corresponding test results showed partial agreement with other conceptual measures of information content but also showed divergence in some respects.

Introduction / Background

The work zone (WZ) is a key location of attempts to improve the California highway transportation system. It is also the location of a disproportionate number of collisions. Safety authorities are thus naturally interested in pursuing means of improving safety in this setting. The CMS, whether in portable form or as a fixed installation, is potentially a useful tool on California highways. Messaging in this medium can be used to enhance safety and reliability on the highway by assisting motorists to gracefully alter travel patterns around the WZ and to reduce delays that presently spring from this cause. Performance of the California transportation system relies on the ability to be able to repair or to install innovative solutions to pressing problems and to minimize disruption while doing so. The CMS is one of a series of ITS tools that has potential to assist in these efforts. But a recent study of CMS effectiveness has revealed a number of areas requiring additional research. A recent White Paper (Dudek, 2002) concerning the operation of Changeable Message Signs (CMS) identifies a number of unanswered questions concerning their optimal use both in a WZ and in general use.

The use of the CMS in or near a work zone is a significant advance as it offers the planner a flexible means of communicating with the driving public in a timely and focused manner. Explicit information, including instruction and suggestions can be conveyed using a CMS. However, the CMS medium is problematic in several ways. First its location, while under the control of the operator, is not necessarily changeable with rapidity. Next, its format requires brevity of message. Finally, complexity of message, even if brief, can have an effect on traffic all by itself, as for example by slowing traffic just so that passers-by have an opportunity to assimilate the message.

The CMS is a seriously constrained medium for communication. A typical CMS will have the ability to display 3 lines of sixteen characters (letters or numbers) each but no graphical pictures. Early research was used to argue against graphics. For example, Pretty and Cleveland (1970) showed that signs based upon diagrams were misinterpreted and later, Dudek and Jones (1971) reported a preference by motorists for word messages.

Beyond a preference for words, there is a subtler issue of how many characters to use to convey the words. Not surprisingly, it has been shown that fewer characters make signs more intelligible (Dudek, et al., 1981a). It will be appreciated that motorists will have limited time during which to grasp the message on the CMS as they are driving past its location. In addition, some messages, owing to their complexity, require more that one ‘frame’ of information. CMS users have attempted to introduce codes as a way of shortening character count. But motorists have proven themselves unable to learn the code (Stockton, et al., 1976). This may be due in part to our increasingly illiterate population (Proffit and Wade, 1998), to say nothing of the barrier imposed by multilingualism and the normally occurring spectrum of individual preferences (Miller et al., 1995) and individual needs (Huchingson et al., 1977). Work zone deployment of the CMS carries with it some very specific problems. Huchingson et al. (1977) showed that displaying the speed that California drivers would encounter ahead in a WZ caused them to slow prematurely. In any case it is recommended that the design of messages adhere to recommendations in the New Jersey DOT *Variable Message Sign Operations Manual* (Dudek 2001).

Exposure Duration: A key variable available to users of the CMS is the time during which each of two frames is left on (use of more than two frames is vigorously warned against. Another variable is the time during which the sign is entirely off (between frames). Little mention of this appears in the literature. Effects of the off-time are explored within this study.

***Flashed messages**:* Very little attention has been paid to developing strategies to give the text on a CMS additional attention-getting capability. Flashing is a traditional technique of supplying emphasis or of attracting attention, but if the cost is longer time to acquire, then the cost is too great in many circumstances, particularly those where the message requires two full frames to convey. However, there are many other ways by which attention can be attracted, and emphasis added, that may not cause slowing of the acquisition time requirement. These techniques have been utilized in a number of studies (Cohn and Nguyen, 2002a, b) and they spring from applying contemporary knowledge of the workings of the visual nervous system. Dudek et al. (2000) and Dudek and Ullman (2002) have observed that flashing a line of text during presentation requires longer for the message to be read. Beyond these matters a more general inquiry is presently underway. Westat, Inc. is conducting a study to find if color and/or animation have been used successfully with CMS technology. Bushman and Taylor (2000) have argued that static signage is less effective than signage with active elements.

The Blank CMS: The FHWA white paper defines the blank sign problem as one of deciding whether to impart information when critical messaging is not needed. Apparently some authorities like to do this and others think that it detracts from the importance of a later message. The phenomenon of ‘change blindness’ (CB) has been asserted to be a key reason to not use the CMS for low priority messaging. This assertion is overly conservative in at least one important way. The CB phenomenon refers to an inability to perceive that some elements in a scene have changed and would have less application in a scenario in which the entire message is replaced.

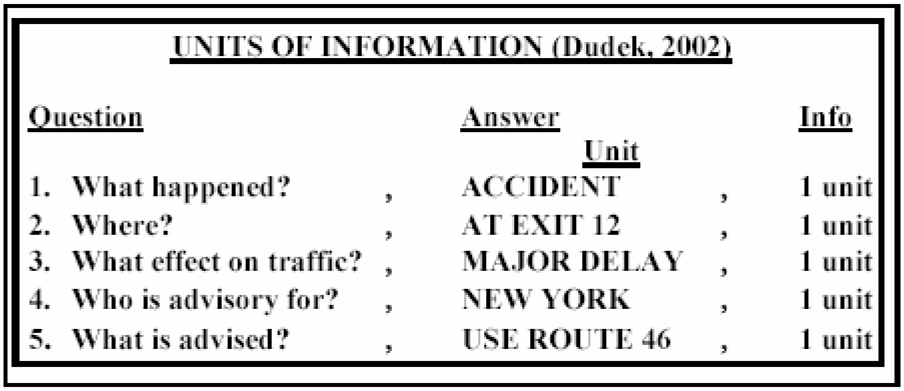
Geometrical Configuration of CMS: Currently, the Caltrans standard for CMS message is center-justified (CJ); however, our research shows that different geometric configurations improve the ability of the driver to understand the sign.

Amber Alert: The use of CMS for the Amber Alert system poses a new type of problem, that of information overload. This is highlighted in the white paper, which describes the state of research in this area as “very weak” as follows:

> “The amount of information that is generally requested to display during and (sic) AMBER alert generally exceeds the maximum allowable units of information. Research should be conducted to determine the role of CMSs in AMBER alerts and the most effective messages. It is speculated that a license plate number and a telephone number usually displayed in an AMBER alert message is equivalent to two units of information each or more. There is a need to

conduct studies to determine the unit of information equivalencies of license plate numbers and telephone numbers.”

The Concept of the Informational Unit: It is thought that the design of the CMS message should be guided by the number of informational units that it needs to contain. An informational unit has been defined as the answer to a single question such as shown in the table below taken from Dudek (2002). Too many such informational units, it is argued, prevent the motorist from grasping the entire message, and may cause the motorist to slow in order to study the CMS. It is clear that answers to some questions are more informative than to others. Since prevailing practice is to avoid the use of the CMS for anything but traffic alerts, the answer to question #3 in the table below can hardly be anything but a “delay”. However, many different things can be the answer to #1. The number of recommended units varies with the speed of traffic flow past the CMS. For example, a CMS containing three units of information for 75 mph passing traffic could be increased to four units for 55 mph traffic.



UNITS OF INFORMATION (Dudek, 2002)

Question		Answer		Info Unit
1. What happened?	,	ACCIDENT	,	1 unit
2. Where?	,	AT EXIT 12	,	1 unit
3. What effect on traffic?	,	MAJOR DELAY	,	1 unit
4. Who is advisory for?	,	NEW YORK	,	1 unit
5. What is advised?	,	USE ROUTE 46	,	1 unit

Table 1 – Definition of Informational Unit

This assumption, that the answer to a question constitutes a single informational unit, has never been tested and reliance upon it may be problematic. Information processing by human observers obeys information theoretic rules. For example, the number of bits of information in a message is the number of answers to yes-no questions in which yes and no are *a priori* equally likely. The computation and processing necessary to deal with a one-bit message is much less than that required for a three-bit message. So for example, if there are eight possible and equally likely reasons for which a lane closure is necessary, and if it is essential to alert motorists to the precise nature of the closure, then three bits must be conveyed. It is possible that it will take the human observer longer to assimilate the three bits than a simpler, say, one bit message.

Methodology of Experiments

The laboratory study tested the ability of human observers to discern CMS-like messages while conducting a separate task akin to lane keeping. Observers were selected from two populations: young adults whose vision places them in the licensable segment of the population (and who are likely to have driver’s licenses) and older adults (age > 60 who are licensed drivers).

The tests studied the accuracy with which CMS-like messages (restricted to at most three lines of 16 characters in one frame for a simulated permanent CMS, or at most three lines of 8 characters on at most two separate frames for a simulated portable CMS) can be acquired as a function of various manipulations of the structure and presentation qualities of the CMS messages. Accuracy is measured in a memory recall task where text of a CMS message is briefly displayed and the observer must write down what was presented. Accuracy is estimated by percent correct identification (percentage of words correct (WC) and the percentage of numerals correct (NC)). Word percentage correct measures the ratio of the total number of words that the observer got correct in *all* trials in a given test sequence to the total number of words contained within those trials. Certain words such as articles (e.g. "a", "the") that did not contribute to message understanding were excluded from the scoring. Numeral percentage correct measures the ratio of the total number of digits that the observer correctly recalled over all the trials to the total number of digits within all the CMS message frames for those trials.

A lane-keeping-like task was presented and the observer had to keep his 'vehicle' within lane boundaries while at the same time looking for messages on signs at the 'side' of the road. The sign increased in size over time as if the observer were passing by it at the design speed of 65mph. The task recorded the ability of the driver to stay within the simulated lane and quantified the data with a total percentage of time spent within the lane across each trial.

In order to assure that the observers did both tasks well, a cost benefit was attached to the tests conducted. The test proposed that the observers would get paid one penny for every word element (WC) and number element (NC) they got correct, provided that they keep within the simulated lane keeping task more than 95% of the time for each trial. If they failed in the lane-keeping task, they would not get any extra pay. Observers were paid a basement level of ten dollars an hour for their time in addition to the incentive pay.

Two sets of experiments were conducted for the sign configuration tests. The first set of experiments, considered preliminary, presented the distracter task on one computer monitor while the animation was displayed on another monitor. No distracter task data was recorded for these experiments because it was a quick test to see what the most significant data is. Only young observers were tested for these experiments. The distracter task akin to lane keeping was originally programmed to move randomly (e.g. white noise with filter). This was deemed too haphazard and not a good simulation of driving for the final set of experiments. The preliminary experiments only tested the following manipulations of the CMS message:

- timing of successive frames (on times and off times) – portable CMS only
- spatial structure of the message display – permanent CMS only
- movement vs. steady presentation of the message display – permanent and portable displays.

The main set of sign configuration experiments moved the distracter task and sign animation onto a single computer, which was deemed preferable in order that the simulated signs not appear at too eccentric a visual angle relative to the distracter task. The ‘white noise with filter’ for the lane movement within the distracter task was changed to a sum of sinusoids. The distracter task then also recorded the ability of the observers to stay within the lane. The final sign configuration experiments tested all the following manipulations of the CMS messages:

- timing of successive frames (on times and off times) – portable CMS only
- spatial structure of the message display – permanent CMS only
- movement vs. steady presentation of the message display – permanent and portable displays.
- changing one element within message display – permanent CMS only
- compacting the CMS information – permanent and portable CMS

After the sign configuration tests had been completed a separate set of tests was done to better delineate the concept of “informational unit” as it applies to certain kinds of CMS messages. In particular, the ability to correctly recall a string of alphanumeric characters (a randomly generated truck license plate) that has no intrinsic meaning (i.e. it doesn’t consist of words; the string forms no obvious pattern) was tested against the ability to recall a single character denoting a direction (‘E’ for East or ‘W’ for West).

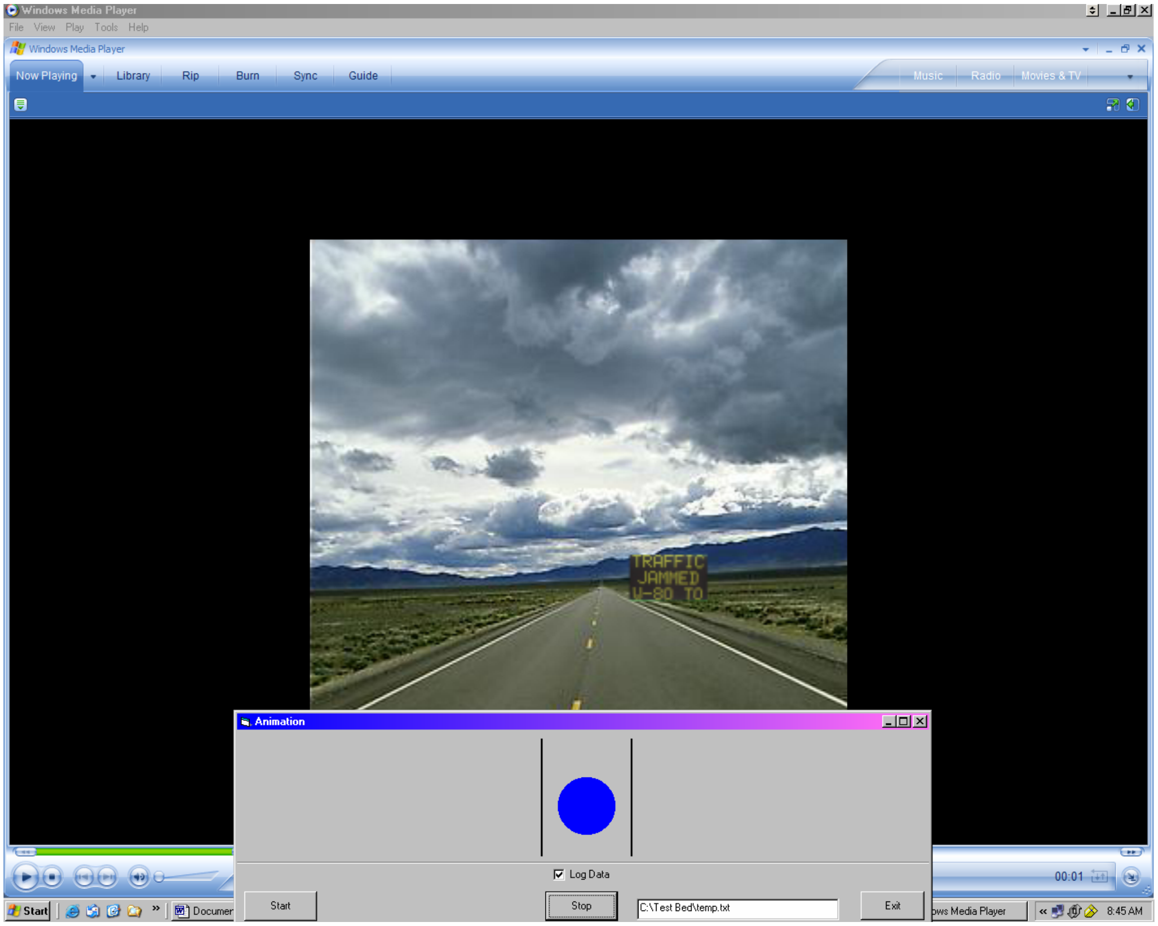
Ten observers were used for this final series of tests. There was some overlap with the prior set of observers but this new set was not broken down by age. Each observer performed forty trials. The observers had to record “E” or “W” if it was an E/W trial and they had to record the license plate characters if it was a license plate trial. There were twenty E/W trials and twenty license plate trials with the two types of trials being interleaved at random. In addition the choice of either “E” or “W” being presented in an E/W trial was randomized. The video presentation method was identical to that for the permanent CMS message simulations. The scoring is discussed in the final results section. The distracter task was identical to that for the final sign configuration tests.

Tests were conducted at the Visual Detection Laboratory on the University of California, Berkeley campus. This laboratory has extensive light measuring instrumentation plus proven pre-simulator apparatus (high-speed monitor, optical equipment) for developing testing modalities. Light-baffled doors, black-out shades, and controllable (dimmable) lighting are available. Observers were positioned a fixed distance from the CMS-like display. It was structured to subtend the same angular subtense as it would at the specification distance for 65 mph travel (Dudek, 2002). The display color and font simulated contemporary CMS design.

Distracter Task

In order to prevent the observer from devoting all of his/her attention to reading the message signs, a distracter task was implemented simultaneously with the moving message display. The task was an analog of steering.

It was displayed at the bottom of the same computer screen that showed the message signs (see Figure 1 below).



☑ Log Data

C:\Test Bed\temp.txt

Figure 1: Screen capture shot of test. The distracter task is the uppermost or foreground (active) window. The CMS message is shown on the lowermost or background window (*Media Player*^{TM}).

The task consisted of the following procedure. Upon clicking on the “Start” button on the distracter task window, the video of the relevant message sign would begin playing on *Windows Media Player*™ on the background window and, at the same time, the vertical black bars on the distracter window would start moving horizontally. They would shift left and right (see Figure 2) together. That is, the two bars would shift in lockstep, both either moving right or both moving left, always maintaining a fixed horizontal separation. The mouse cursor (not shown in any of the figures here) would “lock” onto the disk shown in the distracter window and be able to move the disk by virtue of moving the mouse (no mouse button depressions were necessary for this). The observer was told to use the mouse to keep the disk between the vertical bars. If the observer was successful at keeping the disk between the vertical bars (“in the lane”), then the disk would be blue. This is shown in the larger image of the distracter window alone in the figure below.

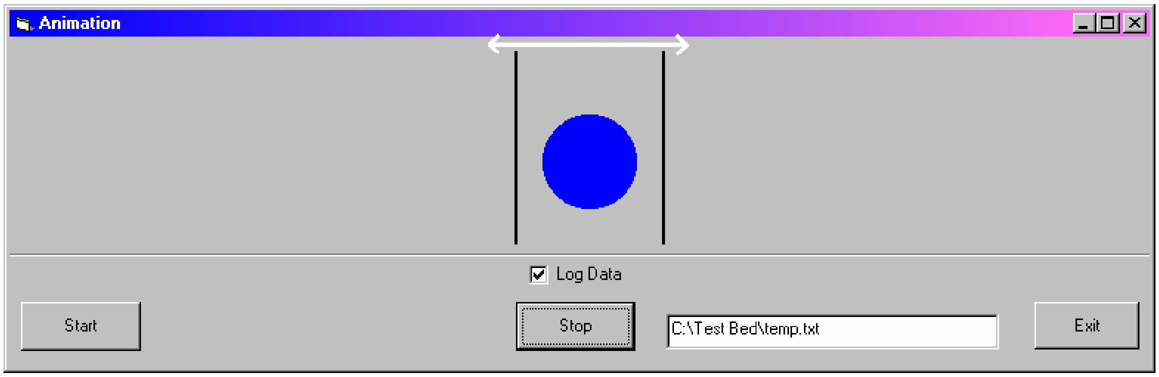


Figure 2: Disk is blue when kept "in the lane". The white arrows indicate the directions the two bars can move. The white arrows are not present in the actual distracter task.

When the observer fails to keep the disk “in the lane”, the disk becomes red (see Figure 3). The disk turns from blue to red when one quarter (or more) of the disk’s width passes beyond either bar.

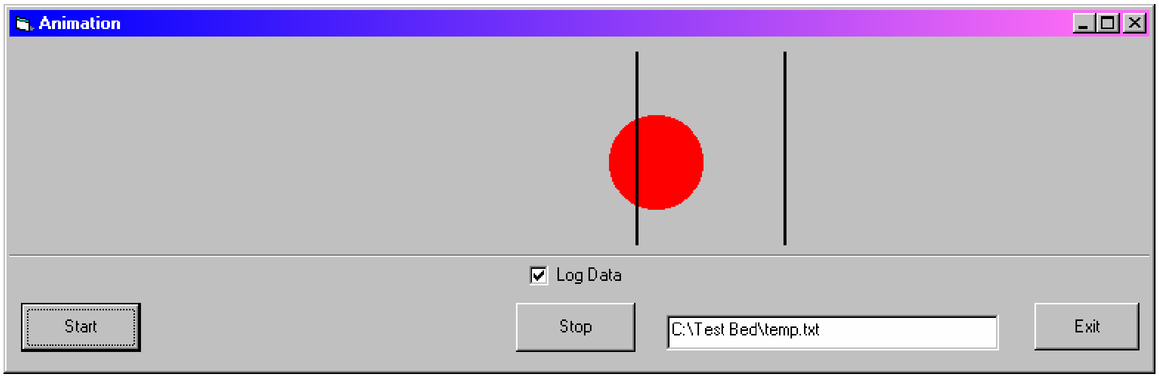


Figure 3: If the observer doesn't keep the disk between the bars then it turns red. Here the disk has just turned red as ¼ of its diameter has passed outside the left bar.

The computer program running the distraction task keeps a running tabulation of the amount of time the disk is “in the lane” or “out of the lane” while the message sign video is playing. The tabulation begins 600 ms after the “Start” button is pressed in order to allow the mouse cursor to be moved from that button to between the bars without incurring a penalty. The tabulation ends when the message sign video ends. The “Stop” button ends the movement of the bars before the observer begins the next trial. The next trial is begun when the observer again hits the “Start” button launching the next video in the queue and starting the bar movement again. Clicking on the “Exit” button ends the programs and the series of trials.

The series of trials (i.e. message sign videos) were in an external playlist that was read by the program. This allowed the experimenter to change the order of presentation of the messages or break the experiment up into a series of sets of trials if the observer was getting fatigued from a very long set of messages. The recorded data for “time in lane” for a given trial was associated to the corresponding video by its order in the playlist.

The back-and-forth movement of the bars along the horizontal was intended to simulate a curving road since the road in the video was straight and making that latter road curve would have been a formidable programming task. The bars were originally programmed to move randomly. This was deemed too haphazard and not a good simulation of driving.

The problem then arose as to how to avoid observer anticipation of the bar movement. Since the observer had to undergo many trials, a repeated pattern (especially a very simple one) would allow the observer to "learn the road" and thereby lessen the effectiveness of the distraction in later trials. There was therefore a tension between needing a smooth, deterministic pattern and keeping the observer from anticipating the pattern.

The following solution was used. A sum of three sinusoids of differing frequency and amplitude was used. The value of this, as a function of time, gave the position of the center between the bars (which move in synchronicity) in "grid units"; these are distance units along the screen that are proportional to pixels. The actual length can vary with screen size or resolution. By way of scale, the horizontal distance between the two vertical bars is 100 of these units. The function is shown in the next figure.

$$f(t) = 20\left\{\sin(2\pi \cdot 0.10t) + 2\sin(2\pi \cdot 0.25t) + 3\sin(2\pi \cdot 0.14t)\right\}$$

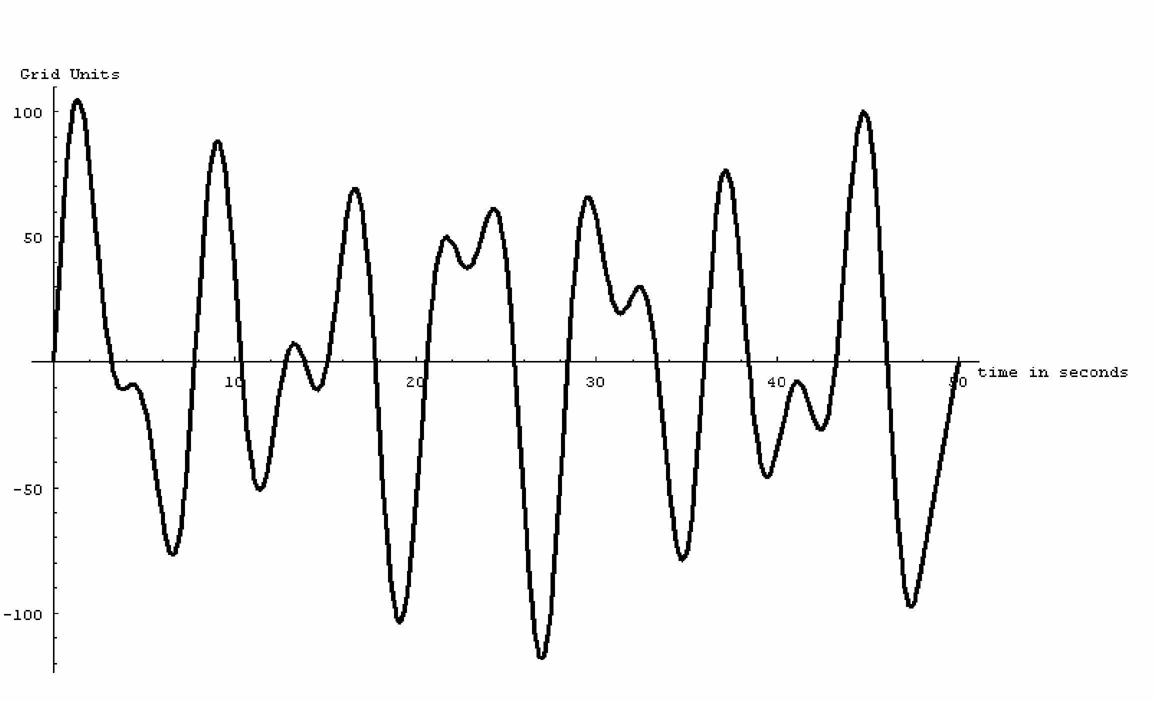


Figure 4: Horizontal movement as a function of time. The equation describing the graph is above it. One hundred "grid units" represents the distance between the vertical black bars in the distracter window.

An additional precaution was taken to prevent learning from one trial to the next. Each ten-second trial had the motion in the distracter window begin ten more seconds into the above plot than the last trial. Thus the first trial ran had the bar motion follow the above plot from 0 to 10 seconds; the second trial had the bar motion follow the above plot from 10 to 20 seconds, etc. This scheme provided smooth, deterministic yet challenging bar (“lane”) movement without the observer being able to memorize or anticipate that movement.

Sign Animation

The tested CMS messages were selected from a CMS archive given from Caltrans district IV for the second and third quarters of the 2002-2003 fiscal year. The archive was broken down into 17 categories and messages were selected from the categories based on their frequency within the archive. The digital images of the signs were made from *SIGNView*^{TM}. The research was restricted to simulated signs displaying 16 characters in each of three lines, representing permanent CMS displays, or signs containing only 8 characters in each of three lines, as is the case in portable CMS displays. The animation was matted on top of a background with a road extending towards the horizon. Figures 5, 6A and 6B show a typical permanent CMS and portable CMS respectively.

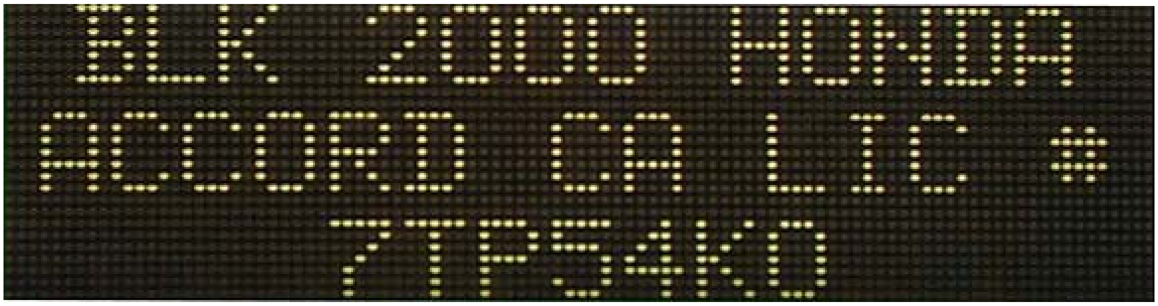


Figure 5 – An AMBER Alert message typically seen on a permanent CMS.

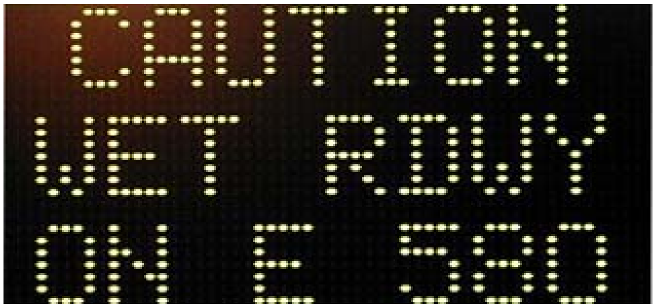


Figure 6A – The first frame in a two-frame CMS message.

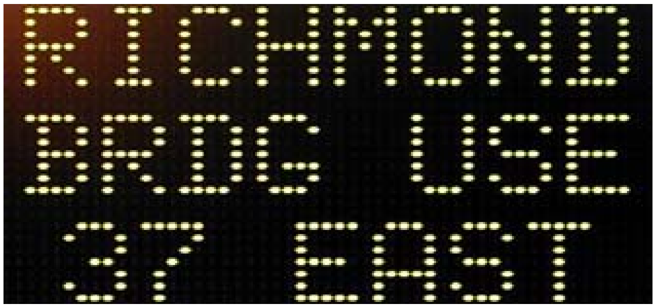


Figure 6B - The second frame in a two-frame CMS message.

All CMS messages were structured to subtend the same angle as would be the case for an actual sign at the specification distance for 65 mph travel (Dudek, 2002). The calculations for computing the approach distance and size of the sign can be seen in Appendix A.

The images were finally compiled to run in succession on an Adobe software platform to create the animations visible on the lowermost window in Figure 1.

Experiments

The computer controlled CMS display program quantified a observer's ability to stay within a simulated lane while measuring the accuracy of his/her response (comprehension of message content) as a function of various manipulations of the CMS message display including:

- timing of successive frames (on times and off times) – portable CMS only
- spatial structure of the message display – permanent CMS only
- movement vs. steady presentation of the message display – permanent and portable displays.
- changing one element within message display – permanent CMS only
- compacting the CMS information – permanent and portable CMS

The first set of experiments examined the effect of inserting a blank "off screen" between successive frames in a portable CMS. The duration of the off period between frames ranged from 0 ms (no blank screen, the current standard advised by the California DOT) to 500 ms. The following figures 7A, 7B, and 7C indicate the typical time flow of the two-frame portable CMS message on a trial run. 18 trials were run for each iteration of off time (0 ms – 500 ms). The data was compared against the Caltrans standard of 0 ms off-time.

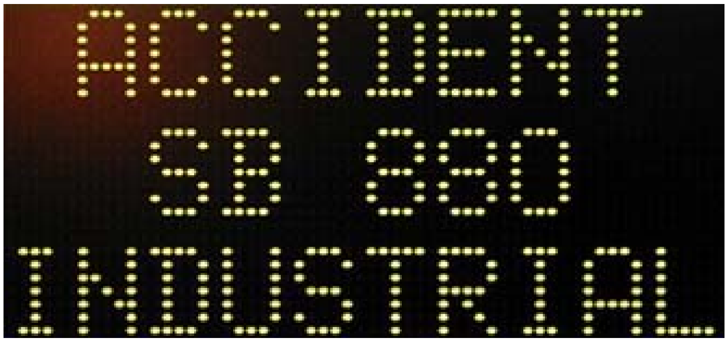


Figure 7A: The first frame in a two frame portable CMS sign left on for a duration of 1 second.

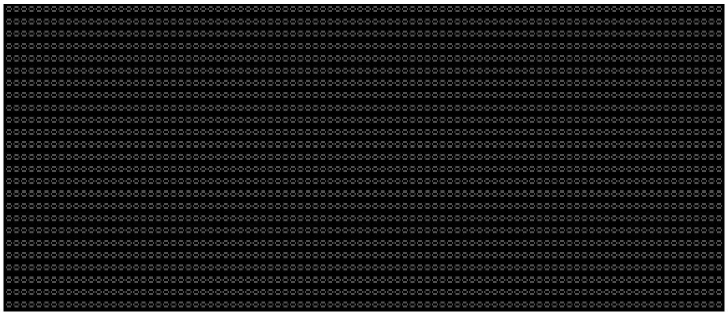


Figure 7B: The 'off screen' between the 1st and 2nd frame of a two-frame CMS message. The exposure duration is varied from 0 ms to 500 ms.

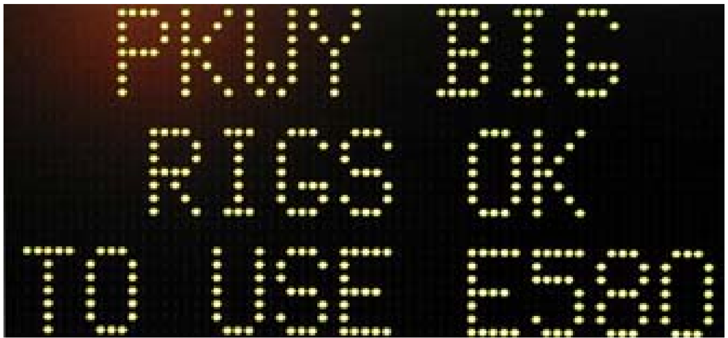


Figure 7C: The 2nd frame of a two-frame CMS message. The exposure duration is 1 second.

The next set of experiments examined the effect of varying the spatial structure of a permanent CMS. The portable CMS was not explored for this exercise because it does not have the room available on its small 8 characters by 3 rows for geometrical manipulation of the text. The message content is presented left-justified (LJ), center-justified (CJ), right-justified (RJ), or what we term staircase-justified (SJ). In the SJ case, the top row is LJ, the middle row is CJ and bottom row is RJ. 35 trials were run for each particular case. The results from LJ, RJ and SJ were compared against the Caltrans standard of CJ. The following figures 8A, 8B, 8C, and 8D represent LJ, CJ, RJ and SJ, respectively.

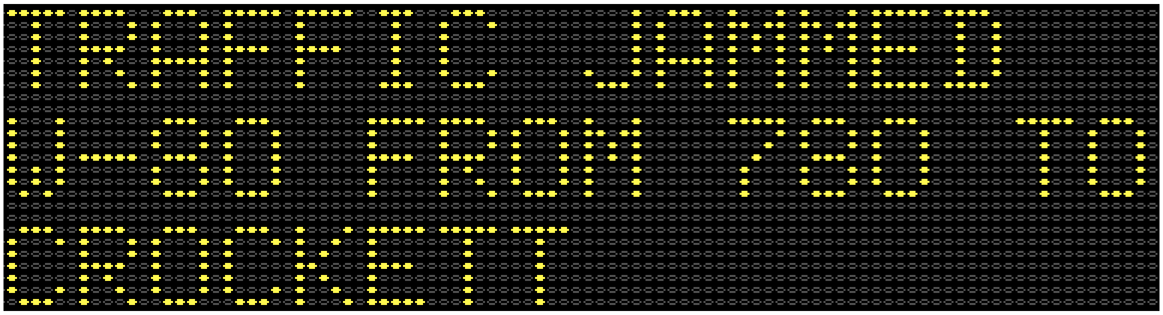


Figure 8A: The permanent CMS sign for traffic information is left-justified (LJ).

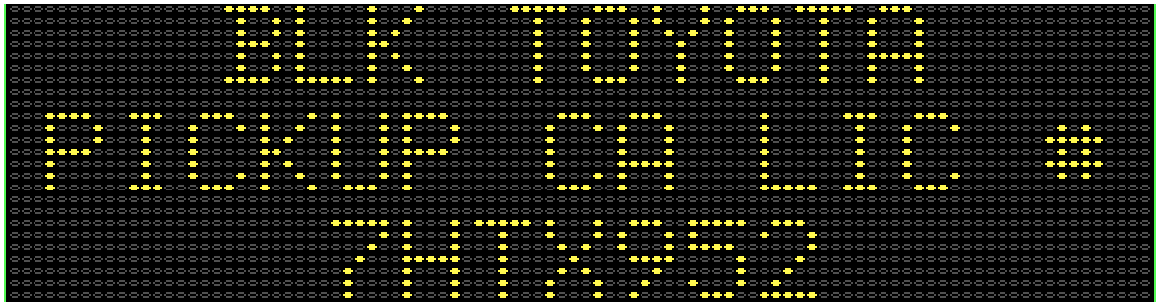


Figure 8B: The AMBER Alert CMS has the text center-justified (CJ).

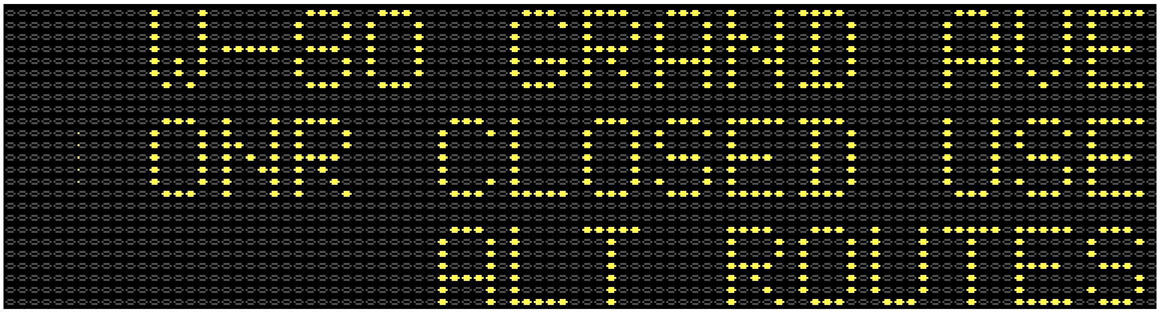


Figure 8C: A typical on-ramp closure CMS has the text right-justified (RJ).

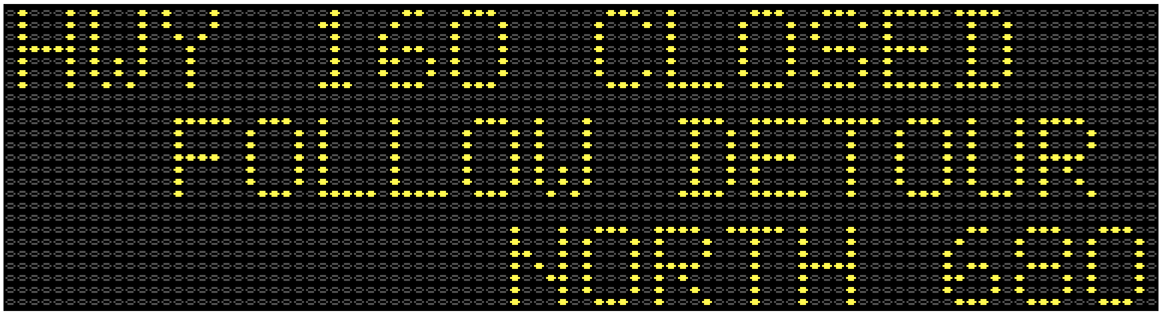


Figure 8D: A highway closure CMS has the text staircase-justified (SJ).

The third set of experiments tested whether artificially expanding the text contained within a CMS message has an effect on intelligibility. CMS displays with looming text were tested against displays in which the size of the text is held constant (relative to the overall dimensions of the display). Looming refers to the text within the CMS expanding from about 16% to 100% in size. Thirty-five trials were run with the permanent CMS and 18 trials were run with the portable CMS. For each trial, the message loomed twice to give a driver time to read and comprehend the message. The looming CMS displays were compared against the Caltrans standards for permanent CMS and portable CMS. Figures 9A-9D show a time sequence run of one looming sequence for a permanent CMS. Keep in mind that two such sequences were presented in each trial. The portable CMS is not shown because it will be the same process for portable CMS but just smaller dimensions.

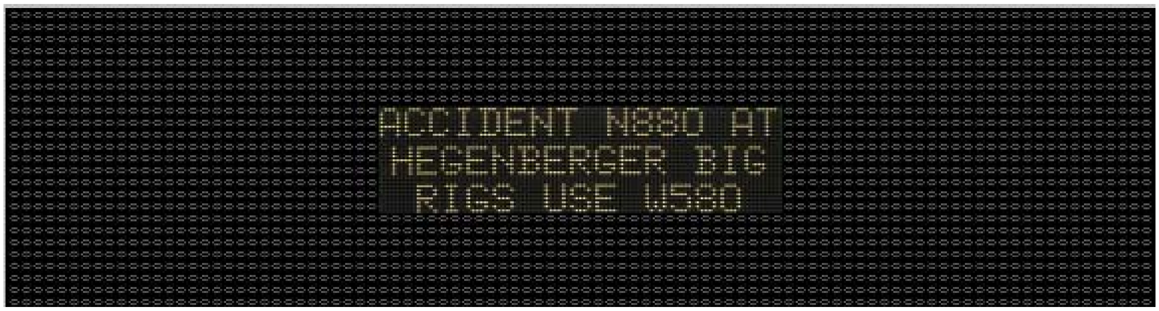


Figure 9A: A looming CMS message shown at 35% of its original dimension.

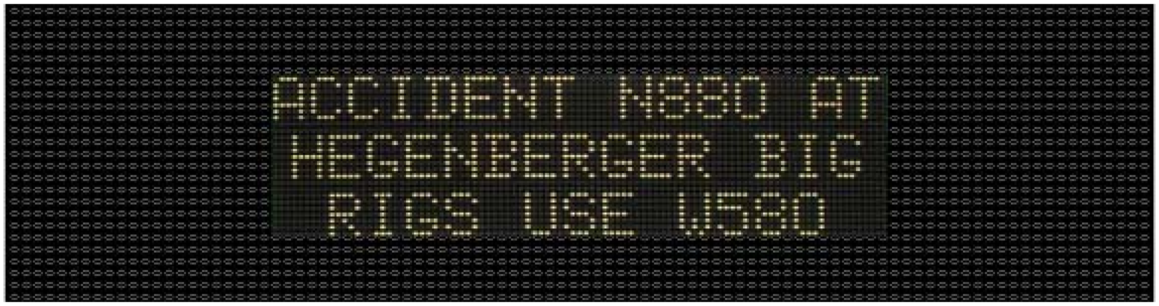


Figure 9B: The same looming CMS message shown at 53.6% of its original dimensions.

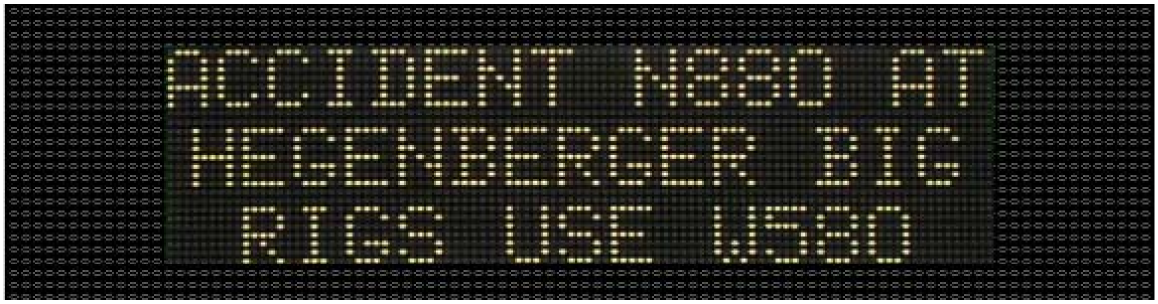


Figure 9C: The same looming CMS message shown at 72.2% of its original dimensions.

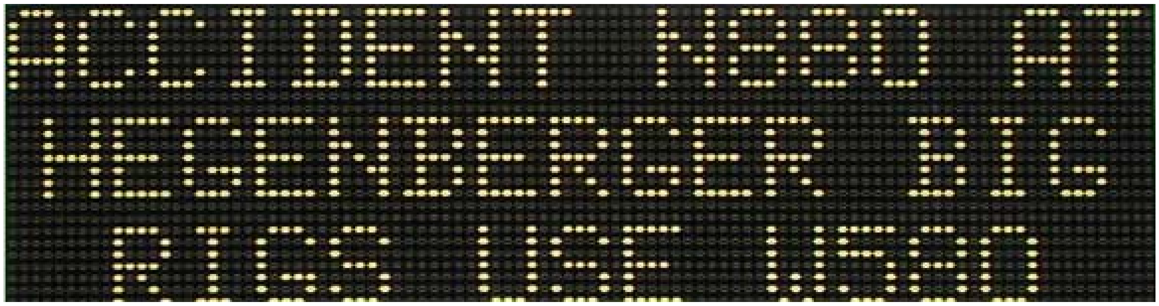


Figure 9D: The same looming CMS message shown at 100% of its original dimensions

The fourth set of experiments tested the effect of varying one specific element like a number or letter on a permanent CMS. These tests are done to see if the change blindness phenomenon occurs on CMS signs when a blank screen is used in between a two-frame message. The CB phenomenon refers to an inability to perceive that some elements in a scene have changed. Thirty-six trials of the permanent CMS were tested to see if this phenomenon occurs. Figures 10A, 10B, and 10C represent one cycle of the CMS message as it cycles through a two-frame message. The permanent CMS message was tested because the sign contains more information and it makes it harder for the observer to grasp the change in the entire image.

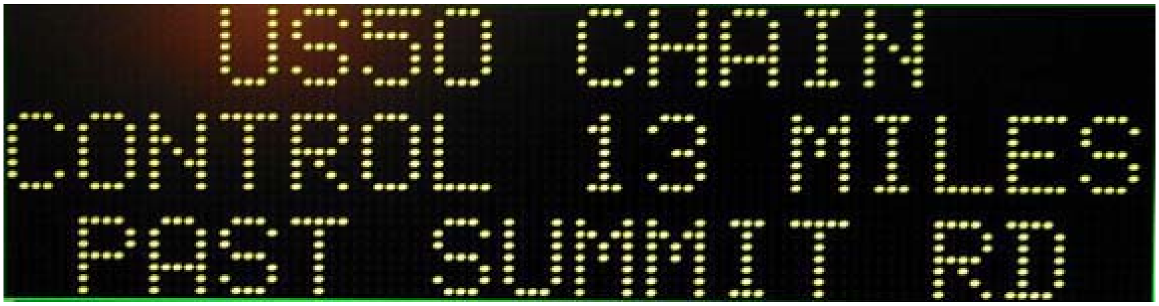


Figure 10A: The first frame of a two-frame CMS message. The frame is left on for 1 second.

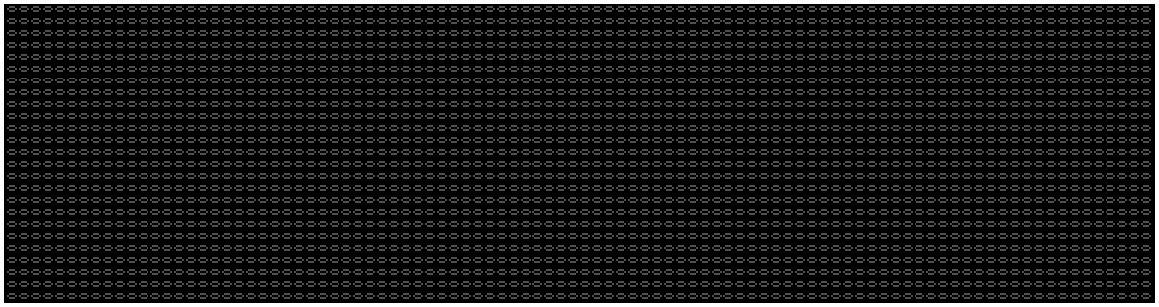


Figure 10B: The blank frame in between the first and second frame of a two frame CMS. The on time for this blank screen is 300 ms.

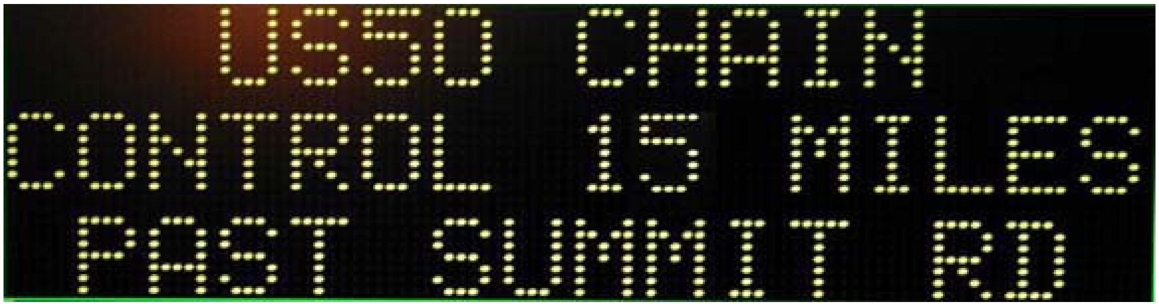


Figure 10C: The second frame of a two-frame CMS message with the '3' changing to a '5'. The frame is left on for 1 second.

The fifth set of experiments tested whether abbreviating words on a CMS in order to fit the necessary information into the limited physical space affected intelligibility. CMS displays with compacted words were tested against CMS displays that had no abbreviation of the words. Thirty-five trials of the permanent CMS and 18 trials of the standard portable CMS were tested to see if the compacted CMS message could help improve intelligibility. The reason behind shortening the CMS display is to provide more information on the limited space on a typical CMS display. The following is a list of all the abbreviations and their long versions that were used:

SR – State Route
E – East
W – West

N/ NB – North/ North bound
S/ SB – South/ South bound
CA - California
NV – Nevada
TX – Texas
FL – Florida
BL – Blue
BLK – Black
RD – Road
MPH – Miles per hour
AM – Amplitude Modulation (Radio)
BRDG/ BRG – Bridge
HWY – Highway
BETW – Between
CLSD – Closed
LN – Lane
RDWY – Roadway
RDWK - Roadwork
BLKD – Blocked
ALT – Alternate
RTE/ RTES – Route/ Routes
AVE – Avenue
BLVD – Boulevard
ST - Street
ONR – on ramp
SFO – San Francisco Airport
SAT. - Saturday
LT – Left
RT - Right
MIN/ MINS – Minute/ Minutes
SF – San Francisco

Figures 11A, 11B, and 11C show an example of a permanent CMS sign along with a portable two-frame CMS display where the CMS message has been compacted.

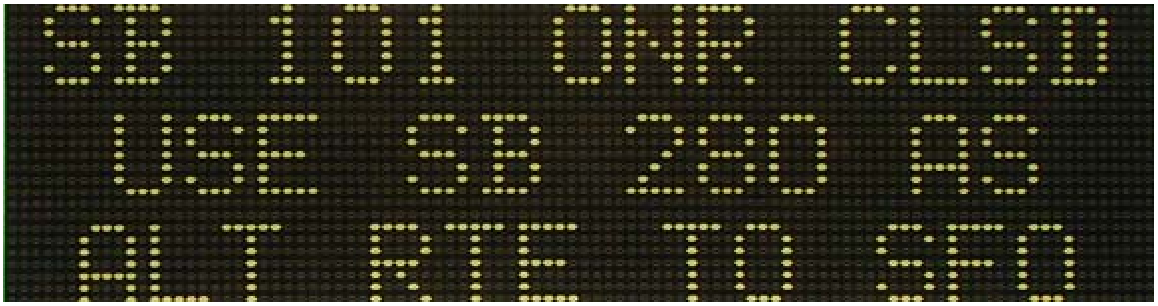


Figure 11A: A compacted permanent CMS sign.

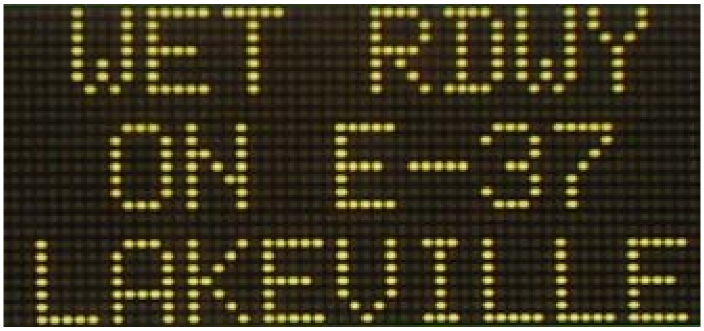


Figure 11B: The first frame of a two-frame CMS that is compacted. The frame is left on for 1 second.

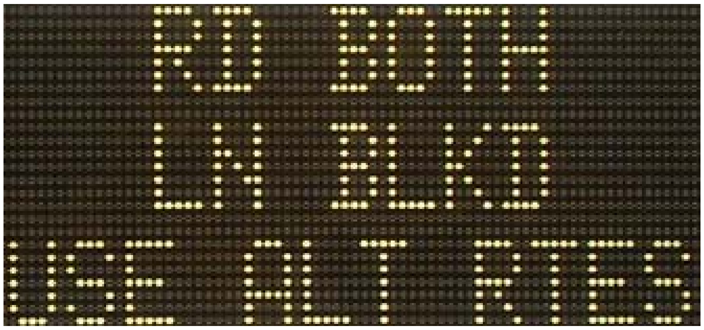


Figure 11C: The second frame of a two-frame CMS that is compacted. The frame is left on for 1 second.

Results

Data Reduction and Analysis

Given the large number of tests undertaken with a variety of conditions and with many observers, it is unwieldy to reproduce all the raw data in this report. Nor is it particularly enlightening to see the details of every calculation based on this mass of data.

Nonetheless, in order that the reader can have confidence in the reduction and analysis procedures used, an illustrative example excerpted from the raw data is shown below.

In Table 2 the results for observer #1 (a young observer) during the 300 ms blank interframe test are shown. During this particular sequence of trials, 18 portable CMS messages were shown. Each message was divided into two frames (e.g. trial number 1.1 and trial number 1.2). Between the frames a completely blank CMS screen was shown for, in this case, 300 ms. Each frame had a certain number of words and numerals that needed to be recalled by the observer. Certain words such as “the”, deemed unnecessary for comprehension, were excluded from the scoring. Some frames had no numerals and thus a “N/A” (not applicable) is shown to denote this. The percentage score for each frame—the number of words [numerals] recalled for that frame divided by the total number possible to be recalled for that frame multiplied by a hundred—is also shown. For example, trial 1.1 had 3 words that needed to be recalled correctly and only two of them were, leading to a word percentage of 66.67. The total possible number of words that could be recalled from all frames is 129. The word percentage in the last row represents the total number correct divided by the total number available, that is 89/129 (times 100) or 68.99%. It does **not** represent the average of the word percentages from each trial, which would be 72.52%.

Trial Number	Word Comprehension	Word Percentage	Number Comprehension	Number Percentage
1.1	2	66.67	0	0
1.2	4	100	2	100
2.1	3	100	4	100
2.2	2	100	4	57.14
3.1	3	100	0	N/A
3.2	3	75	1	100
4.1	3	100	2	100
4.2	4	100	2	100
5.1	3	100	0	N/A
5.2	1	33.33	0	0
6.1	2	100	3	100
6.2	1	25	0	0
7.1	3	100	0	N/A
7.2	3	100	7	100
8.1	2	40	0	0
8.2	0	0	0	0
9.1	3	100	3	100
9.2	4	100	0	N/A
10.1	4	100	0	N/A
10.2	3	100	0	N/A
11.1	2	66.67	3	100
11.2	2	66.67	3	100
12.1	3	75	0	0
12.2	4	100	0	0
13.1	1	20	0	0
13.2	3	100	0	N/A
14.1	5	100	0	N/A
14.2	0	N/A	0	0
15.1	4	80	4	100
15.2	0	0	0	0
16.1	4	100	3	100
16.2	2	40	2	100
17.1	4	100	4	100
17.2	2	50	0	0
18.1	0	0	0	0
18.2	0	0	0	0
Total Correct	89	68.99	47	47

Table 2: Word comprehension (WC) and number comprehension (NC) scores for observer #1 [a young observer] during portable CMS 300 ms blank interframe trials.

The total number of numerals (digits) that had to be correctly recalled in all frames was 100. Thus both the total number recalled and the total percentage are both 47. The average of the number percentages would be 52.04%, the “N/A” values obviously being excluded.

This brings up the question of WC (or NC) scoring methods for a given set of trials (such as the table above). Two possible methods (at least) come to mind: a) the total words recalled divided by the total that could potentially be recalled, or b) the average of the individual percentages. These do **not** give the same numerical result, although both methods seem to yield close scores in practice.

For instance, suppose that only two trials were run and that in the first case three words needed to be recalled and that only two actually were while in the second trial four words were remembered correctly out of a possible five. The percentage score using method ‘a’ is

$$\mathit{Score\ a} = \frac{2+4}{3+5} \times 100 = 75$$

while that using method ‘b’ is

$$\mathit{Score\ b} = \frac{\frac{2}{3}+\frac{4}{5}}{2} \times 100 = 73.3.$$

Also, as mentioned above, the data in the sample table gives a WC score of 68.99% using method ‘a’ and 72.52% using method ‘b’. Arguments can be made for both scoring methods and some further comments are in the second appendix.

Since the methods yielded similar (but not identical) numbers the choice of which to use came down to simplicity. Method ‘a’ requires fewer calculations and thus computational errors were less likely to arise. Thus word and number comprehension scores for a set of trials were based on the total correctly recalled over all trials in the set.

Using the totals for a WC score, a table can be constructed for young observers comparing the 300 ms interframe blanking time WC scores to the 0 ms interframe blanking time (the standard) WC scores. This is shown below.

Observer number	WC total for the 300 ms off-time test (out of a possible 129)	WC total for the 0 ms off-time test (out of a possible 129)
# 1	89	72
# 2	106	90
# 3	127	112
# 4	108	83
# 5	80	61
# 6	128	121
Perc. Avg.	82.43%	69.64%

Table 3: WC scores for young observers in 300 ms and 0 ms tests. These are used in the t-test.

One can see that observer #1's WC score of 89 from Table 2 is used as his score for the 300 ms test. The other numbers are retrieved from the raw data in a similar manner. The last row denotes the average percentage correct (out of 129 words) among the six observers. The average percentage with a 0 ms interframe off-time was around 70%. This improved to over 82% with a 300 ms interframe off-time. The question arises whether this improvement is statistically significant or just a fluke.

The two data columns in Table 3 (without the last row showing percentages) provide the basis for a *paired t-test* that can answer this question. The results of the t-test ("paired two sample for means", 5 d.o.f.) are:

$$t = 6.90$$
$$t_{crit.\,0.99} = 3.36$$
$$p\ (one\ tailed) = 4.89\times 10^{-4}.$$

Since the t statistic of 6.90 is much larger than the (one tailed, 99% level) critical t value of 3.36, the improvement of the observers' average percentage of words correct from 69.64% to 82.43% with the larger off-time is statistically significant. The (one tailed) p-value of $4.89\times 10^{-4} \approx \frac{1}{2000}$ means that if the 0 ms time result was really the same as or bigger than the 300 ms result (null hypothesis) then there is only about a one in two thousand chance of getting the present result through a statistical fluke.

The above calculation is illustrative of the data reduction and analysis that was employed in constructing the tables in the sections that follow.

As mentioned previously, the distracter task software tabulates the time the observer is "in the lane" during a given trial. It does this by recording an "in-lane" (blue) or "out-of-lane" (red) condition at fixed intervals during a given trial (video). The recording interval is constant (typically 78 ms) *during* a run but can change between runs depending on what other software processes the computer is running. Each video lasts ten seconds and recording begins 600 ms after the start. Thus there are approximately $120\ = \frac{10,000\ ms - 600\ ms}{78\ ms}$ recorded lane conditions per video. The number of "in-lane" conditions is divided by the total and multiplied by 100 to get the time in-lane percentage.

For example, during trial 1 for observer #1 in Table 2 (consisting of the cyclic repetition of CMS message frame 1, the 300 ms blanking time and CMS message frame 2 for a total run of ten seconds) the computer recorded 120 lane conditions every 78 ms. Seven of these were "out-of-lane" conditions and 113 of these were "in-lane" conditions, yielding a time in-lane percentage of $94.167\left(=\frac{120-7}{120}\times 100\right)$. Other "time in-lane" values were computed in exactly the same manner.

Preliminary Results

As mentioned above, a preliminary set of experiments was done before the main series. The purpose of this was twofold: a) to refine the experimental procedures and b) to narrow down the "parameter space" of potential tests in order to concentrate on the most promising areas.

The refinements after the preliminary tests included putting the distracter task on the same screen as the simulated highway and making the distracter task more realistic than it had been in the preliminary tests. Narrowing the scope of the tests was necessary because the tests were quite long and tedious. The flagging of the observer's attention causing biasing of the data was a very real possibility, just as repeated message exposure artificially aiding message recall was. Thus the final tests did not cover as wide a range as the preliminary tests in an effort to reduce these risks and, of course, not waste the observer's time on unlikely prospects for improvement.

There were five observers used in the preliminary tests. Four of them were again used in the subsequent final tests. All of the observers were from the younger cohort.

The first set of experiments examined the effect of inserting a blank "off screen" between successive frames in a portable CMS (this is explained in more detail below in the Final Results section). The duration of the off period between frames ranged from 0 ms (no blank screen, the current standard advised by the California DOT) to 500 ms.

As can be seen below, the word comprehension scores (Table 4) increased for the 100 ms and 300 ms blanking times and the number comprehension scores (Table 5) increased for the 100 ms, 200 ms and 300 ms blanking times. The 400 ms and 500 ms off-times yielded lower scores than the standard and were therefore not included in the final tests.

	Standard (0 ms)	**100 ms**	**200 ms**	**300 ms**	**400 ms**	**500 ms**
One-tailed p value from paired t-test against the standard	**NA**	0.250	0.220	0.284	0.226	0.131
Two-tailed p value from paired t-test against the standard	**NA**	0.500	0.440	0.567	0.451	0.262
Average of percent correct among all observers	74.21	75.70	71.96	77.94	68.97	62.99

Table 4: Word comprehension scores (average of percent correct) and the associated p values from comparing against the standard with a paired t-test [preliminary trials for off-time—portable CMS].

	Standard (0 ms)	**100 ms**	**200 ms**	**300 ms**	**400 ms**	**500 ms**
One-tailed p value from paired t-test against the standard	**NA**	0.082	0.137	0.304	0.217	0.186
Two-tailed p value from paired t-test against the standard	**NA**	0.163	0.274	0.608	0.434	0.373
Average of percent correct among all observers	68.39	74.19	73.23	73.23	58.39	54.84

Table 5: Number comprehension scores (average of percent correct) and the associated p values from comparing against the standard with a paired t-test [preliminary trials for off-time—portable CMS].

The p-values shown in these tables would seem to suggest that the greater scores might not be significant, or, more properly, that the null hypothesis (i.e. no difference between the given off-time testing and the standard) cannot be rejected. Surprisingly, the final results for these types of tests did show significance ($p < 0.01$) in some cases—see the *"Final Results"* section below.

In the case of the "geometry" tests (Tables 6 and 7) left, right and staircase aligned all showed higher scores with those scores showing possible significance in the number comprehension case. Staircase alignment showed likely significance in the word comprehension case as well and if a $p < 0.05$ criteria were used instead of a $p < 0.01$ criteria, then all three would have qualified as worthy of examination in word comprehension testing as well as number comprehension testing.

	Standard (Center Justified)	**Loom (Perm. CMS)**	**Left aligned**	**Staircase aligned**	**Right aligned**
One-tailed p value from paired t-test against the standard	**NA**	0.254	0.028	0.006	0.011
Two-tailed p value from paired t-test against the standard	**NA**	0.508	0.057	0.013	0.022
Average of percent correct among all observers	72.67	75.43	83.45	87.33	86.81

Table 6: Word comprehension scores and the associated p values from comparing against the standard with a paired t-test [preliminary trials for geometry and permanent CMS loom].

	Standard (Center Justified)	Loom (Perm. CMS)	Left aligned	Staircase aligned	Right aligned
One-tailed p value from paired t-test against the standard	**NA**	0.414	0.010	3.31×10^{-4}	0.010
Two-tailed p value from paired t-test against the standard	**NA**	0.828	0.020	6.62×10^{-4}	0.020
Average of percent correct among all observers	68.36	69.27	82.73	89.64	86.91

Table 7: Number comprehension scores and the associated p values from comparing against the standard with a paired t-test [preliminary trials for geometry and permanent CMS loom].

In addition to the alignment changes, looming (discussed next for the portable CMS sign) was examined for the permanent CMS sign. It showed only a minor increase in scores and had a high p-value; thus looming for the permanent CMS sign was not included in the final results section.

Looming, artificially expanding the text contained within a CMS message, was examined in more detail for the portable CMS. Here preliminary experiments not only tested "pure" looming but also tested it in combination with an interframe blanking time (off-time). The results of these tests are shown in Tables 8 and 9. While none of the preliminary results showed signs of significance, the scores for looming combined with an off-time of 400 ms or 500 ms were lower than the scores for the standard. Thus these were dropped from final testing. Only looming alone ("pure loom") and looming combined with off-times of 100 ms, 200 ms and 300 ms were used in the final stage of testing.

	Standard No Loom 0 ms	Pure Loom (0 ms)	Loom-100 ms	Loom-200 ms	Loom-300 ms	Loom-400 ms	Loom-500 ms
One-tailed p value from paired t-test against the standard	**NA**	0.168	0.290	0.500	0.356	0.445	0.069
Two-tailed p value from paired t-test against the standard	**NA**	0.337	0.580	1.00	0.712	0.891	0.139
Average of percent correct among all observers	74.21	80.93	77.94	74.21	76.45	73.27	64.30

Table 8: Word comprehension scores and the associated p values from comparing against the standard with a paired t-test [preliminary trials for loom and loom + off-time—portable CMS].

	Standard No Loom 0 ms	Pure Loom (0 ms)	Loom-100 ms	Loom-200 ms	Loom-300 ms	Loom-400 ms	Loom-500 ms
One-tailed p value from paired t-test against the standard	**NA**	0.054	0.274	0.478	0.260	0.328	0.102
Two-tailed p value from paired t-test against the standard	**NA**	0.107	0.549	0.957	0.521	0.657	0.204
Average of percent correct among all observers	68.39	84.19	74.52	69.03	74.84	63.87	49.03

Table 9: Number comprehension scores and the associated p values from comparing against the standard with a paired t-test [preliminary trials for loom and loom + off-time—portable CMS].

The preliminary results showed that certain parts of each criterion tested had merit and warranted further studies in the final set of experiments. Only those criteria that showed higher scores than the standard in the preliminary results were examined in the final tests, (with the exception of the right-alignment testing, due to a data analysis anomaly).

Final Results—Sign Configuration Tests

The results for the main series of experiments testing different sign configurations can be grouped into five sections: (1) interframe blanking tests (or "off-time"), (2) geometry tests, (3) looming tests (including combining looming with off-time), (4) change blindness tests, and (5) compacted message tests.

- Interframe Blanking ("off-time")

The first set of experiments examined the effect of inserting a blank "off screen" between the first and second frames carrying a message in a portable CMS. In other words, as the video played, the observer saw the cycle: frame 1, blank time, frame 2, frame 1, blank time, frame 2, etc. There was no blanking when frame 1 *followed* frame 2. The duration of the off period between frames ranged from 0 ms (no blank screen, the current standard advised by the California DOT) to 300 ms. Based on the preliminary results, the only off periods of interest were the ones from 0 ms to 300 ms.

It should be noted that while the 0 ms off-time trials were used as the standard of comparison, *two* different sets of these trials were run and used. The two sets are very comparable. The only difference between them is the different messages being used. We felt that running several tests where the messages were the same in every test would lead to the observers "learning" (memorizing) the messages after a few tests, thereby skewing the results. Instead, two sets of (quite comparable) messages were created. One set was used for blank times of 0 ms, 100 ms and 300 ms. The other set was used for blank times of 0 ms and 200 ms. By interleaving the two sets in our presentation to the observers we felt that memorization would be attenuated. The only difficulty arising out of this approach is remembering to compare the 300 ms test results (say) against the results of the comparable 0 ms test. In the tables that follow a "(1)" or a "(2)" indicates which message set was used for the comparison.

The 0 ms (1) and 0 ms (2) individual observer scores were turned into percentages and then the two resulting sets were compared against each other (via a paired t-test) to validate the working assumption that the two standards did not differ in their effects upon the observers. Percentages were used because the possible word and numeral totals differed between the two sets. For example, observer #1 had a WC total of 72 out of a possible 129 under the 0 ms (1) test (see Table 3). This is 55.81%. Under the 0 ms (2) test his WC total of 64 out of 142 yields 45.07%. The values for other observers were computed in a similar vein. The average of those percentages for WC involving young observers was 69.64% for the 0 ms (1) test and 66.20% for the 0 ms (2) test. As one would expect, the difference in these values was not statistically significant, as measured by a t-test pairing off the percentage correct for each observer under both 0 ms test sets. The one-tailed p value was 0.15 and the two-tailed p value was 0.30—not even close to being significant. Thus we had no qualms in using the two different sets of 0 ms tests for comparison rather than a single set.

The greatest improvement in WC was found for off screens having a duration of 300 ms. As already shown in the *"Data reduction and Analysis"* section above, the addition of a 300 ms off screen caused the average WC score for the young observers to improve from 69.64% to 82.43%, as compared to no off screen. As can be seen in Table 10A below, the improvements for WC scores for young observers for the 100 ms and 200 ms blank times were not statistically significant, at least not at the 99% level ($p < 0.01$).

As is seen in Table 10B, word comprehension scores for the older observers also had their maximum value and largest increase (from 48.84% to 58.66%) for a blanking time of 300 ms. This result was not significant at the 99% level ($p < 0.01$) but would be (just barely) significant at the 95% level ($p < 0.05$). The large p value for the test between the 0 ms sets shows again that assuming the two standards are comparable is justified.

	0 ms (1)	**0 ms (2)**	**100 ms (1)**	**200 ms (2)**	**300 ms (1)**
p value (one-tailed) from paired t-test	0.153		2.56×10^{-2}	0.120	4.89×10^{-4}
Average of percent correct among all young observers	69.64	66.20	78.55	71.71	82.43

Table 10A: Word Comprehension (WC) scores and associated p values for portable CMS tests with an off-time between the frames (young observers only). The notation "(1)" or "(2)" denotes which message set was used for the comparison. The p value for the 0 ms tests is explained above.

	0 ms (1)	**0 ms (2)**	**100 ms (1)**	**200 ms (2)**	**300 ms (1)**
p value (one-tailed) from paired t-test	0.161		0.170	0.181	4.45×10^{-2}
Average of percent correct among all old observers	48.84	44.60	51.94	46.95	58.66

Table 10B: Word Comprehension (WC) scores and associated p values for portable CMS tests with an off-time between the frames (old observers only).

Number comprehension among the younger observers had its greatest increase of 12.5% in moving from 58.33% under the 0 ms standard to 70.83% under the 300 ms off-time test (see Table 10C). The result was not significant at the 99% level ($p < 0.01$) but was at the 95% level. Once again the large p value between the 0 ms sets justifies having two standards to help reduce the memorization problems that these kinds of experiments are prone to.

	0 ms (1)	0 ms (2)	100 ms (1)	200 ms (2)	300 ms (1)
p value (one-tailed) from paired t-test	0.160		0.108	6.66×10^{-2}	1.63×10^{-2}
Average of percent correct among all young observers	58.33	62.22	66.67	72.89	70.83

Table 10C: Number Comprehension (NC) scores and associated p values for portable CMS tests with an off-time between the frames (young observers only).

The only remarkable thing about the NC results for the older group (Table 10D) is how *little* variation there was with changing off-time. All the scores were in the low to mid thirties and the p value for the 300 ms blanking time (compared to the 0 ms standard) was exactly ½! Clearly the blanking time had almost no effect on the older observers number comprehension.

	0 ms (1)	0 ms (2)	100 ms (1)	200 ms (2)	300 ms (1)
p value (one-tailed) from paired t-test	0.456		0.335	0.443	0.500
Average of percent correct among all old observers	34.33	34.67	32.33	36.00	34.33

Table 10D: Number Comprehension (NC) scores and associated p values for portable CMS tests with an off-time between the frames (old observers only).

In addition to the data for WC and NC, the distracter task data was also recorded for the young and old observers. Young observers had an easier time meeting the 95% threshold criteria in order for them to get paid one penny for each WC element and NC element they got correct. On average, most of the young observers stayed within the lane 95% of the time. Table 10E has the results from the young observers.

YOUNG	0 ms (1)	0 ms (2)	100 ms	200 ms	300 ms
Percent of time within the lane (avg. of observers)	95.36	94.74	96.28	95.66	96.30

Table 10E: Distracter task results for portable CMS with a blank screen in between frames (young observers only)

Older observers (age > 60) had a difficult time adjusting to the sum of sinusoidal change for the lane-keeping task. Consequently, their results for the amount of WC and NC elements that they got correct were highly reduced. Table 10F contains the results for the older observers.

OLD	0 ms (1)	0 ms (2)	100 ms	200 ms	300 ms
Percent of time within the lane (avg. of observers)	56.89	59.39	60.87	59.71	60.29

Table 10F: Distracter task results for portable CMS with a blank screen in between frames (old observers only)

- Geometry

The next set of experiments examined the effect of varying the spatial structure of a permanent CMS message. The message content is presented left-justified (LJ), center-justified (CJ), or what we term staircase-justified (SJ). In the SJ case, the top row is LJ, the middle row is CJ and bottom row is RJ. Center-justified is the current CMS standard and formed the basis of our comparison.

Staircase-justified presentation provided the greatest improvement in word comprehension over the standard (CJ) presentation for both young and old observers (see Tables 11A and 11B). Young observers had their WC score increase to 85.67% from 76.41% with the standard. The increase in score for the young observers was less for the LJ presentation, which did not rise to statistical significance at the 99% level ($p < 0.01$) but was significant at the 95% level.

The increase in WC scores among older observers was greatest for the SJ geometry but it was significant at the 95% level not at the 99% level. Unfortunately, the check on the compatibility of the two standards (CJ) gave some cause for concern as the p value was 0.043 and the difference in scores was fairly large (73% vs. 59%). This is in marked contrast to the check between standards for the young observers where the scores differed by about one part in 7600 and the p value was nearly exactly ½. The reason for this difference in response to the two standards among the old observers is unknown.

	Standard (1) Center Justified	Standard (2) Center Justified	Staircase Justified (1)	Left Justified (2)
p value (one-tailed) from paired t-test	0.496		5.34×10^{-3}	3.38×10^{-2}
Average of percent correct among all young observers	76.41	76.42	85.67	81.09

Table 11A: Word Comprehension (WC) scores and associated p values for permanent CMS tests with changes in the message "geometry" (young observers only). To avoid memorization from repeated exposure two different message sets were used. The notation "(1)" and "(2)" denotes which sets were compared. Thus, for example, the p value for "left justified" comes from a paired t-test against the second standard: Standard (2). The p-value between the standards is the exception to this and comes from comparing the two with a paired t-test after first converting the individual responses to percentages.

	Standard (1) Center Justified	Standard (2) Center Justified	Staircase Justified (1)	Left Justified (2)
p value (one-tailed) from paired t-test	4.32×10^{-2}		2.18×10^{-2}	0.212
Average of percent correct among all old observers	58.85	73.12	76.13	76.68

Table 11B: Word Comprehension (WC) scores and associated p values for permanent CMS tests with changes in the message "geometry" (old observers only).

In the case of number comprehension score changes among the young observers (Table 11C), the same kind of result holds as for WC, namely SJ presentation led to the largest and most significant increase in score over the CJ standard. The LJ presentation led to a smaller increase and was significant at the 95% level but not the 99% level. There were no statistically significant changes in NC score among the older observers (Table 11D).

	Standard (1) Center Justified	Standard (2) Center Justified	Staircase Justified (1)	Left Justified (2)
p value (one-tailed) from paired t-test	0.473		2.11×10^{-3}	3.75×10^{-2}
Average of percent correct among all young observers	68.82	68.60	79.14	73.77

Table 11C: Number Comprehension (NC) scores and associated p values for permanent CMS tests with changes in the message "geometry" (young observers only).

	Standard (1) Center Justified	Standard (2) Center Justified	Staircase Justified (1)	Left Justified (2)
p value (one-tailed) from paired t-test	0.269		8.15×10^{-2}	0.445
Average of percent correct among all old observers	54.40	58.91	61.35	58.40

Table 11D: Number Comprehension (NC) scores and associated p values for permanent CMS tests with changes in the message "geometry" (old observers only).

Nonetheless, given the strong results among the young for both WC and NC score increases for SJ geometry and the moderate results for the older observers (95% level) for WC score increases, it is fair to say that staircase-justified messages definitely result in greater comprehension. It is likely that the geometric alignment of the staircase-justified text fits the natural progression of an observer's eye movements as he tries to read the content from a message.

As with the off-time tests, the distracter task data was also recorded to see if the observers were able to stay within the lane while trying to read the signs. Once again the young observers averaged around 95% of the time staying within the lane. Consequently, they were paid one penny for each WC element or NC element that they got correct. Take note that the young observers under both LJ and SJ showed a substantial increase in intelligible elements in comparison with the standards, while maintaining the 95% threshold of staying within the lane. The old older observers on the other hand had a hard time attaining the 95% criteria of staying within the lane, their observer average never reaching 65%.

YOUNG	Standard (1)	Standard (2)	Staircase Justified	Left Justified
Percent of time within the lane (avg. of observers)	95.65	95.80	95.72	96.20

Table 11E: Distracter task results for permanent CMS with a geometrical configuration change (young observers only)

OLD	Standard (1)	Standard (2)	Staircase Justified	Left Justified
Percent of time within the lane (avg. of observers)	62.06	64.48	63.59	63.94

Table 11F: Distracter task results for permanent CMS with a geometrical configuration change (old observers only)

- Looming – Off-time Combinations

The third set of experiments tested whether artificially expanding the text contained within a CMS message has an effect on intelligibility. CMS displays with looming text were tested against displays in which the size of the text is held constant (relative to the overall dimensions of the CMS). The results indicate that for the portable CMS displays, artificial looming in the message did not improve the observers' ability to reproduce the information contained within the message.

Looking at Tables 12A through 12D, there is no evidence of any statistically significant improvement from looming with the possible exception of young observers' word comprehension under a loom-300 ms off-time combination (at the 95% level). But since looming *alone* shows no statistically significant improvement and since a 300 ms off-time alone shows a greater improvement, the most plausible explanation is that the looming, if doing anything, is *detracting* from the positive effects of the 300 ms off-time.

	No Loom (0 ms) Standard (1)	**No Loom (0 ms) Standard (2)**	**Loom Only (0 ms off-time) (2)**	**Loom-100 ms (1)**	**Loom-200 ms (2)**	**Loom-300 ms (1)**
p value (one-tailed) from paired t-test	0.153		0.187	9.91×10^{-2}	9.49×10^{-2}	1.41×10^{-2}
Average of percent correct among all young observers	69.64	66.20	70.07	74.55	68.78	77.65

Table 12A: Word Comprehension (WC) scores and associated p values for portable CMS tests with looming and looming combined with an off-time between the frames (young observers only).

	No Loom (0 ms) Standard (1)	**No Loom (0 ms) Standard (2)**	**Loom Only (0 ms off-time) (2)**	**Loom-100 ms (1)**	**Loom-200 ms (2)**	**Loom-300 ms (1)**
p value (one-tailed) from paired t-test	0.161		0.500	0.325	0.492	0.321
Average of percent correct among all old observers	48.84	44.60	44.60	51.94	44.37	51.68

Table 12B: Word Comprehension (WC) scores and associated p values for portable CMS tests with looming and looming combined with an off-time between the frames (old observers only).

Although it did not rise to the level of significance (e.g. p = 0.0924 for loom-300 ms in Table 12D), the number comprehension score for the older observers was consistently *less* than when the messages did not loom. The overall impression is that looming definitely did not improve intelligibility and possibly reduced it.

	No Loom (0 ms) Standard (1)	No Loom (0 ms) Standard (2)	Loom Only (0 ms off-time) (2)	Loom-100 ms (1)	Loom-200 ms (2)	Loom-300 ms (1)
p value (one-tailed) from paired t-test	0.160		0.441	0.459	0.270	6.66×10^{-2}
Average of percent correct among all young observers	58.33	62.22	62.89	59.17	66.44	66.83

Table 12C: Number Comprehension (NC) scores and associated p values for portable CMS tests with looming and looming combined with an off-time between the frames (young observers only).

	No Loom (0 ms) Standard (1)	No Loom (0 ms) Standard (2)	Loom Only (0 ms off-time) (2)	Loom-100 ms (1)	Loom-200 ms (2)	Loom-300 ms (1)
p value (one-tailed) from paired t-test	0.456		0.141	9.55×10^{-2}	0.186	9.24×10^{-2}
Average of percent correct among all old observers	34.33	34.67	13.78	19.67	24.00	22.33

Table 12D: Number Comprehension (NC) scores and associated p values for portable CMS tests with looming and looming combined with an off-time between the frames (old observers only).

Comparing the distracter results for looming and looming plus off-time (Tables 12E and 12F below) with those for off-time alone (Tables 10E and 10F) a great deal of similarity can be seen. The percentage of in-lane time for young observers is very close to 95% in both cases and the older observers have in-lane percentages in the high 50's to low 60's in both cases.

Yet the (young) observers got paid less in the looming tests (the older observers seldom passing the 95% payment threshold in any case). This is, of course, because of the higher intelligibility scores (WC and NC) for off-time alone. This suggests the possibility that with the observers devoting about the same amount of attention to lane-keeping, the looming factor actually worked as *another distraction* and therefore the (young) observers had less time to spend comprehending the message.

YOUNG	No Loom 0 ms (1)	No Loom 0 ms (2)	Loom Only	Loom-100 ms	Loom-200 ms	Loom-300 ms
Percent of time within the lane (avg. of observers)	95.36	94.74	96.18	95.81	95.49	95.87

Table 12E: Distracter task results for portable CMS with the attention getting quality of looming and looming plus off-time (young observers only)

OLD	No Loom 0 ms (1)	No Loom 0 ms (2)	Loom Only	Loom-100 ms	Loom-200 ms	Loom-300 ms
Percent of time within the lane (avg. of observers)	56.89	59.39	56.34	61.44	62.28	62.91

Table 12F: Distracter task results for portable CMS with the attention getting quality of looming and looming plus off-time (old observers only)

- Change Blindness

The fourth set of experiments tested the effect of varying one specific element like a number or letter on a permanent CMS. These tests are done to see if the change blindness (CB) phenomenon occurs on CMS signs when a blank screen is used in between a two-frame message. The CB phenomenon refers to an inability to perceive that some elements in a scene have changed. Change blindness would not be expected to occur when there is no blank screen in between frames, because any change in the display would be easy to detect under the circumstances.

Given the overall best improved intelligibility scores for a 300 ms off-time, this was chosen as the blanking time between the two frames for the CB tests. There was *always* a change in some message elements between frames, but the observers were led to believe that changes between frames only occurred in *some* of the trials. This meant the observer could be asked whether he saw a change between the frames without possibly biasing his answer. A observer was given a “true/false” score for each pair of frames. If he said he saw a change, he was given a score of 1 (= “True”) for that trial, since indeed there always was a change. If he saw no change between frames, then he scored zero (= “False”) for that trial.

The word and number comprehension scores were awarded in a slightly different manner than the previous tests and the method is best explained by example. If the first frame said “DETOUR TURN LEFT NEXT EXIT” and the second frame said “DETOUR TURN RIGHT NEXT EXIT” then the observer was given credit for the words common to both frames (namely: ‘DETOUR’, ‘TURN’, ‘NEXT’, ‘EXIT’) if he wrote those words down *for either frame*. He got one word comprehension point when he wrote down the word from the first frame that got changed (‘LEFT’) and he got one point when he wrote down the word in the second frame that it was changed to (‘RIGHT’). Thus in this example, the observer could score a maximum of six: the four “overlap” words recalled in either or both frames plus the two words corresponding to the changed element. Another way of looking at the scoring is to say that only *the first* occurrences of the words were counted. The recall of numerals (NC) was scored in a similar manner.

This method of scoring may seem identical to the way of counting WC and NC in other two-frame tests such as the off-time trials. But there is a difficulty with a straight comparison because the “overlap” words in the CB tests appear in both frames thus giving the observer twice the exposure time with (typically) a lower overall distinct word count compared to the off-time only trials. Nonetheless, some basis of comparison is needed and the off-time trials offer the closest metric.

The results showed that young observers were able to correctly identify that the sign changed 84% of the time (table 13A below); however, their intelligibility scores declined (WC (71%) and NC (69%)) relative to the 300 ms off-time test (WC (82%) and NC (71%)—Tables 10A and 10C). Given the remarks of the previous paragraph, which suggest that CB tests might, all else being equal, have an advantage in scoring, the decline in scores suggests that devoting attention to looking for change elements might impair intelligibility. The slight decrease in lane-keeping percentage is also suggestive of this possibility (compare the last column of table 13A to Table 10E).

Young Observers	Percent [True=1; False=0]	Word Comprehension Total (out of 260)	Word Comp. Percentage	Number Comprehension Total (out of 215)	Number Comp. Percentage	Perc. time w/i Lane
Observer # 1	72.22	142	54.62	125	58.14	89.44
Observer # 2	66.67	190	73.08	158	73.49	93.96
Observer # 3	97.22	257	98.85	212	98.60	95.64
Observer # 4	100	179	68.85	125	58.14	97.17
Observer # 5	75.00	105	40.38	68	31.63	89.54
Observer # 6	94.44	234	90.00	203	94.42	96.72
Averages:	84.26	184.50	70.96	148.50	69.07	93.74

Table 13A: CB Phenomenon final results and distracter results for permanent CMS (young observers only).

For the older observers, the results showed that they correctly identified that the sign changed 62% of the time (Table 13B); however, their word intelligibility score declined drastically (WC (21%) versus WC (59%) in table 10B). Surprisingly, their NC score was, at 41%, higher than the corresponding value (NC (34%)—Table 10D) for the 300 ms off-time case.

Old Observers	Percent [True=1; False=0]	Word Comprehension Total (out of 260)	Word Comp. Percentage	Number Comprehension Total (out of 215)	Number Comp. Percentage	Perc. time w/i Lane
Observer # 7	60.11	119	45.77	90	41.86	80.14
Observer # 8	86.11	41	15.77	119	55.35	61.90
Observer # 9	38.89	2	0.77	53	24.65	13.77
Averages:	62.04	54.00	20.77	87.33	40.62	51.94

Table 13B: CB Phenomenon final results and distracter results for permanent CMS (old observers only).

Overall, the impression is that the change of message elements is fairly likely to be noticed, but at a cost of general intelligibility, the older observers' increases in NC score not withstanding.

- Message Compaction

The procedure used for the message compaction tests was slightly different than that of the other tests. In the interests of experimental efficiency the message compaction tests were not done as a separate run of trials as the other experiments were. Instead, data from the trials involving the standards (no loom, 0 ms off-time, CJ) was extracted from those previously run tests. If the associated CMS video had no abbreviations in it the trial became part of the "new" standard for the compaction tests. On the other hand, if the associated CMS video did have one or more abbreviations in the message then the trial was categorized as a "compact" one. Thus the data from previously run standard trials was turned into either new standard trials or message compaction trials depending on whether abbreviations were in the messages or not. This allowed data previously gathered for other uses to obviate a new set of trials dedicated solely to testing message compaction.

The only disadvantage to this procedure was making sure that the (new) standard messages were comparable (in number of words and types of messages) to the compact ones. This was accomplished by subjective but judicious selection of messages. This had the side effect of not having the same total number of words or letters for both the standard set of trials and the compact set of trials since, unlike the other experiments, the message content of a standard trial was not exactly identical to the message content of a corresponding "configuration change" trial. In other words, a staircase-justified trial (say) had exactly the same message content as its counterpart center-justified (standard) trial. That was not the case here. For example, the standard trials for the permanent CMS had a total possible word score over all trials of 226 while the compact trials had a total possible word score of 256. The numbers are "close" but not identical and given trials from each category cannot be paired up one-to-one.

Thus scoring for this section was slightly different from the usual methods of the previous sections. It proceeded along the lines used for comparing the two standards, 0 ms (1) and 0 ms (2), in the "off-time" section. The total number of words/digits recalled correctly in a given set of trials was divided by the total number possible and that percentage was compared to a similarly calculated percentage for the other set of trials. For example, observer #1 recalled 147 out of 256 words from the compact message set for the permanent CMS giving him a score of 57.42%. He got 163 out of 226 words for the standard set giving him a score of 72.12%. This is shown in Table 14A below. This table was then used to compute the p-value from the paired t-test shown in Table 14B (young observers). Here the average score on the standard (80.24%) as against the average score on the compact set (72.59%) would, if due solely to chance, only be higher 0.924% of the time.

YOUNG OBSERVERS	% Correct Standard WC	% Correct Compact WC
Observer #1	72.12	57.42
Observer #2	86.73	73.83
Observer #3	98.23	94.14
Observer #4	74.78	71.09
Observer #5	51.77	42.58
Observer #6	97.79	96.48
Average:	80.24	72.59

Table 14A: Sample table from the permanent CMS compaction tests illustrating scoring.

Table 14B (the young observers) also shows the standard with a better NC score, 70.93%, than the compact set's NC score of 67.19%. This is not significant at the 99% level as the WC score is but it is significant at the 95% level (p-value of 0.0193).

	Standard WC	Compact WC	Standard NC	Compact NC
p value (one-tailed) from paired t-test against the standard	NA	9.24×10^{-3}	NA	1.93×10^{-2}
Average of percent correct among all young observers	80.24	72.59	70.93	67.19

Table 14B: WC and NC cumulative data for permanent CMS with compacted messages compared against unabbreviated CMS (young observers only).

Table 14C shows the older observers performing better on standard messages than compact ones on the permanent CMS for both words (WC) and digits (NC) but only the WC score increase is significant at the 95% level.

	Standard WC	Compact WC	Standard NC	Compact NC
p value (one-tailed) from paired t-test	NA	4.77×10^{-2}	NA	0.233
Average of percent correct among all older observers	70.06	62.50	55.56	57.44

Table 14C: WC and NC cumulative data for permanent CMS with compacted messages compared against unabbreviated CMS (old observers only).

The same procedure was done for observers on the portable CMS runs. For young observers the standard (unabbreviated) messages were superior in both WC score and NC score to the compact (abbreviated) messages and that superiority was significant at the 99% level in both cases (table 14D).

	Standard WC	Compact WC	Standard NC	Compact NC
p value (one-tailed) from paired t-test against the standard	NA	7.58×10^{-3}	NA	7.74×10^{-3}
Average of percent correct among all young observers	70.28	62.41	67.54	53.51

Table 14D: WC and NC cumulative data for portable CMS with compacted messages compared against unabbreviated CMS (young observers only).

NC scores for the older observers (portable CMS) increased negligibly for the compact messages vis-à-vis the standard messages but that miniscule change was not significant. The WC score though was better for the standard messaging and that was significant at the 95% level (table 14E).

	Standard WC	Compact WC	Standard NC	Compact NC
p value (one-tailed) from paired t-test	NA	3.88×10^{-2}	NA	0.423
Average of percent correct among all older observersw	47.73	43.61	34.21	34.74

Table 14E: WC and NC cumulative data for portable CMS with compacted messages compared against unabbreviated CMS (old observers only).

In fact, the overall results for both portable and permanent CMS, young and old, show that the *compacted CMS made the comprehension of the sign worse*.

This raises an interesting but subtle point. These series of tests have made the quite natural (and implicit) assumption that greater amounts of information are harder to recall than lesser amounts of information. Thus recall has stood as a proxy for, and a way to quantify, the somewhat vague notion of "informational units" (see Table 1) as applied to visual messaging in a realistic environment. But these last results show that this way of measuring can diverge from both Dudek's definition and those from Information Theory (as used in statistics, computer science and electronic communication). Consider 'RDWK' as a replacement for 'ROADWORK'. By Table 1, question 1 both of these would qualify as one informational unit. If recall were a perfect stand-in for this definition then presumably compacted messages would do as well as unabbreviated messages but clearly they do not.

From a strict alphanumeric character count 'ROADWORK' obviously comes out higher than 'RDWK'. Thus a traditional bit count of information (say the dots and dashes to transmit the word over a telegraph line) would leave 'ROADWORK' with a higher information score than its abbreviated version (or have at worst the same score). Yet the abbreviated messages are clearly inferior in NC score and WC score.

This is probably because the brain's processing of the abbreviation detracts from its recall of the rest of the message. But whatever the reason, it seems clear that using recall scores as a proxy for "information units" will not give complete harmony with other proffered definitions. Nonetheless, *some* method of quantification must be fitted to any discussion of "information units" in a practical setting where humans are involved and recall/comprehension scores seem as suitable as any. In fact, recall and comprehension clearly matter more to the end-user (i.e. the driving public) than a more abstract definition of information quantification.

In addition to recording the WC and NC for the compacted CMS, the distracter task data was also recorded. The results showed that even though the drivers did not do well in recognizing the abbreviated CMS, the CMS sign did not distract them from completing the task of lane keeping. The percentage of time spent within the lane for both test cases were comparable to the standards. The younger observers did a vastly better job than the older observers, which is not surprising since the older observers have shown similar results in all the previous experiments for this project. Tables 14E and 14F both show the distracter task data taken for the last set of configuration experiments.

YOUNG	Standard	Compact	OLD	Standard	Compact
Percent Time w/i Lane	95.58	94.91	Percent Time w/i Lane	62.00	62.14

Table 14E: Distracter task results for permanent CMS with compacted CMS compared against the unabbreviated (young and old observers).

YOUNG	Standard	Compact	OLD	Standard	Compact
Percent Time w/i Lane	95.38	94.93	Percent Time w/i Lane	57.24	61.15

Table 14F: Distracter task results for portable CMS with compacted CMS compared against the unabbreviated (young and old observers)

Final Results—Informational Unit Tests

As mentioned above, a separate series of tests was run after the main series of tests (message configuration) was completed. A different set of observers was used with some overlap from the previous set but no breakdown by age was attempted, the bulk of the ten observers being younger.

These tests were run to validate the commonsensical, but unproven, notion that consecutive alphanumeric characters *with no obvious pattern*, treated as one item (a “string” in computer science parlance), have more “information content” than a single character treated as one item. In fact, this indeed turned out to be the case at least with recall scores being the metric (*higher* scores denoting *lower* information content). Unlike the compaction tests, these tests did more closely align with the standard notions from Information Theory.

As noted earlier, the observers experienced forty trials using the simulated permanent CMS. Twenty of those contained a randomly generated truck license plate in the message. The other twenty contained a character denoting direction, ‘E’ for East or ‘W’ for West. The observers were only required to recall the direction character or the license plate characters. They were not asked to recall any other parts of the messages.

The directional character message videos and the license plate message videos were interleaved in a random order. Additionally when a directional character message was called for, the choice of ‘E’ or ‘W’ was also determined at random. Although once the random orders of presentation were determined they were fixed and were the same for all observers.

The truck license plate has 7 characters. It starts with a digit, followed by a letter, followed by 5 digits. A computer program pseudo-randomly generated the plates. The letter ‘O’ (“oh”) was rejected from the “letter position” since it could be confused with the number ‘0’ (zero). In fact if the observer entered ‘O’ the software converted this to a zero before recording the response, the two characters not being distinguishable on a CMS message. Likewise, lowercase keyboard responses were converted to uppercase to allow automatic matching to a key, the case distinction not being relevant here.

The distracter task was the same as for the final configuration tests.

For purposes of scoring to compare informational content, a license plate recall was considered correct if *all* characters were recalled in the correct order. Any deviation from a perfect match was considered a miss. The number of correct responses was divided by the number of license plates shown. This latter number was usually 20 but a computer entry error occurred for a couple of observers that was not noticed until after the experiment was completed (the data being recorded automatically). These unrecorded entries were excluded from the analysis. Thus observer #2’s responses, for instance, were out of a possible 19. Given the very strong results (see below) these unrecorded few entries are not likely to bias the data to any important extent. The fraction correct was multiplied by 100 to get a percentage correct (Table 15A).

The “E or W” responses were obviously graded correct if an ‘E’ appeared on the message video and an ‘E’ was recorded and likewise for ‘W’. The number of correct responses was divided by the number of possible correct responses and multiplied by 100 to get a percentage. It would seem impossible to make a mistake on this but a couple of observers did (Table 15A). Since the observers could be run “automatically” and were not generally observed while entering data, it is possible that an unusual amount of time elapsed between seeing the video and recording the response such as if the observer engaged someone in conversation or took a break.

The results are shown in Table 15A. Every observer was superior at “directional” (E/W) recall in comparison to license plate recall. This is not a statistical fluke: the p-value (one-sided from a paired t-test) for the table is 1.04×10^{-4}.

Observer Number	% E-W Correct	% Lic. Pl. Correct
1	100	95
2	94.74	36.84
3	100	45
4	100	75
5	95	50
6	100	75
7	100	70
8	100	85
9	100	75
10	100	65
Average:	98.97	67.18

Table 15A: Percentage correct for the two types of messages testing informational content

With only a one in ten thousand chance of the directional recall score actually being equal or inferior to the license plate recall score, it is clear that *a “random” (no obvious pattern) license plate possesses more informational content than a single character,* at least when using recall as a (inverse) metric for information content.

Note the method here of *reductio ad absurdum*. Recall is used as a metric or proxy for information content (*inverse* metric actually—better recall shows less information content). This is a reasonable assumption but see the discussion of pitfalls under the message compaction section. Then license plates are treated as a single item of information (1 “informational unit”) by scoring them the same way as the directional characters, simply all right or all wrong with no partial credit. This is the unreasonable assumption that we wish to disprove. Since we are lead to the conclusion that (statistically speaking) the two types of information are always recalled at consistently unequal rates, there is a contradiction in considering the two types of information as equivalent. We can conclude that the directional character has less “information content” than the license plate; they should **not** be regarded as equivalent in terms of information units.

This validation of their unequal information content is in accord with traditional measures of information from Information Theory. Messages containing patterns have to be distinguished from those without patterns. Personalized or “vanity” license plates are usually easier to recall than “random” plates. Similarly, mnemonics can pack the same information into fewer characters than the original message. A formula can reproduce a very long string of numbers and so on. Provided that a string is random though, it cannot be reproduced in fewer bits than the number itself is expressed in. The minimal number of bits needed to reproduce a message is thus a measure of its information content. Obviously, ‘E’ requires fewer bits to record or transmit than ‘9L14317’ say.

This series of tests also examined whether there was any correlation between license plate recall and distracter task performance. For this aspect of the tests, license plate recall was scored in a different manner than that discussed above. A observer’s response in remembering the plate characters was (left) aligned with the correct license tag stored in a grading key. A program then matched the two strings character by character. Every character in the response that did not match its counterpart in the key was counted as wrong. The number of wrong characters was subtracted from 7 and the result divided by 7 and multiplied by 100 to get a percentage for that plate response.

Thus if the observer typed ‘9L14318’ when the actual string was ‘9L14317’, he got one character wrong and his score for that plate was $\frac{6}{7} \times 100 = 85.71\%$. Potential problems arise with this scoring method when other types of incorrect responses occur. For example, a response of ‘89L1431’ in the above example would score zero because none of the characters match their counterpart in the key, yet 6 of the characters match the key and in the correct order, just not in the correct place. Similarly, what if the first 7 characters match up but the response has an extra character on the end (i.e. 8 total characters in the response)? Since there is nothing in the key to match the last character this would count as one error. Should transpositions count as one error or two? And so on. There are, however, a very large number of ways to incorrectly recall a license tag and treating all the possibilities differently would make administering the scoring system too convoluted. Thus the system described above, even with its imperfections, was used.

Only license plate response scores were tallied in this portion of testing. They were put into a column in a spreadsheet and the corresponding “time-in-lane” percentages from the distracter task (computed as before) were put into the spreadsheet’s adjoining column. The goal was to see if there was a correlation between the in-lane percentages and the license plate scores. The thought was that a higher score in license tag recall would be associated with poorer in-lane performance, the observer devoting more effort toward recall and less toward lane keeping.

In the actual analysis though, this turned out not to be the case. The data from all observers was combined and the (linear) correlation coefficient for the two variables (columns) was calculated as r = 0.170. The correlation coefficient was positive! Observers did *better* in lane keeping when they did better in recall (assuming a linear relationship exists between the two variables). Is it possible that the variation in r would allow it to be consistent with a value of zero? The statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

(where n is the number of data points) is known to have Student's distribution with n-2 degrees of freedom. The ten observers saw twenty license plates but there was one computer entry error resulting in unrecorded data making n equal to 199 (= 200-1). The value of t in this case is 2.42. For 197 degrees of freedom, the value of t_{crit} at the 95% level is $t_{0.95} = 1.6526$. In fact, there is less than a 1% chance (0.82%) that the value of r could "really" be zero or negative. Thus the value of r actually is consistent with being positive.

Rerunning the data excluding the 100% license plate response events results in a much lower value of r (= 0.0057) for which the null hypothesis (r = 0) cannot be rejected. This shows that the bulk of the positive correlation comes from the responses where the entire license plate was recalled correctly. But one is hard pressed to come up with a justification for excluding the 100% responses.

In any event, it is clear that the *correlation between time-in-lane and license plate recall is **not** negative*. Since the observers in these tests averaged very close to 95% time-in-lane for the distracter task just as in all the other experiments (young observers) it may be that the distracter task was not sufficiently distracting. But although it is admittedly a subjective view, the distracter task seemed at least as difficult as driving on a smooth highway, in daylight, with good weather in light traffic. Thus to raise the difficulty of the distraction to a much higher level where it might impinge significantly on recall may be an unrealistic simulation, akin to raising the simulated speed to 100 m.p.h.

Conclusions

In general, the computer-based simulation results showed that the CMS can be improved by various manipulations of the CMS message display including:

- timing of successive frames (on times and off times) – portable CMS only
- spatial structure of the message display – permanent CMS only

The observers' comments and results show that certain criteria and changes to the CMS can decrease the intelligibility of the CMS message including:

- movement vs. steady presentation of the message display – permanent and portable displays.
- compacting the CMS information – permanent and portable CMS

For the insertion of a blank “off screen” between successive frames in a portable CMS, the duration of 300 ms for the “off screen” proved to have the greatest (statistically significant) improvement in both word comprehension and number comprehension. The findings show that the “off screen” lasting 300 ms enables the observer to better process the informational content on the CMS screen, possibly because a refractory period is needed in the processing of information in between screens.

The effect of varying the spatial structure of a permanent CMS shows that marked improvements can be made to the existing CMS message display. For younger observers, the staircase-justified (SJ) case and left-justified (LJ) case both showed strong improvements over the Caltrans standard of center-justified (CJ), with the SJ case showing stronger results than the LJ case. Among the older observers the only statistically significant improvement was in word comprehension for the SJ presentation. Still, the overall results indicate that both left and staircase justification improve an observer’s ability to reproduce information content contained within a CMS. It is likely that the geometric alignment of the staircase-justified text fits the natural progression of the observers’ eye movements as they try to read the content from a message, and the left-justified text fits the conditioning of the human eye to read left-justified text in a paper, book, or computer.

The third positive test result was the change blindness test where observers were asked to correctly detect if certain parts of a message have changed given that most of the message has remained constant. This phenomenon refers to the inability to perceive that some elements in a scene have changed. The results showed that the younger observers were able to identify that a change had occurred an average of 84% of the time while the older observers were able to identify a change 62% of the time. However, despite observers being much more likely to notice a change than not, there was a decrease in intelligibility scores compared to tests involving messages without changes.

The artificial expansion (looming) of the CMS text had a minimal effect on intelligibility. The preliminary tests on the pure looming permanent CMS texts showed only marginal improvement in comparison to no looming and that minimal gain was not even close to being statistically significant. The looming, and looming combined with off-time, tests for the portable CMS showed no statistically significant improvement over the standard presentation of CMS messages, with the exception of the looming combined with a 300 ms blanking time. But even here the effect of looming causing an improvement is doubtful as the results were less pronounced than with a presentation with the 300 ms off-time *alone*. This suggests that the looming may have *detracted* from the intelligibility rather than having some type of positive synergistic effect with the off-time. Additionally, while the scores were not statistically significant, number comprehension among older observers fell dramatically for all the looming tests on the portable CMS. Taken in total, our results indicate that these modifications to the manner in which CMS messages are displayed degrade the intelligibility of these messages.

Compacting the CMS message did not improve the intelligibility of the CMS messages. The use of common abbreviations to fit the necessary information into the limited space on the CMS proved to degrade the observers' ability to reproduce the informational content of the CMS message because the abbreviations quite probably caused the observers to stop reading the rest of the message due to confusion with the abbreviated words themselves.

The results of compacting CMS messages along with the results of the later tests done specifically to elucidate the concepts of "informational content", show that recall of CMS messages is a reasonable inverse measure of information content. This method of quantification has both overlap with and divergence from traditional measures of information used in Information Theory. It also shows the (perhaps obvious) need to distinguish between a string that is random (typical license plates) and one that is not (i.e. English words) when it comes to recall. Refinements of Dudek's definitions (Table 1) should take this into account.

Overall, the research showed that the CMS has many avenues in which it can be improved and a field study should be the next progression in the research in order to implement the changes and apply them to a real world setting.

References

Bushman, R. and Taylor, B. (2000) Dynamic safety solutions: the problem with traditional safety warnings is that they tend to be rather static which can reduce their impact., *Traffic Technology International*, August/Sept. p. 72-3.

Cohn, T E and Nguyen, K. (2002a) Tortoise Beats the Hare Again: Turning on Parts of a Warning Signal with Some Delay Makes It Seen Faster, *Proceedings of the TRB Visibility Meeting*, Iowa City, June 2002.

Cohn T. E. and K. Nguyen (2002b) "Turning it on piecemeal makes it seen faster" Abstracts of the Vision Sciences Annual Meeting, http://journalofvision.org/2/7/232/ Journal of Vision, ISSN 1534-7362

Copp, R., (2002) Caltrans. Email message to Conrad Dudek, October 9,.

Dudek, C.L. and H.B. Jones. Evaluation of Real-Time Visual Information Displays for Urban Freeways. In *Highway Research Record 366*, TRB, National Research Council, Washington, 1971, pp 64-76.

Dudek, C.L., and R. D. Huchingson, R.D. Williams, R. J. Koppa (1981). *Human Factors Design of Dynamic Visual and Auditory Displays for Metropolitan Traffic Management; Vol. 2 – Dynamic Visual Displays*. Report No. FHWA/RD-81/040. January.

Dudek, C.L., N. Trout, S. Booth, and G. Ullman. *Improved Dynamic Message Sign (*2000) *Messages and Operations*. Report No. FHWA/TX/-01/1882-2, Texas Department of Transportation, October.

Dudek, C.L. (2001) *Variable Message Sign Operations Manual*. Report No. FHWA-NJ-2001-10, New Jersey Department of Transportation, November.

Dudek, C.L. and G.L. Ullman. (2002) *Flashing Messages, Flashing Lines, and Alternating One-Line on Changeable Message Signs*. Paper presented at the 2002 Meeting of the Transportation Research Board, Washington, D.C. January.

Dudek, C.L. (2002) *Guidelines for Changeable Message Sign Messages.* Report No. FHWA-XX-2002-XX. FHWA, U.S. Department of Transportation, December.

Huchingson, R.D., R.W. McNees, C.L. Dudek. (1977) Survey of Motorist Route-Selection Criteria. In *Transportation Research Record 643*, TRB, National Research Council, Washington, , pp. 45-48.

Miller, J.S., B.L. Smith, B.R. Newman, and M.J. Demetsky (1995). *Development of Manuals for the Effective Use of Variable Message Signs*...Report No. VTRC 95-R15, Virginia Department of Transportation, January.

Pretty, R.L. and D.E. Cleveland. (1970) *The Effects of Dynamic Routing Information Signs on Route Selection and Freeway Corridor Operations*. HSRI Report No. TrS-4, Contract NCHPR 20-3, Highway Research Board, National Cooperative Highway Research Program, National Academy of Sciences,.

Proffitt, D.R. and M.M. Wade. (1998) *Creating Effective Variable Message Signs: Human Factors Issues*. Report No. VTRC 98-CR31. University of Virginia, Charlottesville, Virginia. March.

Stockton, W. R., C.L. Dudek, D.B. Fambro and C.J. Messer. (1976) Evaluation of Selected Messages and Codes for Real-Time Motorist Information Displays. In *Transportation Research Record 600,* TRB, National Research Council, Washington, , pp. 40-41.

Appendix A: Calculating the Approach Speed of CMS Sign (Simulation Calculations)

Real World Representation:

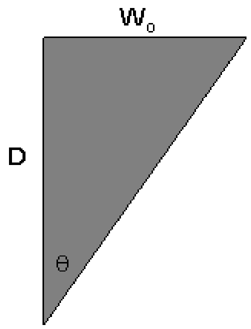
V_0 = constant velocity at freeway speeds [70 mph – 55 mph]

W_0 = width of CMS [306.06 in. for Permanent CMS or 86 in. for Portable CMS]

D_0 = original distance from the sign [1026.7 ft – 800 ft]

$$\boxed{D = D_0 - V_0 {*} t}$$

$$\boxed{\Theta = \tan^{-1}(W_0/D)}$$

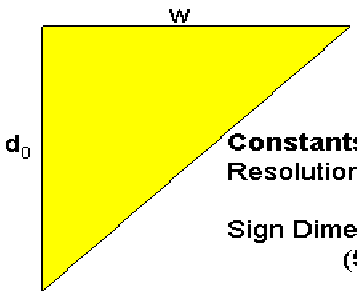


Computer/Simulation Environment:

w = increasing width size of CMS on screen

d_0 = distance from computer screen (19.25 in.)

$$\boxed{(W_0 / D) = (w / d_0)}$$
$$\downarrow$$
$$\boxed{w = (W_0 {*} d_0) / D}$$
$$\downarrow$$
$$\boxed{w = (W_0 {*} d_0) / (D_0 - V_0 {*} t)}$$



Constants:
Resolution: 1024 X768

Sign Dimensions on Screen:
(580 x 150)

.25 mm grille pitch = 100 pixels/ in

From these equations, one is able vary the width of the CMS on the screen using discrete points for the distance (D_0) and time (t) given that the W_0 and d_0 are measured quantities. Using a simple mathematical program, the relative increase in the width of the sign can be calculated over time.

Appendix B: Comments on Word (Number) Comprehension Scoring

If the ith trial has b_i words (or numerals) that *should* be recalled and a_i *actually are* and there n trials in the set, then there are two possible ways of giving an overall WC (or NC) score to the set of trials. One score uses the totals,

$$S_A \equiv \frac{\sum_{i=1}^{n} a_i}{\sum_{j=1}^{n} b_j} \tag{A.1}$$

while the other averages each trial,

$$S_B \equiv \frac{\sum_{i=1}^{n} \frac{a_i}{b_i}}{n} \;. \tag{A.2}$$

As mentioned in the body of the report, these scores are **not** in general numerically equal. When are they numerically equal? If the observer were to recall a consistent fraction, f, of the words of each trial regardless of the number of words in a trial then these measures are equal. In other words, if $a_i = f\,b_i$ independent of i, then

$$S_A = \frac{\sum_{i=1}^{n} f\,b_i}{\sum_{j=1}^{n} b_j} = \frac{f\sum_{i=1}^{n} b_i}{\sum_{j=1}^{n} b_j} = f \quad and \quad S_B = \frac{\sum_{i=1}^{n} \frac{f\,b_i}{b_i}}{n} = f\,\frac{n}{n} = f\,. \tag{A.3}$$

Generally speaking however, a more realistic expectation would be poorer recall for a trial involving more words. Take the case of three trials (n = 3) where the b_i are arranged in increasing order: $b_1 < b_2 < b_3$, and the associated fractional recall is in descending order: $f_1 > f_2 > f_3$. Under these conditions, it can be shown that $S_B - S_A > 0$.

Thus we would likely expect to see S_B (averages) come out larger than S_A (totals) in most cases. This is seen in the real-life case of table 2 where the WC score involving averages came out as 72.52% while the WC score using totals came out as 68.99%.

Of course $S_B - S_A$ has a chance of coming out negative as can be seen in the simple, made-up example using two trials (near table 2 in the body of the report). In practice the two measures seem to be close enough that there is not much to recommend one over the other. Thus we used the simpler scoring method—totals (S_A).

Field Test Report

Optimizing Comprehension of Changeable Message Signs (CMS)

Prepared for

Caltrans TO 5203: Optimizing the CMS
California PATH

By

Visual Detection Laboratory
University of California, Berkeley
360 Minor Hall
Berkeley, CA 94720

Kent Christianson
Daniel Greenhouse

CMS Field Test Report

Introduction

The goal of this research was to optimize the message content and presentation within changeable message signs (CMS) so as to maximize comprehension of those messages with minimal disruption to traffic flow. There were two parts to this research: laboratory tests and a field test.

Previously, the Visual Detection Laboratory (VDL) had conducted tests in the laboratory to maximize comprehension of messages[1]. These tests consisted of an observer viewing an animated, computer simulated roadway scene with a CMS in various configurations as the observer simultaneously performed a "distracter task". The task was an analog of steering and was used to prevent the observer from devoting all of his/her attention to reading the message signs. The observer was then asked to recall the message content after the trial run and scored on the recall accordingly. This process was repeated extensively for many runs, many observers and many CMS configurations and the scores analyzed statistically.

While the results of these lab tests were then used to inform the design of the field test, a duplication of the full set of lab tests was impractical because the number of test configurations was very large, and because we could not allow any field test to compromise the safety of the driving public. If any of the tests were to result in failing to convey an important message to the public (contrary to our expectations), then the experiment could have had a detrimental impact on motorists. Unlike lab volunteers, motorists could not be asked for their recall of the CMS content immediately after viewing it. At best they could be surveyed, if they chose to do so, at some point down the road where their responses could be recorded safely (i.e. a stoplight at an exit ramp). The people agreeing to such a survey would be unlikely to be a random sample, and the delay in time between viewing the CMS and responding to a survey would greatly degrade the accuracy of their responses. Even if it had possible in theory to introduce such a survey, the strict rules of the Committee for Protection of Human Subjects at the University of California would have rendered it impossible to implement. Thus VDL concluded that the optimization of content and presentation would have to be solely based on the in-lab tests and the field work reserved for testing the goal of minimal disruption to traffic flow under changed message conditions.

In order to avoid any inadvertent, detrimental impact on the public, the message used for testing could not be critical to the motorists. In other words, a test message like *"WARNING STOPPED TRAFFIC AHEAD"* was out of the question. The test message could obviously not be false and if true VDL could not, on the very first field test of this sort, risk such a critical message being ineffectively conveyed to the public, despite our

[1] See ***Optimizing Comprehension of Changeable Message Signs (CMS)*** **[Final Laboratory Report - Caltrans TO 5203]**

belief that our message configurations (based on in-lab testing) are superior at conveying the message content. Hence VDL came to the conclusion that there would be only one test message and one test configuration and that that message would be "innocuous". The message chosen was *"REPORT DRUNK DRIVERS CALL 911"*. The details of this testing are given in the next section.

Testing Configuration

The field test was conducted along a stretch of westbound Interstate 80 that constitutes part of the Berkeley Highway Lab (BHL). A schematic of the BHL is shown below.

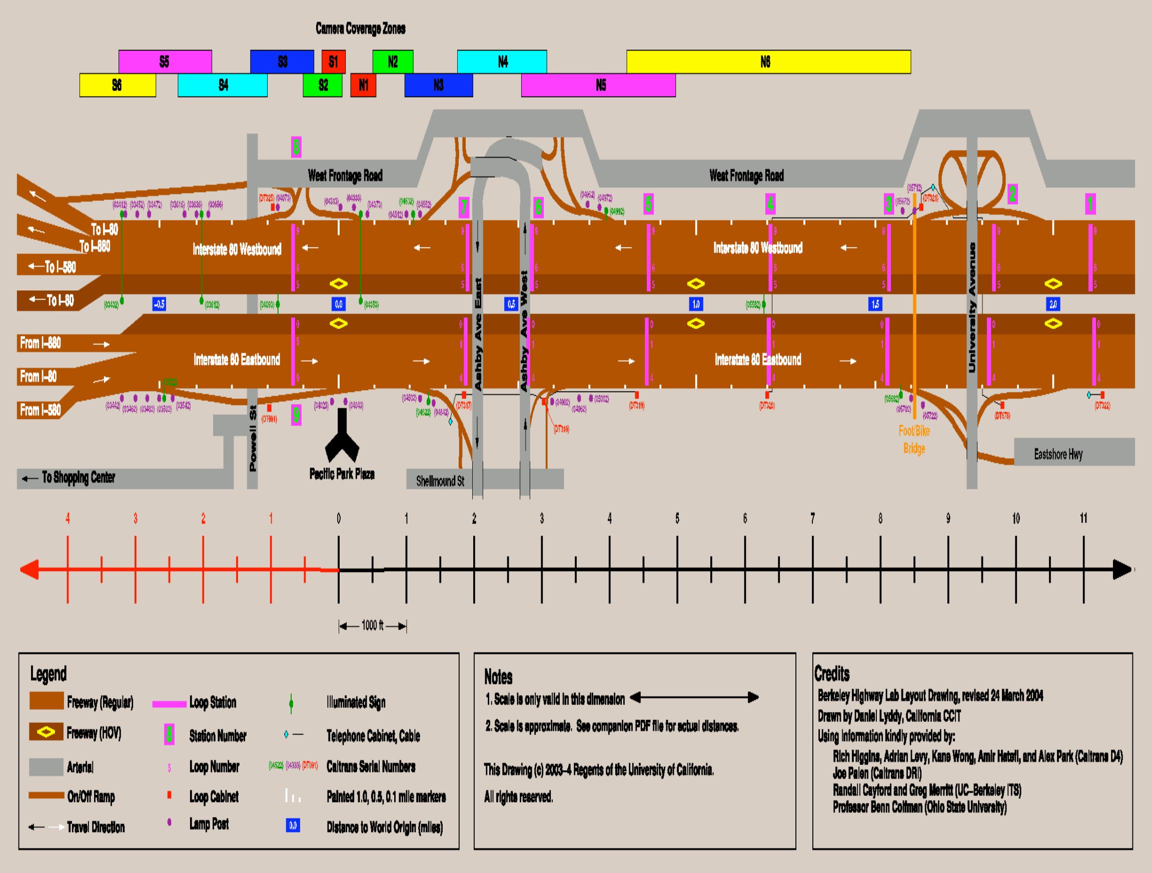


Figure 1: A schematic showing the Berkeley Highway Lab (BHL). It constitutes a portion of I-80 in Berkeley and Emeryville, California that is highly instrumented to measure traffic properties.

The BHL is a 2.7 mile portion of I-80 that has been outfitted with large quantity of instrumentation (video monitoring and loop detectors) in order to learn about traffic properties and serve as a testbed for research in transportation. Data from BHL can be accessed from CCIT (California Center for Innovative Transportation). (A brief description of BHL can be found at: http://www.calccit.org/projects/atms.html.)

A permanent CMS (figure 2) is located on the westbound stretch of freeway between detector stations 5 and 6. A blowup of the relevant roadway section from the schematic is shown in figure 3.

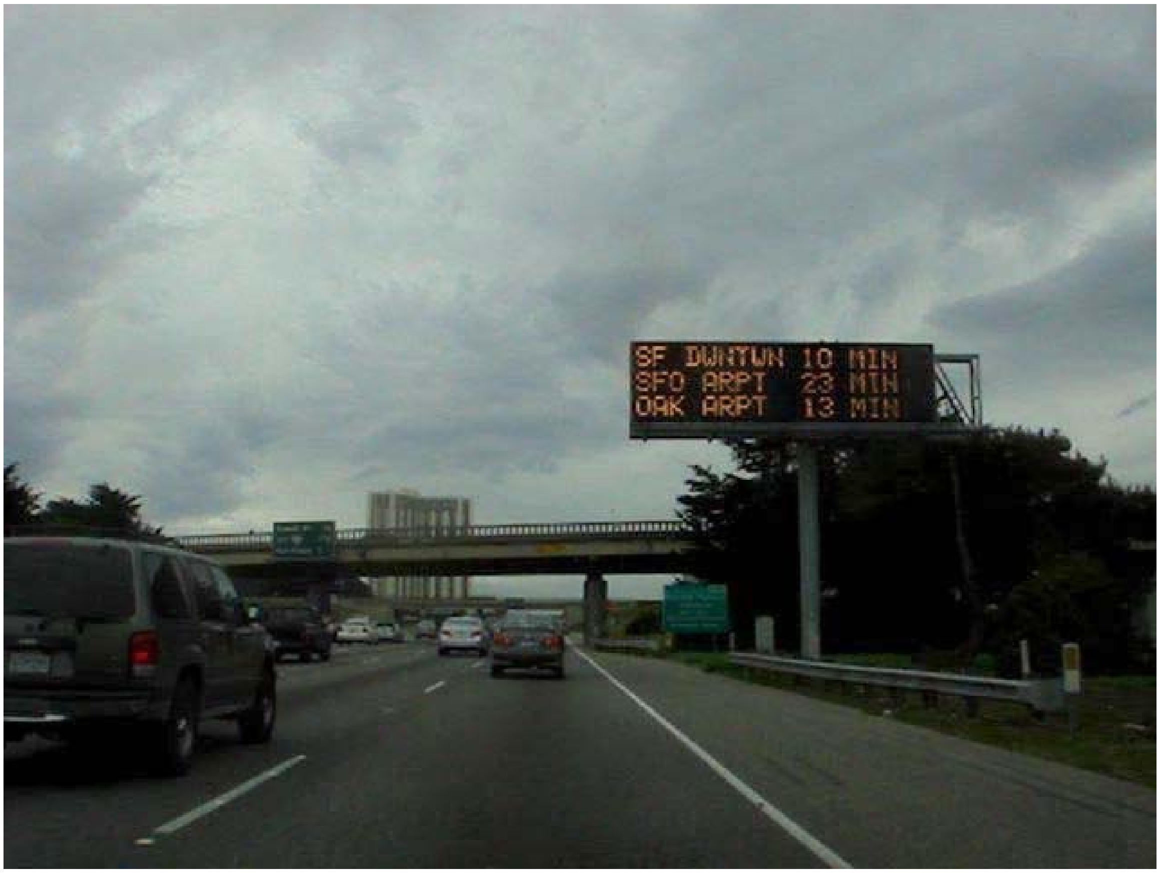


Figure 2: The permanent CMS used for the test as it looks from the rightmost lane (this picture was not taken at test time).

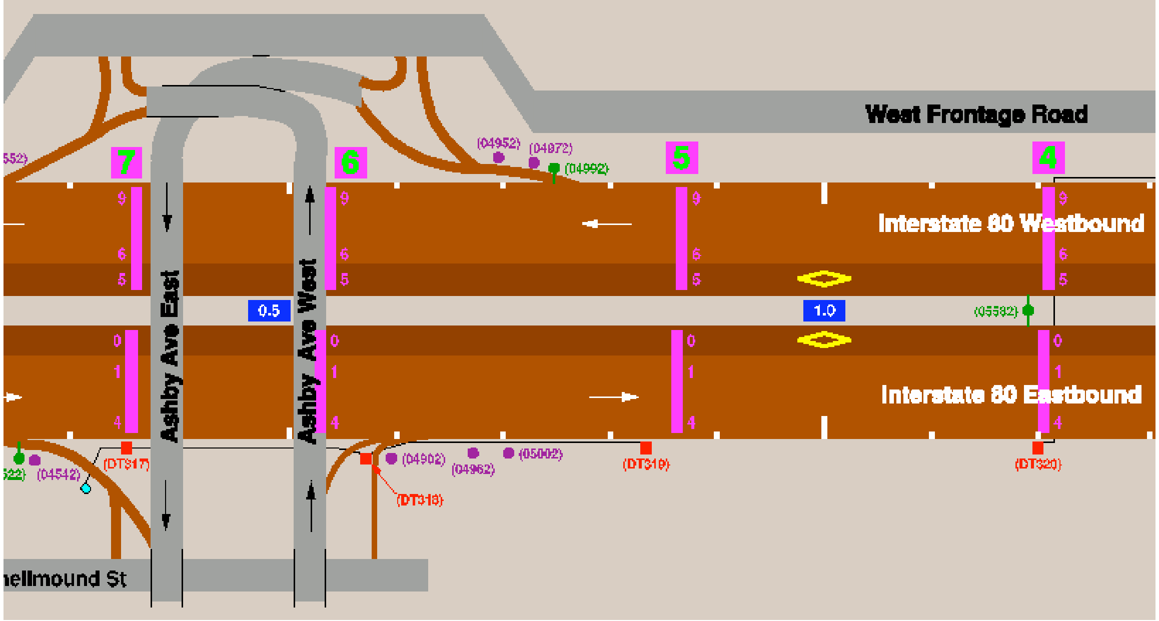
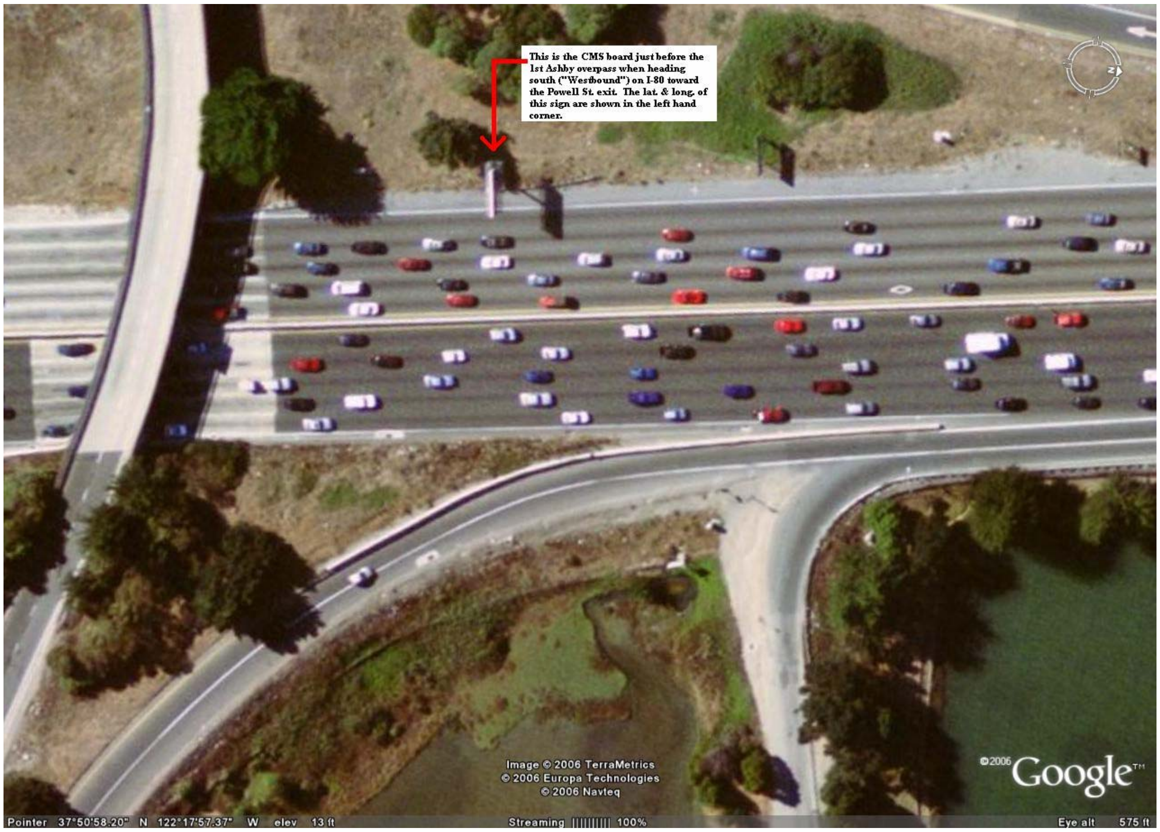


Figure 3: Schematic of the portion of I-80 used for the test. Only data from stations 5 and 6 (labeled in pink) on the westbound lanes was relevant to the CMS test (see text below).

Referring to figure 3, the broad pink stripes represent a series of loop detectors that run across the road. Each set of detectors at a given interval is a station. The station number is shown in green on a pink square. The red squares are the corresponding loop (or station) cabinets, the box housing the electronics associated with the loop detectors. The small numbers in pink are the lane designations for the loop detectors. This numbering system should not be confused with the one *Caltrans* uses for highway design.[2]

The CMS is not shown on the schematic but its location can be seen in figure 4, an aerial view of the actual scene from *Google Earth*™. The annotation with the red arrow pointing to the CMS was added to the scene after capturing the screenshot. At the scale of figure 4, it may be too small to read; the annotation merely notes that although the freeway is "westbound" at that point, the actual direction of travel is much closer to due south and it mentions the CMS coordinates given in the caption for figure 4.

[2] In this latter case, the leftmost lane in a given direction of travel is always the number one lane. See for example, http://www.dot.ca.gov/hq/oppd/hdm/pdf/chp0060.pdf.



This is the CMS board just before the 1st Ashby overpass when heading south ("Westbound") on I-80 toward the Powell St. exit. The lat. & long. of this sign are shown in the left hand corner.

Image © 2006 TerraMetrics
© 2006 Europa Technologies
© 2006 Navteq
©2006 Google™
Pointer 37°50'58.20" N 122°17'57.37" W elev 13 ft
Streaming ||||||||| 100%
Eye alt 575 ft

Figure 4: Location of the CMS (red arrow) as shown on a Google Earth aerial image. The sign is at 37º 50′ 58.20″ N latitude and 122º 17′ 57.37″ W longitude. The station 6 loop cabinet is the white square near where the two roadways at the bottom of the picture merge to become the onramp for the eastbound freeway.

Given the depiction in figure 3 and the scene in figure 4, one may be tempted to conclude that station 6 lies "downstream" of the CMS, nearly coincident with the Ashby overpass (also shown in figure 4). This is in fact, <u>not</u> correct. The drawing in figure 3 is meant to be approximate with respect to position. The station 6 detectors are supposed to lie "inline" (i.e. across the roadway) with the loop cabinet (white square cabinet—see figure 4 caption) or as close to it as possible. According to CCIT, the loop detectors for station 6 are at least 60 feet north ("upstream") of the CMS.[3] This is important, as a detector "downstream" of the CMS is much less likely to register changes in driver speed due to driver reaction to the CMS than one that is just "upstream" of it; common sense suggests most drivers would react to a sign upon reading it—a delayed reaction being unlikely.

In order to test whether station 5 data needed to be used, a member of the VDL staff drove this section of the freeway and was easily able to see the sign well north ("upstream") of station 5 (by an estimated 900 feet). Thus if driver speed could change in reaction to the CMS display, the response could show up at stations 5 or 6. Stations 4 or 7 were deemed very unlikely to be useful for the test.

[3] Personal communication with Mr. Saneesh Apte, Associate Development Engineer at CCIT.

Based on the latitude and longitude of the stations supplied by CCIT[4], the sign and detector positions are shown in figure 5 along with the best estimates of their separations.

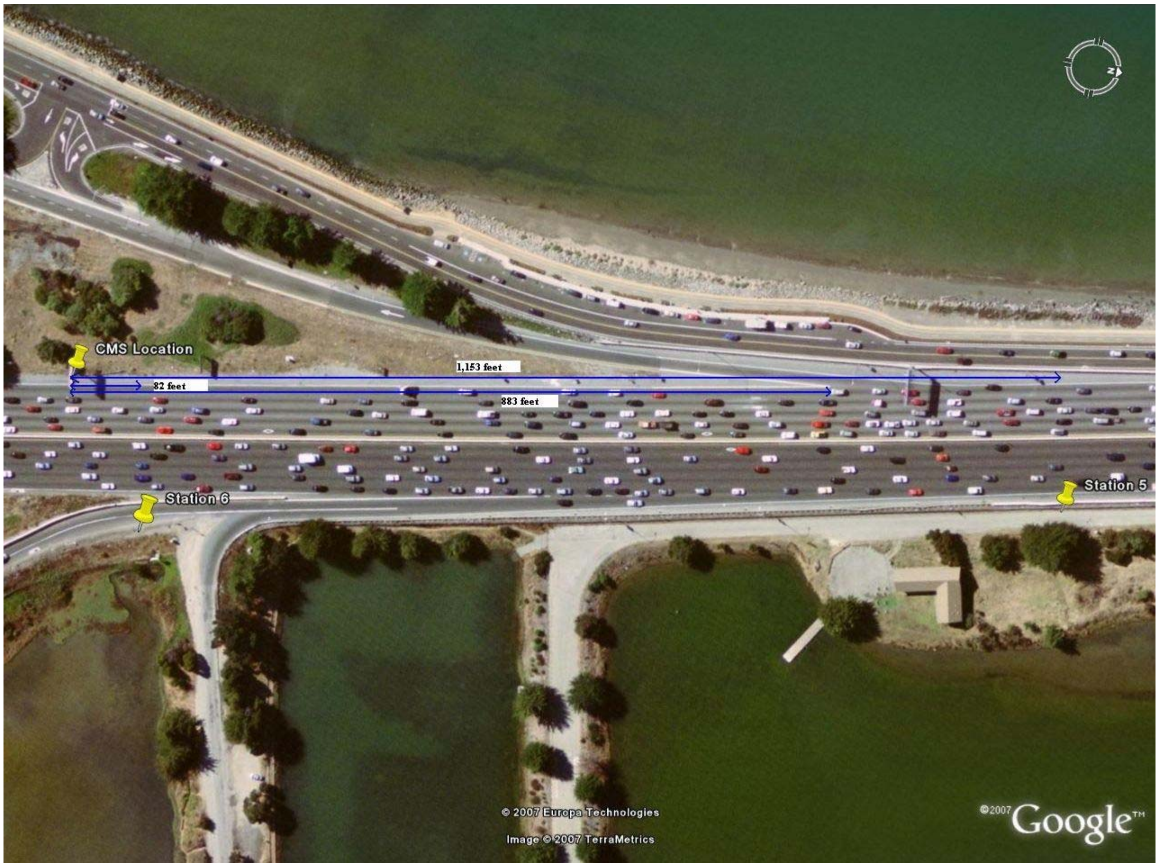


Figure 5: The line of loop detectors at station 6 are estimated to be around 80 feet "upstream" of the CMS, while those of station 5 are around 1,150 feet "upstream" of the CMS. The sign and stations are marked with superimposed yellow thumbtacks.

Speed data for lanes 5 through 9 (numbered as described above) at stations 5 and 6 was therefore taken under the conditions described below. The speed data consisted of vehicle speed averages over 30 second intervals (in miles per hour) for every lane at each station for the duration of the tests.

Two two-hour tests were conducted, one during the day (noon to 2 p.m.) and the other at night (10 p.m. to midnight) on September 13, 2006. During each two-hour test, the CMS operators put a standard version of a message onto the CMS for ten minutes then a blanking period (no message) for five minutes followed by the test version of the message for ten minutes and then this was followed by another blanking period. The cycle then repeated for the full two hours.

[4] For station 5: 37° 51′ 9.72″ N latitude and 122° 17′ 58.2″ W longitude.
For station 6: 37° 50′ 59.28″ N latitude and 122° 17′ 55.32″ W longitude.

This blanking or off period was necessary so that any given set of drivers would not be exposed to both the test and standard messages. Instead they would be exposed to one or the other, five minutes being judged enough for traffic to clear the section of I-80 over which the CMS was visible. Thus a particular driver would see the standard message, or the standard message and no message, or the test message, or the test message and no message, or just no message (depending on his timing). He would be unlikely to see both the standard and test message, unless traffic was unusually slow, in which case any effect from the CMS would be swamped by the traffic conditions.

The test and standard messages were alternated over short periods in order to minimize any time-of-day effects. For example, had the test message been displayed continuously from noon to 1 p.m. and the standard message from 1 p.m. to 2 p.m., then if traffic were heavier (and thus moving more slowly) during the first hour (due to say workers going to lunch) then the resulting data would have been skewed by this effect that had nothing to do with the CMS. This interleaving of the standard and test messages, along with the blanking, allowed a clean separation of the data without confounding factors that would otherwise have been a problem.

The message configurations are shown in figure 6. The actual implementations are shown (albeit somewhat roughly) in figure 7.

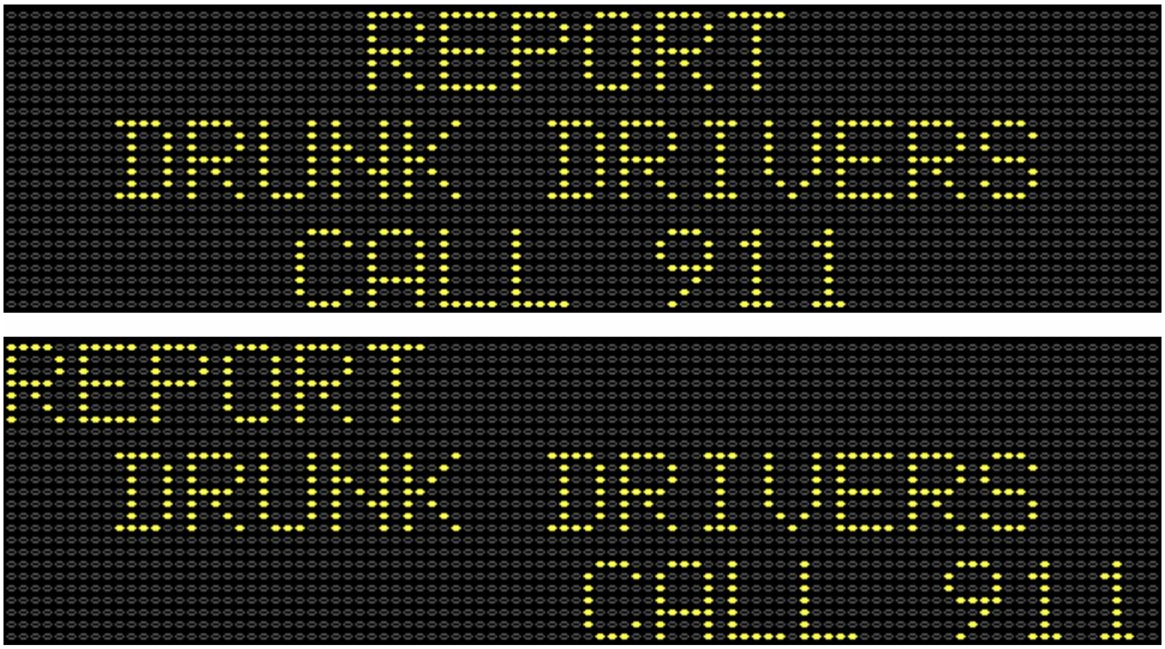


Figure 6: Shown here are the standard configuration – center justified (upper message) and the test configuration – staircase justified (lower message). Although the content of each message is identical and only the configurations are different, they will be refered to as the "standard message" and "test message" for simplicity.

As noted above, the message was chosen to be innocuous. If someone failed to comprehend it in the test configuration there was no chance of any harm as a result. Since VDL could only realistically test one configuration against the standard without unduly burdening *Caltrans* we made the decision to test one of the stronger (statistically significant) configurations found from our in-lab work. This is the so-called “staircase justified” message (see the report listed under footnote 1). In this case (lower portion of figure 6) the first line is left-justified, the second line is center-justified and the last line is right-justified. This was the “test message” while the “standard message” consisted of the same message content but with each line center-justified (the usual method *Caltrans* uses for CMS messages).

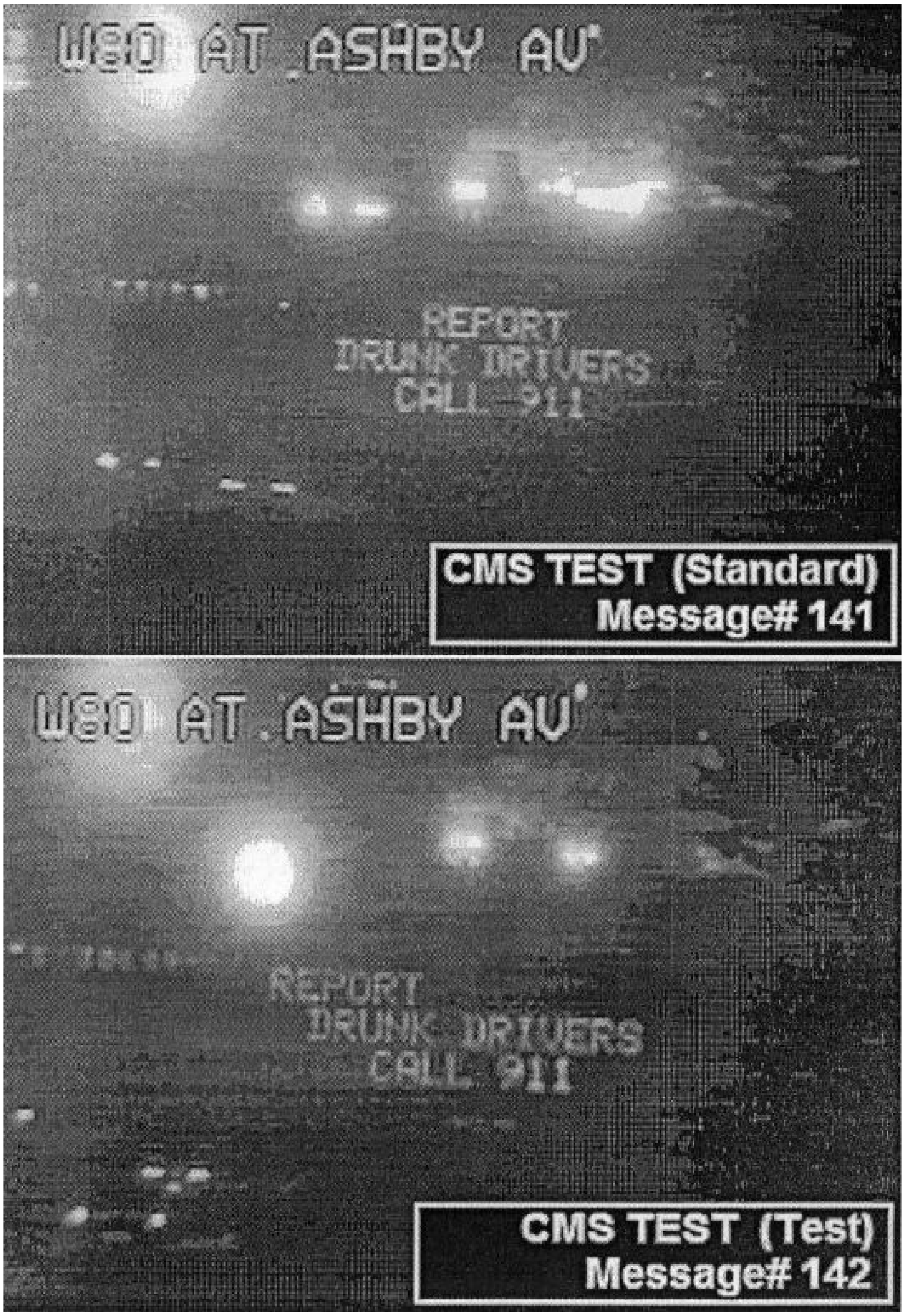


Figure 7: CMS field test as it looked in operation (courtesy of Caltrans district 4).

The point of the field testing was to see if the novel message configuration had any significant impact (positive or negative) on traffic flow. More specifically, the traffic speeds (in each lane for both day and night) were compared under the two CMS message configurations. The null hypothesis was that there would be no statistically significant difference in mean vehicle speed between the two CMS message configurations (for a given lane at a given time of day). The alternative hypothesis was that there would be a difference in vehicle speed under the two conditions.

If the null hypothesis were accepted, one could reasonably be assured that changes in message configurations (at least the kind of a change tested here) would have little impact on traffic. If the null hypothesis were rejected, it would not necessarily mean that the novel message configuration was having an adverse impact on traffic. In the first instance, the change in traffic speeds could have been positive—traffic could have been moving faster under the test message than the standard message. Even if traffic slowed under the test configuration however, this does not necessarily imply an adverse effect. "Statistically significant" isn't necessarily the same thing as "real-life" significance. If the slow down were real but trivially small then it would be hard to claim a serious disruption of traffic flow.

Data Analysis and Statistical Methodology

After we obtained permission from Caltrans District 4, we arranged for the operators of the CMS on westbound I-80 at Ashby Avenue to run the test and standard messages in the manner described above (see figure 7) and to supply VDL with the intervals (times) during which each message (or blank) ran. CCIT gave VDL the detector loop data from stations 5 and 6 for the times during which the field test was running.

This latter data was imported into Microsoft Excel. The speed data from like-condition interleaved intervals (based on the *Caltrans* timings) was aggregated onto a separate worksheet on a per-lane, time of day and station basis. In other words there was a station 5 data file and a station 6 data file. Within each file there was a day worksheet and a night worksheet. Within each worksheet there were three columns for each lane: a column showing speeds in that lane under the standard message, a column showing speeds in that lane for the test message, and a column showing speeds in that lane during the blanking (off) interval.

This kind of data rearrangement, while tedious to perform, was necessary because of all the possible confounding factors that could overwhelm any effect due to the CMS. Traffic patterns are different at night than during the day. The left lanes are meant for higher speed traffic. Station 6 is much closer to the CMS than is station 5 and therefore likely to see more of an effect (if there is one).

Another aspect of the data processing that should be noted is the following. The loop detector data consists of speed averages (in a given lane) over successive 30-second intervals, but it also includes occupancy and flow numbers for that 30-second interval. Thus the number of vehicles passing over that detector in that 30-second interval is known. It is therefore possible in principle to multiply that number by the average speed and thus weight each 30-second interval average speed by the number of vehicles used in that average. This was not done. The justification is that because each interval is small in comparison to the total time of data collection (~ 40 minutes for each message configuration) and because the collection times are spread out over two hours, any given 30 second interval with a high flow number is likely to be matched with a corresponding interval with a low flow number. Thus barring traffic jams, flow number is effectively randomized over the large number of intervals. The one exception to this was removing zero valued speed data; if no vehicle passed over the detector in a 30-second interval (rare during the day but it can happen at night) then a value of zero was reported by the detector. Clearly a zero should not enter into the sample of vehicle speeds and such zeroes were deleted.

Upon processing, there were 5 lanes x 3 conditions x 2 times of day x 2 stations = 60 sampling distributions of vehicle speeds. Initial statistics were run on 40 of these pairwise for an initial 20 tests, the blanking condition not being tested. These were two sample z-tests for means performed on each standard-test message pair for each lane for each time of day for each station[5]. Typical vehicle speed distributions are shown in figures 8 and 9.

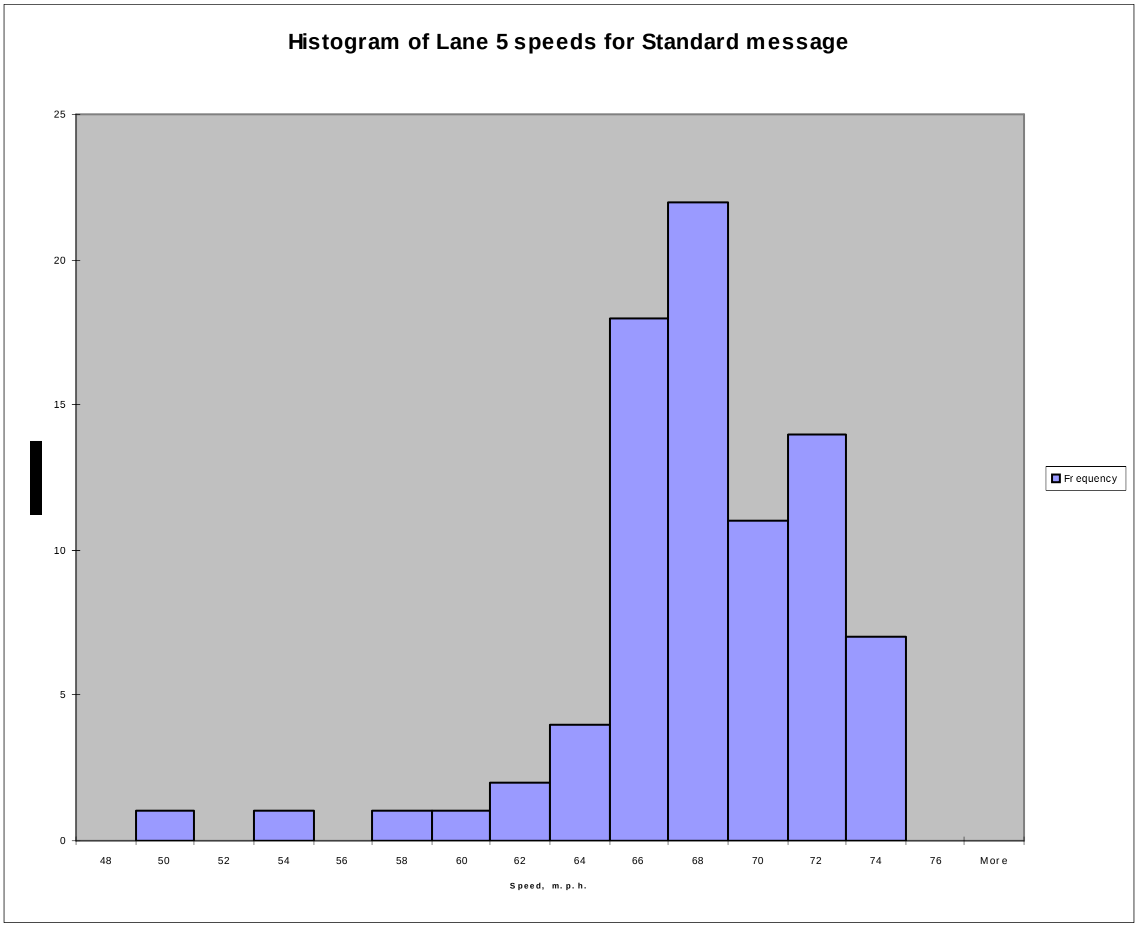


Figure 8: Histogram of lane 5 speeds for daytime testing displaying the standard message (station 6).

[5] After these initial tests, t-tests were done in selected instances—see page 14.

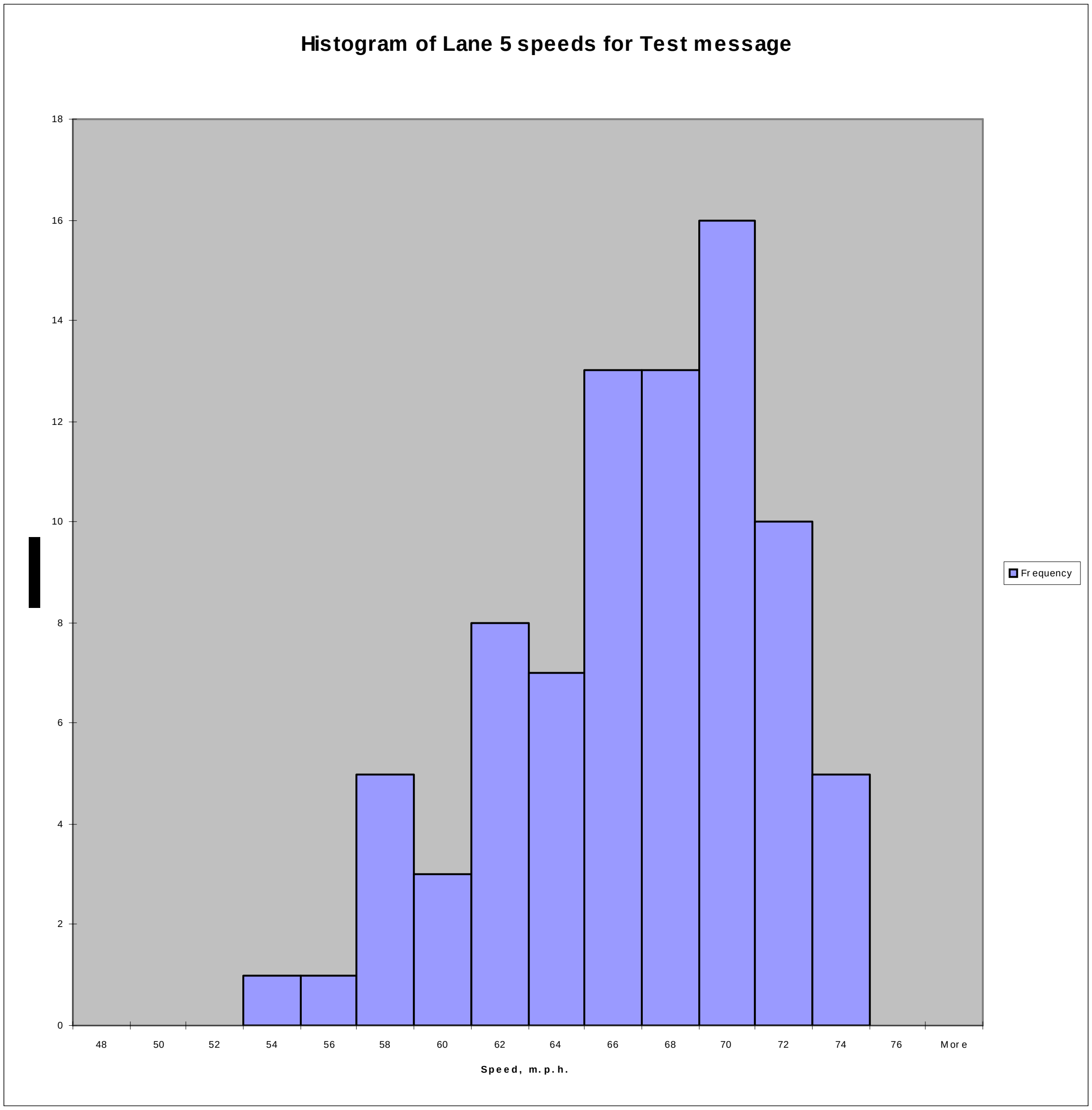


Figure 9: Histogram of lane 5 speeds for daytime testing displaying the test message (station 6).

An immediate problem presents itself: figures 8 and 9 are clearly non-normal. They skew to the left. In the case of the histogram in figure 8 the coefficient of skewness is –1.636 while that of figure 9 is –0.551.

From a driving perspective this makes some sense; larger or a greater number of excursions below the mean are less likely to result in legal sanctions than larger or a greater number of excursions above the mean—the highway patrol is less likely to give a driver a ticket for going slow than speeding. Also drivers may slow down a bit to change lanes or make an exit.

All the speed data for the standard and test configurations was tested for normality. A normal quantile plot was used. A typical example is shown in figure 10.

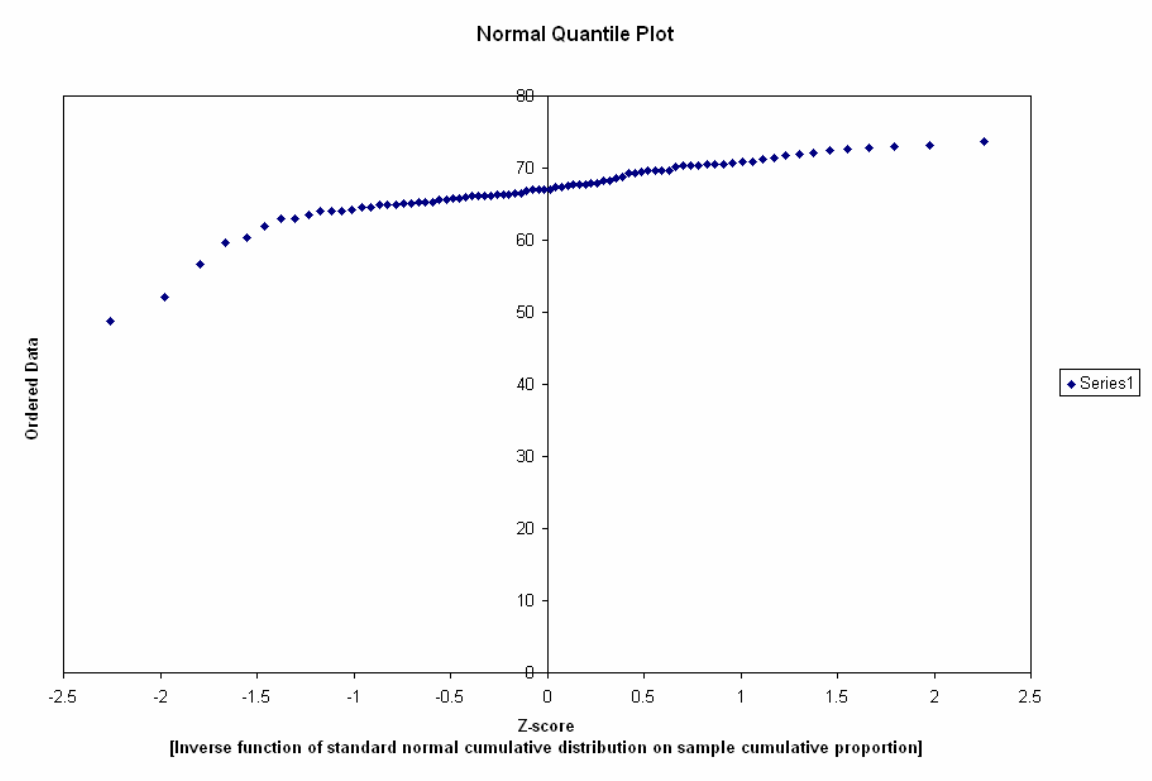


Figure 10: A normal quantile plot—speed data from lane 5 under the standard CMS configuration (shown in histogram in figure 8) plotted (after ordering) against the z-score of the speed data cumulative proportion.

To get such a plot the speed data is rearranged from lowest to highest ("ordered") in one column. Then in a companion column the cumulative proportion of the ordered data is listed. If there are n samples then the lowest value sample gives 1/(n +1), the next highest corresponds to 2/(n + 1) and so on up to n/(n + 1). Then in a third column the inverse of the standard normal cumulative distribution function (essentially the inverse of the error function, *erf*) is used on the second column. The values in the first column are then plotted against the corresponding values in the third column (the "z-score").

If the data are normally distributed then the points should fall on a straight line. If the plot deviates from a straight line then the data are likely non-normal. For example, the downward bending in figure 10 indicates a tail to the left[6] (i.e. negative skewness), which is exactly what is seen in figure 8.

This plot gives a quick visual indication of how non-normal the data is and in what kind of way, but the non-normality can also be quantified from the plot. A "correlation coefficient" can be calculated from the plot data by the usual method, although strictly

[6] See for example the illustrations on the web page at http://www.skymark.com/resources/tools/normal_test_plot.asp.

speaking the resulting number is not really a correlation coefficient in the usual sense (i.e. it does not arise from a bivariate normal distribution) since the z-score is computed from a "theoretical" point of view and is not a random variable. Nonetheless, this "correlation coefficient" can be used as a test statistic for normality. Ryan and Joiner developed this normality test and they supplied formulae and critical values for various sample sizes.[7]

The results of normality testing (using the Ryan and Joiner method) are shown in the tables in the next section. Each table represents a summary of the results for each station, for each time of day, by lane and by message configuration (test or standard). The third row of tables one through four has entries with either a Y (yes, the population is likely normal) or an N (no, the population is likely not normal). More precisely, these answers really determine if the test statistic (the "correlation coefficient") fell into the critical region or whether it stayed outside of that region where the critical value was taken at the level of significance $\alpha = 0.05$ for a *Type I* error under the null hypothesis that the sample was from a normal distribution (rather than a categorical assertion that the distribution is normal or non-normal). As an aside, it is noted that unlike many familiar statistical tests the "correlation coefficient" was in the critical region when it was *below* the critical value; this is because a correlation value of 1 would denote a normal distribution (the null hypothesis) and it forms an obvious upper bound.

For station 5's daytime run, 6 of the 10 samples could be considered normally distributed for purposes of further testing while the nighttime run had 4 out of 10 as likely normal. Station 6's daytime results had 4 out of 10, which could be called normal, but there were an additional two which fell into this category if only one extreme value was removed from each of the samples (out of typically 82 speeds in a sample); this normality test is sensitive to outliers. For station 6's nighttime results there was only one sample distribution that would be classified as normal but if the an extreme value was dropped from each of two other samples they also could classify as normal. (One of these extreme values was an astonishing 13 m.p.h.—on the freeway!)

Unfortunately the z-test, which had been used in every case, is predicated on sampling from normal distributions. Fortunately there is a strong justification for sticking with the z-test results rather than reworking all the data with non-parametric tests such as the Wilcoxon. In every case, there are two very favorable factors working in our favor: i) the sample size is "large" and ii) the sample sizes being compared (pairwise) are equal or nearly equal. In most cases there is the additional propitious circumstance of the two distributions being compared having their departures from normality be of the same character (e.g. both distributions have negative skewness).

[7] Their 1974 technical article is available on the "Minitab" (statistics software) web page at http://www.minitab.com/resources/articles/normprob.aspx and the ideas therein are expanded upon in *Goodness-of-Fit Techniques* edited by Ralph B. D'Augostino and Michael A. Stevens (Dekker, 1986). There is a strong relationship between this "correlation coefficient" normality statistic and the better-known Shapiro-Wilk test W. The advantage of the "correlation coefficient" is the ability to visualize it in the normal quantile plot.

Determining whether a sample size is "large" or "small" means comparing it to some other quantity, such as at what point a too small sample size will begin to give serious errors when taking sample parameters as population parameters, or at what size an asymptotic result holds to within a certain percentage. But as a rule-of-thumb (i.e. guides in statistics textbooks), "large" sample sizes are taken to be 30 or more. Some more conservative authors use 50. The sample sizes used for these z-tests were typically around 80 and none were lower than 76.

When sample sizes are "large" one can show that there is little difference between a two sample z-test (for means), a two sample t-test with equal variances and a two sample t-test with unequal variances (*Welch's test*). In other words, in the asymptotic limit of sample size, these test statistics approach the same values. This is important to note because *the t-test is especially robust against non-normal distributions*[8]. Furthermore the two sample t-test is even more robust than the one sample t-test.

Additionally when sample sizes are equal or nearly equal (especially when they are large), *the t-tests are robust against variance heterogeneity*[9].

Thus we did not replace all the z-tests we had already laboriously done with t-tests, but rather we did "spot checks", to make sure the known asymptotic equivalency was holding, by performing t-tests on selected samples. To give a quick example: on the lane 5 pair for night at station 5, the z-value was –0.298; the critical z (two-tail) was 1.960; the t-value (equal var.) was –0.296; the corresponding critical t (t.t.) was 1.975; the t-value (unequal var.) was –0.298 with a critical t (t.t.) of 1.975. The disagreement can be seen to be parts per thousand—not worth worrying about. Similar agreement held with the other checks.

Thus although some of our sample distributions were not likely normal, a significant fraction were consistent with normal distributions and our z-tests were equivalent (thanks to large sample size) to t-tests which are very robust against non-normality. The nearly equal sample sizes meant robustness against unequal variances so that the variance ratios didn't even need to be tested. Consequently, we can be confident that the statistical tests used are on a sure footing.

[8] See: **The Robustness of two sample tests for Means-A Reply on von Eye's Comment** by V. Guiard and D. Rasch in *Psychology Science*, vol. 46, 2004 (4) p. 549-554, **The Robustness of the Two-Sample T-test over the Pearson System** by H.O. Posten in *Journal of Statistical Computation and Simulation* vol. 6, 1978 p.295-311, and **The Robustness of Parametric Statistical Methods** by D. Rasch and V. Guiard in *Psychology Science*, vol. 46, 2004 (4) p. 175-208.

[9] And the Wilcoxon test is **not**. See the references in footnote 8 and **Robustness of the Two-Sample T-test Under Violations of the Homogeneity of Variance Assumptions** by H.O. Posten, H.C. Yeh and D.B. Owen in *Communications in Statistics: Theory and Methods,* vol. 11, 1982 p. 109-126.

Results

As mentioned above, the results of our analysis are shown in tables 1 through 4.

Each table consists of two parts—lanes 5 through 7 and 8 through 9. For each lane, under each condition (standard configuration and test configuration) the mean speed is given in miles per hour in the first row. The next row marks which condition for that lane had a greater speed. The next row, as already mentioned, deals with whether or not the sample population under those conditions can be considered coming from a Gaussian (or normal) distribution. The final row gives the results of two sample z-test. The level of significance in all cases was taken as α = 0.05. If the z value was greater than the critical value this was noted. The associated p-value is shown in square brackets.

	Lane 5 **Standard** msg.	Lane 5 **Test** msg.	Lane 6 **Standard** msg.	Lane 6 **Test** msg.	Lane 7 **Standard** msg.	Lane 7 **Test** msg.
Mean Speed m.p.h.	**67.16**	**66.05**	**62.70**	**61.56**	**59.68**	**58.54**
Which lane speed greater?	>		>		>	
Normal pop. @ α=0.05? (Y/N)	Y	N	Y	N	Y	N
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.16]		N [p=0.08]		Y [p=0.04]	

	Lane 8 **Standard** msg.	Lane 8 **Test** msg.	Lane 9 **Standard** msg.	Lane 9 **Test** msg.
Mean Speed m.p.h.	**59.70**	**59.43**	**56.71**	**56.11**
Which lane speed greater?	>		>	
Normal pop. @ α=0.05? (Y/N)	Y	N	Y	Y
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.60]		N [p=0.25]	

Table 2: Results for station 5 daytime

For station 5 in the daytime (table 1) only the lane 7 pair shows any hint of significance. None of the lane pairs in table 2 (station 5 at night) show any significance, nor do any of the pairs in tables 3 or 4 (station 6 during the day and night respectively). Thus, for the twenty lane pairs in all four tables, only one lane pair shows any significance (and then only barely at the 4% level). But one potentially significant pair out of twenty is *exactly what one would expect of chance at the* α = 0.05 *level of significance.* Thus these tests do not show any significant impact on average driving speed between the CMS displaying a message in its standard format and its test format.

	Lane 5 **Standard** msg.	Lane 5 **Test** msg.	Lane 6 **Standard** msg.	Lane 6 **Test** msg.	Lane 7 **Standard** msg.	Lane 7 **Test** msg.
Mean Speed m.p.h.	**76.53**	**76.72**	**70.42**	**71.00**	**67.47**	**67.71**
Which lane speed greater?		>		>		>
Normal pop. @ α=0.05? (Y/N)	N	N	Y	N	Y	N
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.77]		N [p=0.25]		N [p=0.68]	

	Lane 8 **Standard** msg.	Lane 8 **Test** msg.	Lane 9 **Standard** msg.	Lane 9 **Test** msg.
Mean Speed m.p.h.	**66.54**	**66.99**	**66.32**	**64.94**
Which lane speed greater?		>	>	
Normal pop. @ α=0.05? (Y/N)	Y	N	N	Y
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.48]		N [p=0.09]	

Table 3: Results for station 5 night

	Lane 5 **Standard** msg.	Lane 5 **Test** msg.	Lane 6 **Standard** msg.	Lane 6 **Test** msg.	Lane 7 **Standard** msg.	Lane 7 **Test** msg.
Mean Speed m.p.h.	**67.05**	**65.76**	**61.48**	**60.87**	**59.17**	**58.27**
Which lane speed greater?	>		>		>	
Normal pop. @ α=0.05? (Y/N)	N	Y	N [Remove lowest speed and it becomes a Y]	N	N	Y
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p = 0.062]		N [p = 0.32]		N [p = 0.10]	

	Lane 8 **Standard** msg.	Lane 8 **Test** msg.	Lane 9 **Standard** msg.	Lane 9 **Test** msg.
Mean Speed m.p.h.	**59.69**	**59.26**	**57.94**	**57.58**
Which lane speed greater?	>		>	
Normal pop. @ α=0.05? (Y/N)	Y	N	Y	N [Remove highest speed and it becomes a Y]
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.39]		N [p=0.48]	

Table 4: Results for station 6 daytime

	Lane 5 **Standard** msg.	Lane 5 **Test** msg.	Lane 6 **Standard** msg.	Lane 6 **Test** msg.	Lane 7 **Standard** msg.	Lane 7 **Test** msg.
Mean Speed m.p.h.	**75.28**	**75.88**	**69.23**	**69.72**	**66.01**	**67.13**
Which lane speed greater?		>		>		>
Normal pop. @ α=0.05? (Y/N)	N	N	N	N [But drop highest speed and it becomes a Y]	N [But drop lowest-> 13.5 mph! And it is Y]	Y
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.44]		N [p=0.36]		N [p=0.21]	

	Lane 8 **Standard** msg.	Lane 8 **Test** msg.	Lane 9 **Standard** msg.	Lane 9 **Test** msg.
Mean Speed m.p.h.	**65.64**	**66.53**	**64.14**	**63.09**
Which lane speed greater?		>	>	
Normal pop. @ α=0.05? (Y/N)	N	N	N	N
(for α=0.05) Z > Zcrit (two-tail) ? [P value]	N [p=0.25]		N [p=0.25]	

Table 5: Results for station 6 night

While the lane-by-lane, two sample z-testing did not show any significant effects from the change in CMS configuration, there is still an interesting observation to be made from the tables. While there was only one significant lane pair (and that quite likely by chance), the *direction* of the change showed a pattern.

In tables 1 and 3 (daytime), the **standard** message configuration consistently had a higher mean speed, regardless of lane. In tables 2 and 4 (nighttime) the **test** message configuration consistently had a higher mean speed except for lane 9, which was the same in both cases.

While it is tempting to perform a *two-factor Analysis of Variance* (ANOVA)[10], it is not clear that there is much insight to be gained from it. Any overall effect that was missed by the lane-by-lane analysis would necessarily be weak, and night and day messages would seem to have the opposite effects.

Conclusions

A lane-by-lane analysis showed no significant differences in mean vehicle speeds between the standard CMS configuration and the tested CMS configuration when the permanent CMS showed the same "innocuous" message under both conditions.

A *possible* overall, weak effect may be present but the time of day would be a confounding factor (see table 5).

	Station 5 [std.-test]	**Station 6 [std.-test]**
Day	0.85 m.p.h.	0.72 m.p.h.
Night	-0.01 m.p.h.	-0.41 m.p.h.

Table 6: Differential between mean speeds under the standard configuration and the test configuration, averaged over lanes.

In any event, unless such small differences, as are shown in table 5, are likely to be perceived as important by a traffic engineer, it is hard to escape the conclusion that <u>*there is no significant disruption to traffic flow when a CMS displays a message in an optimal presentation*</u> (where "optimal" is used in the sense of VDL's previous in-lab results).

In the highly unlikely case that such small differentials as shown in table 5 are important to a traffic engineer, then a future research program could err on the side of caution by testing other messages/configurations at night as a first step. This is assuming that other test configurations would likely have results similar to table 5; in that case a night test of "minimal traffic disruption" for other configurations would likely be successful.

Obviously, if the daytime numbers are problematic then further tests specifically for daytime could be designed based on the above results. We did not calculate uncertainties on table 5 because such small differences seemed unlikely to be important. If a traffic engineer deems otherwise, then given a cutoff value and an uncertainty from the engineer (e.g. 0.8 ± 0.2 m.p.h. as the maximum permissible "standard minus test" difference) VDL could estimate the number of vehicle speed measurements needed in a future test in order to provide sufficient resolution to see definitively if traffic conditions under test meet the given cutoff. A Monte Carlo simulation based on the speed distributions found here could best do such estimation. This approach is quite involved however and would only be worth the effort if an expert deemed the small speed differences in table 5 problematic.

[10] The factors being the *lane*, which will clearly have an effect, and the *message configuration*. Failure to take into account the lane number would give incorrect results.

As was mentioned earlier, VDL’s choice of an “innocuous” message was made so as to have no harmful effect if the optimized message did have an unexpected disruption on traffic. Having seen no significant evidence of such a disruption, the next step for someone wishing to carry the work further would be to test a “real” message. This would not be as straightforward as our work here since a “real” message (e.g. “*ROADWORK AHEAD* “) would have to be a) true and b) coordinated with the roadway speed sensors.

Acknowledgements

The staff at VDL would like to thank Charles Price and Alan Chow of *Caltrans* district 4 for both approving this work and having the CMS operations staff successfully implement the required sign changes. We would also like to thank J.D. Margulici and Saneesh Apte of the California Center for Innovative Transportation who provided us with the requisite speed data and much other useful information about the Berkeley Highway Laboratory.