EVALUATING COMPREHENSION OF TEMPORARY TRAFFIC CONTROL

A Dissertation

Presented to

The Academic Faculty

by

Aaron Todd Greenwood

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in Civil Engineering in the

School of Civil and Environmental Engineering

Georgia Institute of Technology

December 2015

EVALUATING COMPREHENSION OF TEMPORARY TRAFFIC CONTROL

Approved by:

Dr. Michael P. Hunter, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Kari Watkins
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael O. Rodgers
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Gregory Corso
Department of Psychology
Morehead State University
School of Psychology
Georgia Institute of Technology

Dr. Randall Guensler
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: November 13, 2015

To my family.

ACKNOWLEDGEMENTS

This work has been part of a larger project for the Georgia Department of Transportation, and without the help of my project team there is no way I could have produced this work. The team, Dr. Michael Hunter, Dr. Michael Rodgers, Dr. Gregory Corso, Dr. Yanzhi (Ann) Xu, and Chenhao Liu worked with me to develop the experimental procedure and select alternatives for testing.

Dr. Corso arranged for the participants through the School of Psychology and provided key insight, reframing a problem with a huge psychological component for a group of engineers and scientists.

I could not have worked through many of the problems in this dissertation without the support and insight of many of my coworkers. Dwayne Henclewood, Wonho Suh, Popa Pratyaksa, Laura Schmitt Forinash, Lance Ballard, Lakshmi Peesapati, Tom Wall, Alice Grossman, Adnan Sheikh, Stefanie Brodie, Atiyya Shaw, Jong In Bae, Ross Wang, James Anderson, and all the others slugged through some of these problems with me.

On a personal note, I need to thank my board game group, my trivia team, and my dinner club for providing routine away from the lab. And I can't thank my committee enough for their invaluable help with this dissertation. Dr. Hunter, Dr. Rodgers, Dr. Guensler, Dr. Watkins, and Dr. Corso have worked with me through my unreasonably tight time constraints, and for that I am supremely grateful.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1148903.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS AND ABBREVIATIONS

β	Beta, Item Difficulty Parameter
θ	Theta, Person Ability Parameter
μ	Mu, mean
σ	Sigma, variance
χ^2	Chi-squared distribution
2PL	2-Parameter Logistic Model
3PL	3-Parameter Logistic Model
AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials
D10A	Aligned drums spaced 10 feet apart
D10M	Misaligned drums spaced 10 feet +/- 2 feet apart
D40A	Aligned drums spaced 40 feet apart
D40M	Misaligned drums spaced 40 feet +/- 2 feet apart
GEMS	Generic Error Modeling System
ICC	Item Characteristic Curve

ID	Identification
IRT	Item Response Theory
ISO	International Standards Organization
LCD	Linear Channelizing Device
LLTM	Linear Logistic Test Model
MLE	Maximum Likelihood Estimation
MUTCD	Manual on Uniform Traffic Control Devices
NA	Not Applicable, a placeholder for missing data
NHTSA	National Highway Traffic Safety Administration
PCB	Portable Concrete Barrier
Pr(x)	Probability of x
WZ	Work Zone

SUMMARY

There are over 5 million reported motor vehicle collisions annually in the United States, and while crash rates and fatality rates have declined in the past decades, rates in work zones are disproportionately high. There are strict standards for evaluating the crashworthiness of temporary traffic control devices, but not for evaluating drivers' comprehension of existing or novel device deployments. This dissertation presents a series of three experiments evaluating driver comprehension for existing and novel traffic control devices conducted in a work zone setting. This evaluation is further expanded by decomposing the task of comprehending traffic control into the three subtasks of detection, localization, and identification. Methods are proposed for conducting a computer-based experiment with still image stimuli to measure participant performance at each of these subtasks. Next, procedures for categorizing localization responses and accounting for variation in participants physical responses are explored. Lastly, an application of Item Response Theory toward the evaluation and comparison of participant comprehension is demonstrated. It is hoped that these methods and procedures can be used by future researchers and experimenters to compare novel temporary traffic control devices and systems to inform future design.

CHAPTER 1

INTRODUCTION

When a driver operating a vehicle to the best of their abilities in a work zone crashes because they did not understand the temporary traffic control, the cause may be written off as driver error. In reality, there is often a disconnect between a traffic control designer's message and a driver's reading of that message, with no clear way to measure comprehension. While there are extensive tests for determining how much physical harm a crash can cause, there are simply not tests for determining the potential harm caused by the confusion a system of traffic control devices may generate. To ensure that a car is safe, it is tested; NHTSA certifies every model of automobile by crashing it before it is ever sold and measuring the effect on simulated people, the dummies. Similarly, there are standards for testing the crashworthiness of the devices we place on the road, and the Federal government requires each device be certified through physical testing (Ré & Carlson, 2012; Ross, Sicking, & Zimmer, 1993). The objective of this dissertation is to explore issues related to the testing of human perception of systems of traffic control devices and to develop guidance for such testing through the use of an extensive work zone related case study. This guidance accounts for visual search processes, physical response error by test participants, strategies for categorizing responses, and methods for modeling and interpreting test participants' performance. That is, understanding how users interpret and respond to a device rather than the crash worthiness of the device.

Understanding these issues is critical. The driving task is a complicated process requiring divided attention; drivers must simultaneously control a vehicle, navigate a route, and communicate with other drivers, all while avoiding potentially lethal collisions. A work zone case study is utilized to discuss and demonstrate these issues as a work zone represents one of the most common yet challenging locations to design effective traffic control. Work zones add an extra level of complexity to the driving task by creating a temporary second layer of traffic control devices which supersede permanent striping and signage on routes that are new to some drivers and driven daily by others. It is no surprise, then, that decades of research shows work zones are disproportionately dangerous environments for drivers (Graham, Paulsen, & Glennon, 1978; Khattak, Khattak, & Council, 2002; Rouphail, Yang, & Fazio, 1988; G. Ullman, Finley, & Bryden, 2008). To better design work zone and non-work zone traffic control systems, practitioners need solid guidance on methods for ensuring that drivers comprehend their traffic control plans.

Safety is Improved by Reducing Errors

A significant body of literature focuses on improving safety by focusing on crashes. The recently published Highway Safety Manual illustrates crash-centric safety analysis with its use of Crash Modification Factors (Transportation Research Board, 2010). CMFs, numbers indicating the change in odds of a crash if safety measures are implemented, are useful for practitioners weighing the implementation costs of existing devices and programs, but are not helpful for evaluating new devices and novel configurations. Crashes are rare events, and untested devices do not have the in-field use history needed to evaluate their effectiveness at reducing crashes.

While the ultimate measure of safety is the frequency and severity of crashes, often in terms of crash severity, the actions and events prior to a crash must be addressed to improve safety. A successful safety improvement program encompasses the reduction of events that may lead to a crash (i.e. driver error, lack of clear message for traffic control device, etc.) While only in rare instances will these events lead to a crash, a measure of their occurrence may help act as a hazard indicator (Chin & Quek, 1997). Unlike the airline or nuclear energy industries, there is no "near-miss" reporting system for road transportation. However, with advances in remote eye tracking, survey design, in-vehicle sensors, and simulation, technologies allowing for the capture of smaller correctable errors are becoming readily available. The psychological and ergonomics theories to describe and approach these events have been refined for decades. For instance, Rasmussen's (1983) Skills, Rules, Knowledge (SRK) framework is well suited for driving. The framework states that humans use automatic, learned skills in normal operations, defined rules in response to previously encountered problems, and knowledge about a system to deal with previously unseen problems. Rasmussen's taxonomy has three levels of task responses: skill-based, rule-based, and knowledge-based. Changes in traffic control devices in a roadway constitute a sufficient change to the operating environment that they move drivers from skill-based work (typical in the common physical aspects of driving, such as lane-keeping or making a turn) into rule-based decision making or knowledge-based problem solving. Drivers encountering simple traffic control can maintain their routes in response to rule-based decisions (e.g. lane shifts or lane closures). Drivers performing tasks in more complex environments must use knowledge-based problem solving to navigate a unique environment based on schema from past experience (e.g. taking an exit or turning into an unmarked driveway). Easily understood traffic control is critical on all three levels. Effective traffic control can quickly cue a driver to use a skill. Effective traffic control can also help recall a clear rule to follow in a more difficult situation. And in very complicated work zones, effective traffic control can provide sufficient information to guide the driver to the correct decision.

Better traffic control is not a panacea for driver safety, but it can lead to a reduction in specific types of errors. In Reason's (1990) taxonomy, errors can be classified into four types. *Slips* are errors where a plan of action was appropriate but physically executed incorrectly. *Lapses* are errors where a plan of action was appropriate, but a step or critical piece of information was forgotten. *Mistakes* are errors where a plan of action executed correctly but was unknowingly inappropriate. *Violations* are errors where a plan of action was intentionally inappropriate and executed correctly. Better traffic control can address lapses and mistakes by making important information easier to maintain in working memory to avoid lapses and clearer to avoid the improper assessment of the operating environment that leads to mistakes. While not directly addressed in this study, violations may also potentially be reduced by traffic control which presents a more forceful message, reducing the willingness of a driver to intentionally commit a violation.

Better Design Principles and Testing Reduce Errors

The first edition of the Manual of Uniform Traffic Control Devices was published in 1935, when the United States had fewer than 130 million citizens and only 20 states required driver license examinations. At the time, local municipalities were erecting signs using their own standards, and engineer studies were not widely required for traffic control implementation. It is easy to see how standardization was of utmost importance. However, AASHO's 1927 *Manual and Specifications for the Manufacture, Display, and Erection of U.S. Standard Road Markers and Signs*, which predated the MUTCD, selected colors and shapes before modern testing methods and understanding of humans' visual perception were well understood and widespread. (AASHO - the American Association of State Highway Officials is the predecessor of AASHTO - the American Association of State Highway Transportation Officials.) The priority at the time was not device design, but rather standardizing a mix of inconsistent signs erected by local municipalities.

While the standardization of signage has been successful, new challenges exist today. The US has over 350 million citizens, highly standardized traffic control, and 150 million drivers with driver licenses which all required knowledge tests. However, the comprehension rate for several signs is less than 50% (Stokes, Rys, & Russell, 1996). It is likely true that there are network effects for some traffic control devices, and uniformity has helped make an otherwise meaningless symbol gain uniform recognition; `devices which may not be optimally designed are well known because they are common, much the same way the QWERTY keyboard has persisted despite other designs being faster for beginning typists. The STOP sign is a good example; the STOP sign is almost universally recognizable, even though nothing about an octagon conveys the idea of "Stop," and red does not consistently mean danger or warning across cultures. While some such signs may benefit from uniformity, it is not clear that the network effects from legacy uniform traffic control devices uniformly outweigh the need to develop signs and striping with more easily understood meanings. Traffic engineers present traffic control devices in an effort to help drivers assess and safely traverse the environment. If those devices do not convey accurately interpreted messages, they are not working to prevent driver error and may in fact be contributing to it. Using modern usability assessment techniques (i.e. studies of how actual users respond to devices) and technologies, this research will focus on understanding key issues for the testing and assessment of design principles for practitioners developing traffic control, with a work zone case study, from a human factors standpoint (ISO, 2010; Johnson, 2010; Nielson, 1993). While others have developed general principles for traffic control design using uniform devices (Campbell et al., 2012; Lunenfeld & Alexander, 1990), focus on design principles specifically for work zones has been limited or overly broad (e.g. MUTCD 6B.01). Further, inherently complicated work zones may not be best explained to drivers using existing, uniform devices.

Even with better design principles, practitioners must design traffic control for each site uniquely. In the field of computer interface design, Carroll (1989) described the Infinite Detail requirement: "bridging representations must, in principle, incorporate *all* the details of the situation of use." If principles are not specific enough, designers will "design-by-emulation," a concept already observed in traffic control as temporary traffic control plans are designed by modifying previously used plans. In such cases, designers who are presented with a universe of infinite use cases will emulate previous designs without an understanding of how they will be interpreted by new users. In computer science, "design-by-emulation" has been overcome by user-centered design and iterative testing. User-centered design calls for construction of a test case using general design principles, and then iterative refinement by testing use cases with typical users. As 3D site modelling becomes less resource intensive, practitioners will be able to quickly generate test cases and employ rapid testing for understanding the comprehension of traffic control traffic control in a specific environment. This research develops the testing method practitioners and researchers can use to evaluate users' understanding of traffic control devices and systems.

Methods for Administering Spatial Traffic Control Tests and Processes for Analyzing their Results Are Contributions

This work details a method for testing traffic control device systems for driver comprehension. In order to give the reader the necessary information to consider this method, Chapter 2 will provide background. Chapter 2 will give details on traffic safety, work zone safety, design for perception, and visual search. The third chapter will give an overview of a work zone case study which explored channelizing devices. Chapter 3 will outline a of traffic control at work zones consisting of three experiments. Chapter 3 will include an evaluation of various existing device configurations, a developed novel treatment, and work zone conditions. Chapters 4 through 7 will then discuss and address some of the issues realized in the work zone case study.

Chapter 4 begins to re-evaluate of the underlying assumptions of the case study, showing how the factors contributing to response complexities are not discernable in the case study. Using signal detection theory for deeper analysis, these data suggest that the task of comprehending traffic control devices is more than just a simple yes/no detection task, but a task requiring detection, localization, and identification.

Chapter 5 will explore how to design a new experiment to capture these three parts of perception: detection, localization, and identification. The suggested perception experiment separates out each of these three parts as a subtask for individual analysis.

Chapter 5 then further explores the categorization of participant responses. Each step of the detect, localize, and identify task deals with nominal response categories. However, binning of localization responses into spatially defined categories complicates the analysis of participants'' responses; Chapter 5 discusses ways to ensure an unbiased system when defining categories for localization responses.

Chapter 6 considers ways for taking the proposed method in Chapter 5 and reducing the data noise caused by response errors, as opposed to comprehension errors. Chapter 6 will investigate strategies others have used to account for physical error in response systems. Methods for mathematically accounting for the physical error of participants through measuring their response distributions will also be proposed.

Chapter 7 describes a model to give analytical and predictive power to the collected and categorized data. This chapter overviews how to use a form of logistic regression, the Rasch model, to assess the difficulty participants will have detecting, localizing, and identifying traffic control devices. Also in the chapter, the Linear Logistic Test Model is applied to account for practice and fatigue effects over the course of an experiment. Chapter 7 further discusses ways of using these models to predict the likelihood of a driver comprehending a device once it is implemented on the road.

Finally Chapter 8 summarizes the work completed and connects each chapter to show how they work in concert to develop a robust methodology for testing and analyzing traffic control. Chapter 8 explicitly shows the contributions of this work, which are an analysis of prior traffic control testing, an experimental design with categorization for testing traffic control, a method for accounting for physical error in participant responses, and the application of the Linear Logistic Test Model to data analysis.

CHAPTER 2

BACKGROUND

The results of a comprehension test are not useful in and of themselves; they are only practical in the context of improving design of systems. In those contexts, it is important to understand the problem of safety which traffic engineers are trying to solve. The first section of the background will focus on safety literature. The road safety literature largely focuses on car crashes, but crashes do not just happen; crashes are the result of human action. Thus, after detailing the safety problem on public roads (especially in work zones), the next section will discuss the concept of human error. While in the vernacular, “human error” implies fault with the human, a human’s response and behavior is strongly influenced by its environment. This section provides an overview for how design principles can be used to reduce the chance of human error, or maybe more importantly, increase the chance of the desired (i.e. safe) response. To develop temporary traffic control design principles, it is important to understand the perceptual processes underlying visual search.

Safety

Road safety is a serious problem on both a national and global level. In 2011, there were approximately 5.3 million crashes in the United States and 32,367 fatalities (NHTSA, 2013). The United Nations estimates that nearly 1.3 million people die worldwide in traffic crashes (United Nations, 2011). While research shows that drivers operate their vehicles in a manner they perceive to be safe (Theeuwes & Godthelp, 1995), all elements of the roadway including the traffic control system impact driver safety. The following focuses on freeways, ramp junctions, and work zones as this is the case study scenario. However, the observations for these locations offer important insights into understanding issues surrounding safety and help highlight the need for effective traffic control, particularly in complex situations.

Safety on Freeways

While general road safety is important, the unique design of freeways with their full control of access, high traffic volumes, and high speeds necessitates special safety consideration. Wang, Cao, Deng, Lu, and Zhang (2011) evaluated truck-related crashes at exit ramps in an attempt to develop a model for determining safety at diverges. Wang et al. found that collisions increased as AADT increased, both for trucks and overall. Wang et al. also found a significant improvement in safety from an increase in the length of deceleration lanes and from using ramps without lane drops or with option lanes (in the case of 2-lane exits). Lastly, they saw a significant improvement in safety with an increase in shoulder width.

Chen, Zhou, Zhao, and Hsu (2011) investigated safety at left side exit ramps in Florida, and found that there was an elevated crash risk for these types of exits. While Chen et al. did not explore why left exits were correlated with an elevated crash risk, the potential exists that left hand exits could also present increased hazards in work zones and require improved traffic control device systems. Lu et al. (2009) evaluated diverges in Florida, investigating how ramp type and ramp characteristics influenced safety. They found that exits without lane drops had the lowest crash rates and that free flow loop ramps significantly increased crash rate. There is value in knowing that different types of ramps can influence crash risk, and ramps should be designed knowing that underlying characteristics of the ramps themselves could contribute to collisions. Khorashadi (1998) found that 15% of incidents in the State of California between 1992 and 1994 occurred on ramps. Analyzing those incidents, he found that ramp AADT, freeway AADT, whether the ramp was urban/rural, the type (on/off), the configuration, the length of the speed change lanes, and the ramp length to be significant. Of note were that off-ramps had more collisions and more severe (injury and fatality) incidents than on-ramps.

McCartt, Northrup, and Retting (2004) examined 1,150 crashes at ramps and found that about half of crashes happened when drivers were exiting the freeway; however, they found that congestion and speed were contributing factors to all crash types. Speed was mostly a factor in run-off-the-road crashes and congestion was a strong factor in rear-end collisions. Thus, from this study and the previous, it is clear that many factors may contribute to crashes. Traffic control device system guidance that can be applied to specific sites in an effort to improve driver responses and reduce crashes is critical.

Safety in Work Zones

In the roadway system, construction zones represent some of the most visually intense and complex environments, requiring drivers to deviate from usual driving behavior to deal with new traffic patterns and devices to indicate an elevated level of risk. Khattak, Khattak, and Council (2002) estimate that there are approximately 24,000 non-injury crashes and 52,000 property damage-only crashes in work zones annually. The Fatality Analysis Reporting System statistics for 2010 show that there were at least 576 fatalities (2% of total reported facilities, including workers killed by traffic) in work zones in 2010 alone (National Work Zone Safety Clearinghouse, 2012). Several studies have shown specific dangers of work zones to drivers.

Daniel, Dixon, and Jared (2000) found that there was an elevated risk of fatal incidents in Georgia work zones. Specifically, they found that even though work zones make up a relatively small amount of overall roadway mileage, they account for more freeway fatal freeway crashes than in areas without road work. The types of collisions where fatal crashes occur are also telling: nearly half of all crashes were single-vehicle collisions, and 12.1% of collisions were rear-end collisions, compared with 56% single vehicle and 5% rear-end collisions in non-work zone fatal crashes. Most of the crashes took place in construction zones that were idle and the type of construction was typically resurfacing or roadway widening. These conditions suggest that relatively common work zones that may be perceived as being lower risk still lead to an unacceptable number of fatalities. These areas, typically delineated by drums and often having temporary diverges through changes in the pavement surface, could benefit from improved methods of work zone delineation.

Work zone intrusions are especially worrisome when considering diverges as the ultimate goal of an exiting driver at a diverge is to depart from the current roadway. These types of error will prove to be especially relevant to the work zone case in Chapter 3. The decision to diverge from the travelled way is, in effect, the decision to intrude upon the work zone *in the proper location*. Bryden, Andrew, and Foruniewicz (2000) evaluated 290 intrusions between 1993 and 1998 in New York State. Of these observed intrusions, 10 occurred where drivers were trying to cross the work zone to enter or exit “a driveway or other roadside location.” While this type of incident is rare, the study demonstrates that it is an issue in work zones and that there is room for improvement in delineation methods. The same methods and devices for improving channelizing device comprehension at diverge locations could also be used in tangent sections to reduce other unintentional intrusions.

Some guidance exists on work zone design (Roadway Safety Consortium, 2010), but it is mostly concerned with maintenance. State level plans exist to help with temporary traffic control plans, but these are largely regulatory, rather than providing general guidance. Further work has compared nighttime and daytime work zone operations, but like much research it was focused on crash reports, not driver error (G. Ullman et al., 2008).

Traffic Control Devices

Design of Traffic Control Devices

Pain, McGee, and Knapp (1981) explain a problem with temporary traffic control design: "Devices described in Part VI of the *Manual on Uniform Traffic Control Devices* (MUTCD), have developed simply as an evolvement from other devices, rather than as a result of scientific testing as to what best stimulates driver awareness of work zone situations." For instance, the nearly ubiquitous channelizing drum's patent was not filed until 1976 (Florsheim & Kulp, 1978). The plastic drum was deemed a safer alternative than the filled metal 55-gallon drums previously in use. Little research has been found prior to this patent exploring how drivers interpreted these devices. Some research has been found from after the patent filing, such as a discussion of their visibility characteristics (Pain et al., 1981). However, drums (and other devices) are now used extensively and were integrated into design standards without any testing or certification of whether drivers understood drums' intended meaning.

Evaluation of Traffic Control Device

Crashworthiness

In the United States, the Federal Highway Administration mandates that all temporary traffic control devices comply with the crashworthiness standards outlined in NCHRP 350 (Ross et al., 1993). These standards require that manufacturers certify their devices prior to use on the National Highway System.

Comprehension

Modern research into comprehension of channelizing devices has largely focused on existing systems. Several studies have looked at how channelizing devices in work zones affect driver performance, both at exit ramps and through work zones in general. Dudek, Finley, and Ullman (2001), for instance, investigated how sequential flashing lights placed on top of drums aided driver comprehension of a lane closure. They evaluated driver understanding through a traditional survey after participants drove through the scene, though others have used simple computer surveys to gauge comprehension. Ullman, Trout, and Ullman (2012) for instance, showed drivers still images of mobile painting operations to evaluate comprehension of signs. They used a questionnaire to evaluate the use of "Your Speed/My Speed" signs on the back of slow moving trucks, and they found that drivers were confused by the two sets of numbers.

User-Centered Design

The use of user-centered design techniques to ensure comprehension in human-computer systems is well documented and established in the field of Human-Computer

Interaction. Carroll (1990) was one of the first to identify issues with requirement-driven design rather than design guided by general principles. Nielson's (1993) foundational work evaluated usability engineering in the context of both requirement gathering (collecting information about the context of both users and the operating environment) and iterative design. Following ISO 13407, ISO 9241-210 (ISO, 2010) outlines the process for user-centered design, focusing on user requirements gathering, iterative design, and user involvement in the design process. Rodgers, Sharp, and Preece (2011) and others have modern textbooks and guidebooks on the methodologies and processes refined in the field of user-centered design.

For designers to shift away from design using past principles and employ user-centered design, current user-centered design methods require an iterative testing method. The method presented in this work can be deployed quickly enough that several iterations can be performed early in the design process for new traffic control systems. It is not enough, though, to have iterative tests unless they can measure the right performance metric. A common thread between most of the previously mentioned studies of safety is that they evaluate crash data. However, there are weaknesses in the use of crash data. From a practical perspective, crashes are often not reported (M. Davis & Co., 2015), which can bias the data. Using crash data does not necessarily show which designs are "good" or "better", but rather which designs reach a minimum threshold; nothing is known about the how drivers perceive a device, only that it is present nearby an arbitrarily defined 'acceptable' number of crashes. Also, because crashes are rare events using crash data to observe a safety problem may mask a design problem, especially where the risk rate is high but the traffic exposure is low. It is therefore useful to supplement or precede crashes with errors as the metric for a test of comprehension. The next section will investigate human errors and how they relate to the traffic system.

Human Error

In order to investigate driver error, it is critical to have a solid understanding and model of what human error is and its underlying causes. There are several complementary theories on this subject. Senders and Morray (1991) describe error as something which was "not intended by the actor, not desired by a set of rules or an external observer, or that led the task or system outside its acceptable limits." Rasmussen's (Rasmussen, 1983) presented a Skill-Rule-Knowledge model, which suggests that humans take shortcuts to perform actions based on their level of comfort with "automatically" performing an action or responding to a situation.

Reason's Generic Error Modeling System further extends Rasmussen's SRK model into a method for discovering at which point in the decision-action process an error occurs (Reason, 1990). This model applies to traffic control specifically within Rule-Based Mistakes where drivers incorrectly "Consider local state information" and Knowledge-Based Mistakes where drivers are unable to form a working mental model of the situation or are unable to relate abstract concepts to specific traffic control. Figure 1 demonstrates Rasmussen's GEMS model. The "Consider local state information" stage is the second step of the second box representing Rule-Based problem solving.

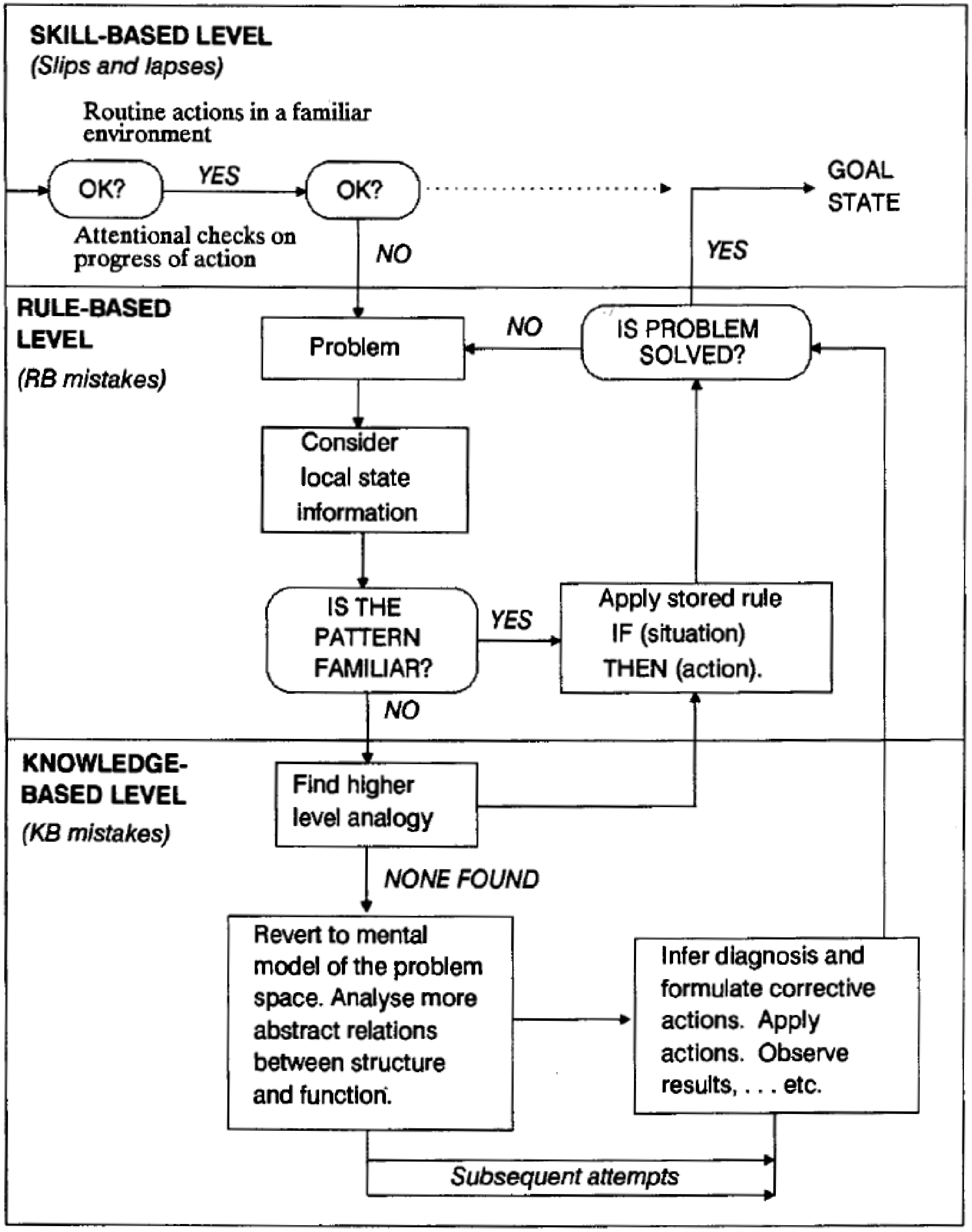

Figure 1: The GEMS model shows the links where errors can occur. Reproduced from Reason, 1990

Building on Rasmussen's work, Reason (1990) developed four classifications for errors: *slips, lapses, violations,* or *mistakes*, subdividing mistakes further into rule-based mistakes and knowledge-based mistakes.

A *slip* is an error of execution, where the actor makes a physical error that leads to a problem. An example of a slip in driving would be where a driver intends to shift from fifth into fourth gear, but instead shifts into reverse--the plan of action was correct, but the error was in the physical execution. A *lapse* is an error of memory, where an actor forgets a key piece of information about the task such as a needed number or what step they are in of a process. An example of a lapse in driving would be where a driver stops at a fuel station, turns the car off, opens the gas tank, but has not realized that they have forgotten to put the car in Park. A *violation* is an error where the actor, accurately knowing the system state and the rules, chooses to act contrary to the system rules. A *mistake* is an error caused by a perception problem in observation of the system state which leads to the selection of improper action. These can further be divided into rule-based mistakes (where an improper rule is selected because of a perception) and knowledge-based mistakes (where a person acts improperly because they do not have the knowledge needed to select a rule or do not know of a rule to follow). An example of a rule-based mistake would be taking a left HOV freeway exit when intending to stay on the freeway; the driver chooses the rule that keeping left stays on the main road, and they act in a manner that would be correct if their observation was not wrong. An example of a knowledge-based mistake would be reacting to a skid on ice without proper training; a driver might turn out of a skid rather than into a skid because their mental model of how skidding works does not align with the physics.

Designing for Error

Each of the errors listed as examples in the previous section lead back to the operator; ultimately the action that can be classified as an 'error' is initiated entirely by a single actor. However, designers can work both to reduce the likelihood of an error and to reduce the negative consequences of an error. For example, for the critical slip error of shifting into reverse while moving forward, most H-pattern shifters prevent the 5th to Reverse movement. For the lapse error of forgetting to put a car in Park, a shift interlock device reminds the driver of their lapse by keeping the key stuck in the ignition. For the rule-based mistake of taking a left exit, designers use redundant signage and lights to indicate that a different rule must be selected, and crash barriers to reduce harm if such an error occurs. For the knowledge-based mistake of skidding on ice, electronic stability control and antilock braking can reduce the impact of such an error.

The analysis in this dissertation deals with reducing mistakes caused by improperly perceiving the traffic control system. The testing method developed here is rooted in knowledge about the human visual system, and this knowledge will inform the experimental design herein. The following information gives an overview of the concepts of visual search and design principles needed to understand how a perception test can indicate that a traffic control system would reduce the likelihood of mistakes caused by not comprehending the visual environment.

Visual Search

Understanding how drivers search the scene has been a priority for decades in road safety research. Mourant and Rockwell (1972) used primitive eye-tracking to determine differences between novice and expert drivers in the early 1970s and late

1960s. Shinar (Shinar, 2008) furthered this study by looking at the locations where drivers focus using modern techniques. Others (Crundall et al., 2012) noted that novice drivers and experienced drivers fixate differently on potential and actual hazards. They found that novice drivers fixate longer on actual hazards than experienced drivers, but experienced drivers better identified hazard precursors, hence modern technology can help differentiate between scanning and study of a scene. This is consistent with findings that younger drivers tend to fixate on nearer points for longer times (Mourant & Rockwell, 1972).

Outside the field of road transportation, there are several interesting findings regarding visual search. Chang, Kinshuk, Chen, and Yu (2012) found that information presentation matters in recall tasks, occasionally more than information density. Specifically, when presented with patterns to remember simultaneously rather than sequentially, participants had better recall. This runs counter to the concept of spreading in traffic control, instead suggesting that some levels of information, simultaneous presentation may be more memorable.

While not directly related to visual search, another important consideration for this research is Trick, Brandigampola, and Enns' (2012) finding that images can affect drivers' emotions, which in turn affect both their steering and hazard response time. Just by showing different images, an individual's emotion's valence (the attractiveness or desirability of an emotion) impacted the time to being braking in response to a lead vehicle deceleration event. This further support's Carroll's (1990) observation that designers cannot accommodate the infinite level of detail in each situation, and further justifies the use of user centered design methods for work zone testing.

Principles of Grouping

In work zones, it is often physically difficult or very costly to use a single object to indicate the perimeter of a work zone. Since it would be difficult to place a solid fence up in an active travel way, most jurisdictions depend on separate channelizing devices to "simulate" a single wall of objects in the mind of drivers. These point devices, e.g. orange and white retroreflective channelizing drums, depend on the Gestalt principles of grouping for drivers to take the individual drums, panels, or other channelizing devices and mentally associate them with a group. Johnson (2010) explains the six non-moving Gestalt principles of proximity, similarity, continuity, closure, symmetry, and figure/ground, as demonstrated in Figure 2.

Individuals use proximity to interpret that separate objects are grouped because of how close they are to each other. Similarity indicates that separate objects are grouped because they appear to be in some way the same. Continuity indicates grouping through a linear pattern common to all objects in the group. Closure makes overlapping objects appear to be grouped together and also allows separate objects appear to construct a single object. Symmetry helps group wireframe objects that overlap, and figure/ground helps individuals group objects together based on a common background.

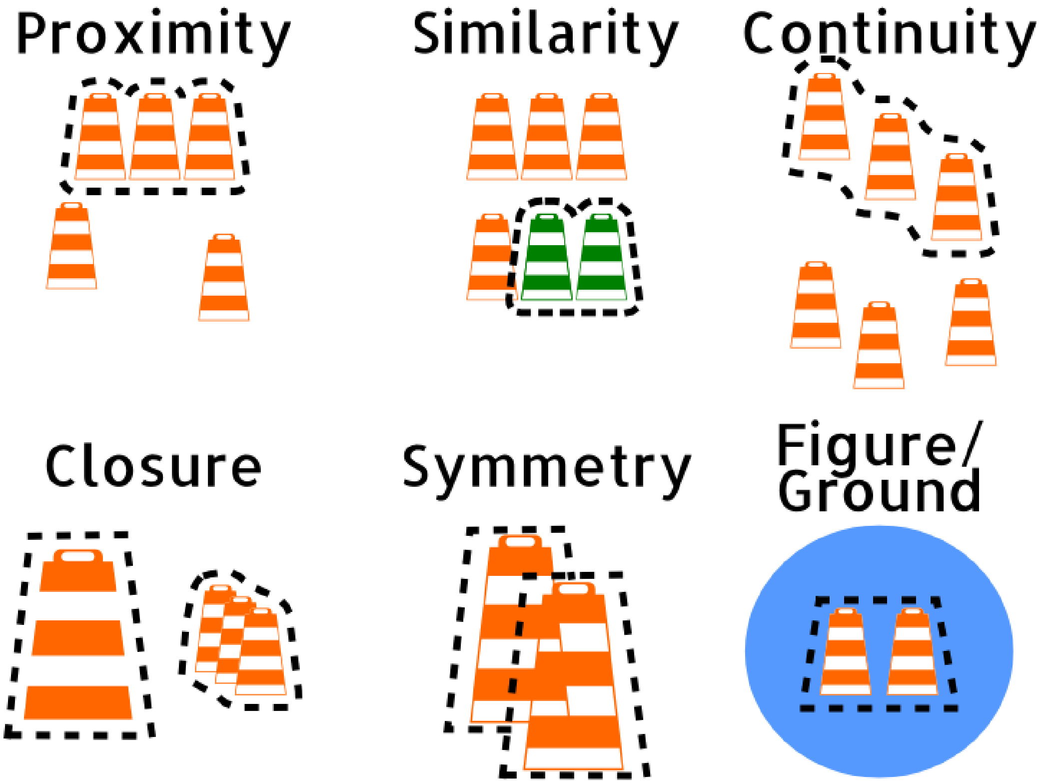

Figure 2. Gestalt Principles of Grouping (Groups Shown with Dotted Lines)

Work zone traffic control designers implicitly employ these grouping principles to maintain the appearance of a single closed area through point-based channelizing devices. Several problems arise with this system, however. Different states have different standards on how close drums should be spaced, illustrating how there is no consensus on an appropriate level of proximity. Continuity can be degraded due to variability in device placement or natural shifting from wind or traffic. Drums or cones appear closed when at a distance because they overlap in a driver's frame of view, but as the driver approaches these devices the closure is broken, shifting the burden of grouping to the other three Gestalt principles. Unique to diverges, similarity creates a problem because there are two appropriate and safe traveled ways (the main road and the ramp) that are both indicated with the same devices, making it difficult to identify that there are actually two groups of channelizing devices.

The effect these principles have on perception can significantly affect how an individual responds to stimuli in the world. In a series of five experiments, Coren and Girgus (1980) found that when some objects were grouped through Gestalt principles, the distances between objects in the group was perceived to be smaller than the distance between objects outside the groupings, even, though the distances were identical. This could have profound impacts on work zone design if perceived distances vary from actual distances in a way that negatively impacts safety. O'Shaughnessy and Kayson (1982) further investigated these concepts by including the time an individual is shown the tested scene. O'Shaughnessy and Kayson found that both proximity and time had an effect on how individuals accurately assessed distances, with improved accuracy with shorter times and improved accuracy with smaller distances. O'Shaughnessy and Kayson did not find the same effects with similarity and closure, however, indicating that while the Gestalt principles are a good heuristic, they cannot be applied as "laws" and testing is still necessary to predict perceptual performance.

Feature Integration Theory

An important piece of understanding visual search is Feature Integration Theory. Feature Integration Theory suggests that individuals pre-attentively identify potential targets based on features, and then search serially to identify those targets (A. M. Treisman & Gelade, 1980). Features are thought to be singular characteristics of an object, such as color, shape, contrast, orientation, etc. Objects defined by one feature can be searched quickly and efficiently, while objects defined by a conjunction of features require inefficient, serial search for identification. Further, when attention cannot be given to an object, ‘illusory conjunctions’ may be formed, where the subject incorrectly identifies an object by mixing up features present in the scene.

Summary

This chapter was an overview of the literature and previous work associated with temporary traffic control and evaluating temporary traffic control. The literature suggests that that there is a safety problem in not only on the roads in general, but especially in work zones. This concept of safety is generally measured as the number of crashes, even though crashes are the result of human errors. The most relevant error to misinterpreting traffic control is a mistake, where a driver misreads the environment and that incorrect assessment shapes their actions. By designing devices that account for how drivers perform visual search, designers can attempt to reduce the mistakes drivers make in work zones. However, there are many methods in the literature used to test comprehension in the literature, and research suggests that the test used can impact the results. The work zone case in the next chapter was designed in that context of attempting to measure comprehension without a standardized methodology.

CHAPTER 3

WORK ZONE CASE

The following chapter overviews a specific case that is the foundation for the work performed in this dissertation. The purpose of this case study was to investigate current designs for work zone traffic control at freeway diverges and to develop potentially novel treatments. This question was a direct result of observations by the state Department of Transportation. Using existing guidance and traffic control best practices, drivers were still making mistakes in the work zones. The work performed can be summarized into a working knowledge in the next few paragraphs.

The goal of this study was to see how different traffic control potentially impacted drivers' understanding of the exit ramp. To investigate potential conflicts at diverges in freeway work zones, a series of still images of work zones were generated and shown in rapid succession to participants on computer monitors. Participants were asked to click on the ramp if the ramp was open, and click a button labeled "EXIT CLOSED" if the ramp was closed. The X,Y coordinates of their clicks were recorded for later analysis. There were three experiments, each showing different traffic control devices. These devices were Aligned Drums 40 feet apart (D40A), Aligned Drums 10 feet apart (D10A), Misaligned Drums 40 feet apart (D40M), Misaligned Drums 10 feet apart (D10M), a Portable Concrete Barrier (PCB), and a novel Linear Channelizing Device (LCD). Examples of these images and other stimuli presented throughout the experiment are available in Appendix A.

Results were analyzed in aggregate using response rates and through ANOVA, blocking for participant. These results indicated that participants were likely to correctly identify a ramp with the PCB or LCD alternative, less likely to correctly identify the ramp with the D10A and D40A alternatives, and unlikely to correctly identify the ramp with the D40M and D10M alternatives, especially at further distances from the ramp.

For more information about the case, continue through this chapter. For the technical report on the work, see Hunter, Rodgers, Corso, Xu, & Greenwood (2014).

Methodology

The goals of reducing risk to participants observing novel devices, the high cost of field research, and the need for rapid testing with existing equipment dictated that the evaluation of the effectiveness of delineation devices would be conducted under laboratory conditions. The method chosen was to test the ability of volunteer participants to identify the location and condition (i.e., open or closed) of a ramp diverge within a freeway work zone from a brief view of a still image (scene). The images were varied to reflect various work zone configurations, distances from the ramp, and in the types and spacing of delineation devices used.

Participants and Protocols

Since this study used human subjects, all experimental protocols were vetted and approved by the Human Subjects Institutional Review Board (IRB) for both the Georgia Institute of Technology and Morehead State University. Study participants were recruited from the pool of students in an introductory psychology course at either the Georgia Institute of Technology in Atlanta, GA or Morehead State University in Morehead, KY. As an elective this course includes students from departments across each campus. Participants were excluded from participation if they had not held a valid driver's license for at least two years. Demographic data about participants was not collected; without knowledge only that the participants were enrolled in a college psychology course, readers should use caution generalizing the results of this work to the broader driving population.

For this study three sequential experiments were conducted, with different participants, over the project duration. In each experiment participants were shown a variety of scenes that varied features such as roadway geometry, ramp condition, roadside vegetation, placement of construction equipment, and work zone traffic control devices and layout patterns. Each image shown to the participants contained a diverge area, either within a work zone or a base case with no work zone. Multiple alternative channelizing devices and layouts were provided (e.g. drums at different spacing, barriers, etc.) in each set of images shown to the participants. After viewing each image, participants were asked to indicate if the ramp was open or closed and, if open, to identify the location of the ramp entrance. The accuracy of the participants' responses in identifying the ramp location and condition (open/closed) were subsequently analyzed to determine the effectiveness of the particular treatment for delineation of the ramp.

Experimental Series

Each of the three study experiments built on the knowledge gained from the previous. Each experiment is described as follow:

Experiment 1 – This experiment tested existing channelizing devices and layouts in an uncluttered environment at five different distances from the ramp. This experiment evaluated the participant's (driver) perception (location and condition) of the ramp while limiting the influence of potential confounding factors not related to the channelization devices themselves (e.g. presence of construction equipment, roadside vegetation, signage, etc.).

Experiment 2 – Provided additional investigation into potential findings from the first experiment, such as the impact of minor device misalignment. In addition, this experiment added a new channelizing device (the linear channelizing device (LCD)) developed to address driver (participant) errors observed in Experiment 1.

Experiment 3 – Evaluated selected channelizing devices in environments with various roadside vegetation and construction equipment combinations, increasing scene complexity to better reflect potential field conditions.

Figure **3**, Figure 4, and Figure **5** provide example images used in Experiments 1, 2, and 3, respectively.

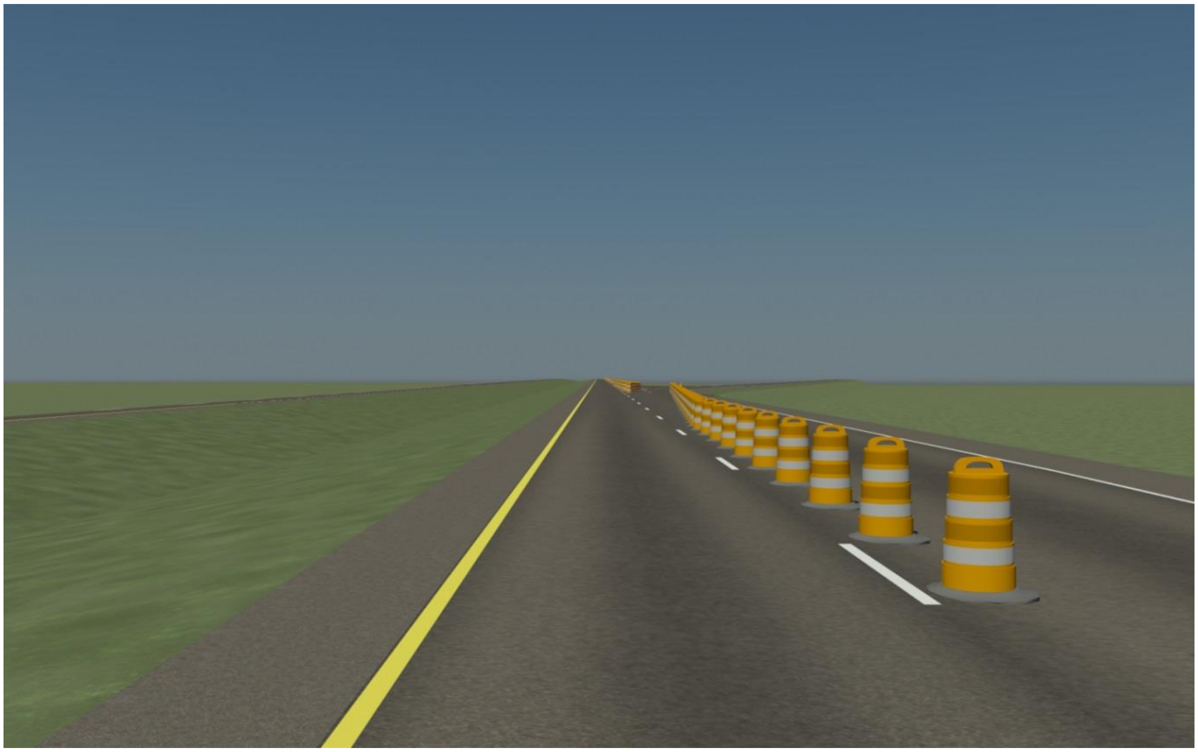

Figure 3**: Example Experiment 1 Rendering: Drums 10 ft. Apart, Ramp Open, Straight Freeway Alignment, 1 Second Travel Time to Diverge**

Figure 4: Example Experiment 2 Rendering: LCD, Ramp Open, Straight Freeway Alignment, 1 Second Travel Time to Diverge

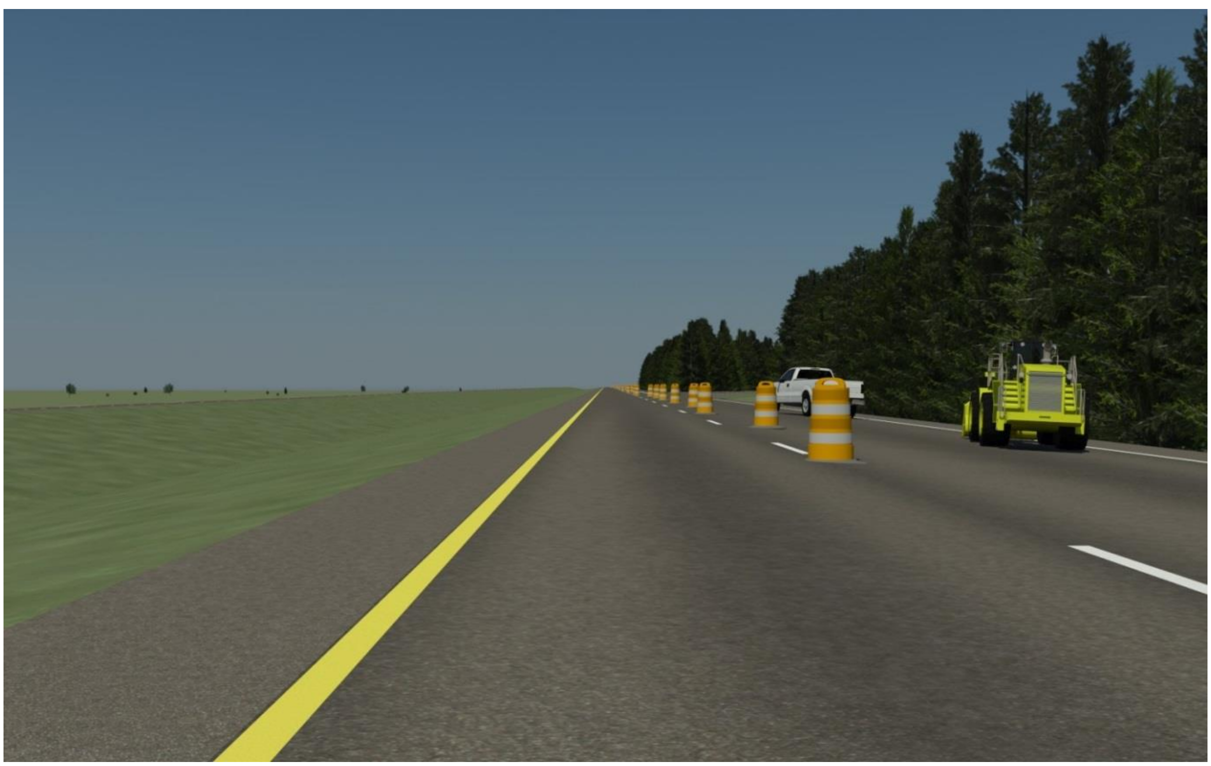

Figure 5: Example Experiment 3 Rendering: Ramp Closed, Straight Alignment, with Roadside Vegetation and Construction Equipment

Design of Linear Channelizing Device

Following the analysis of results from Experiment 1 that highlighted continuity and closure as critical aspects of channelization (these results will be discussed in detail later in this report), the research team developed a device for virtual testing that incorporated those principles while avoiding the physical size of a portable concrete barrier.

The design of the Linear Channelizing Device (LCD) was based on existing devices in the field, such as the MUTCD defined "Temporary Lane Separators" (MUTCD: 6F.72, FHWA 2009), and the engineering judgment of the project team. The base of the device has an overall trapezoidal configuration with a bottom width of 2 feet (60 cm) in contact with the pavement. The two sloping sides are each 9 inches wide and colored orange. The top surface is colored white and is 6 inches wide. The color scheme was developed using MUTCD standard colors to simulate a white lane edge line along with the orange to indicate construction. The rise in the sloped section was 3 inches, based on GDOT Standard 9032B ("Concrete Curb & Gutter, Concrete Curbs, Concrete Medians," October 2011) for a "Raised Edge with Concrete Gutter." These raised edges are approved for use on high speed arterials and freeways.

The visibility of the trapezoidal base is augmented by the periodic introduction of vertical pylons. The pylon design followed the specifications outlined in Section 6F.65 Figure 6F-7 of the MUTCD, "Tubular Markers" (FHWA 2009). The material of the linear channelizing device is not specified since it has only been represented virtually, though it is intended to be highly flexible when traversed providing minimal to no physical resistance to impact.

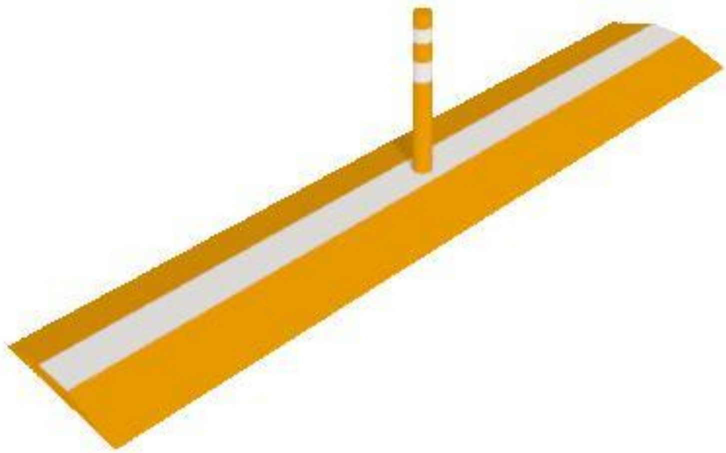

Figure 6: Illustration of the Linear Channelizing Device used in Experiment 2. These segments would be placed end-on-end to create both closure and continuity.

Experiments

For each experiment, participants were seated at individual computer workstations in the same room as other participants. After some brief comments from the proctor and a few introductory slides to familiarize the participants with the computer configuration, participants began marking responses on the stimulus images.

During the course of each experiment, participants were shown a series of static images and asked to identify if the ramp shown was open or closed to traffic. If the ramp was open, they were asked to move the cursor to the ramp location and click the left mouse button. If the ramp was closed, the participants were asked to identify this condition by clicking on “Exit Closed” icon on the lower left corner of the image (Figure 7). In Experiments 2 and 3, an additional “Don’t Know” icon was added to the top left of the image to allow participants an additional response option.

Between images, participants were asked to click in a region on a transition image (Figure 8) to return their mouse cursor to a consistent starting position. Having a fixed intial cursor position allows for consistent measurement of response latency (i.e. time from initial image display to participant response) that can also be used in analysis of participant responses. If, for any reason, a participant did not respond to an image within an allotted time (3 seconds in Experiment 1 or 3.5 seconds in Experiments 2 and 3), the image would time-out and the transition image would be displayed. The transition image did not time-out. The participants were required to click on the + sign (see Figure 8) to exit the transition image.

Also as stated, with the exception of a base case image without a work zone, each test image showed a particular freeway alignment with a ramp and a work zone defined by delineation devices in one of the various configurations. The ramp was closed in half of these images. The number of delineation device configurations and time-to-exit locations (i.e. travel time from image view point to beginning of diverge taper) varied by the experiment, as did the number of replicate images. However, the total number of test images shown in each experiment was restricted to a range of 800 to 1000. Within an experiment all participants where shown the same images before and after the rest period although the image order during each time period was randomized for each participant. In an effort to control for a practice effect, each participant saw all of the images exactly once before any image was repeated; e.g. for Experiment 1, the 90 slides were presented in ten fully randomized blocks, with random draw without replacement until all images had been presented. The overall time required varied by participant, but ranged from less than 45 minutes to one-hour. A more detailed description of the images used in each experiment is provided in the next section.

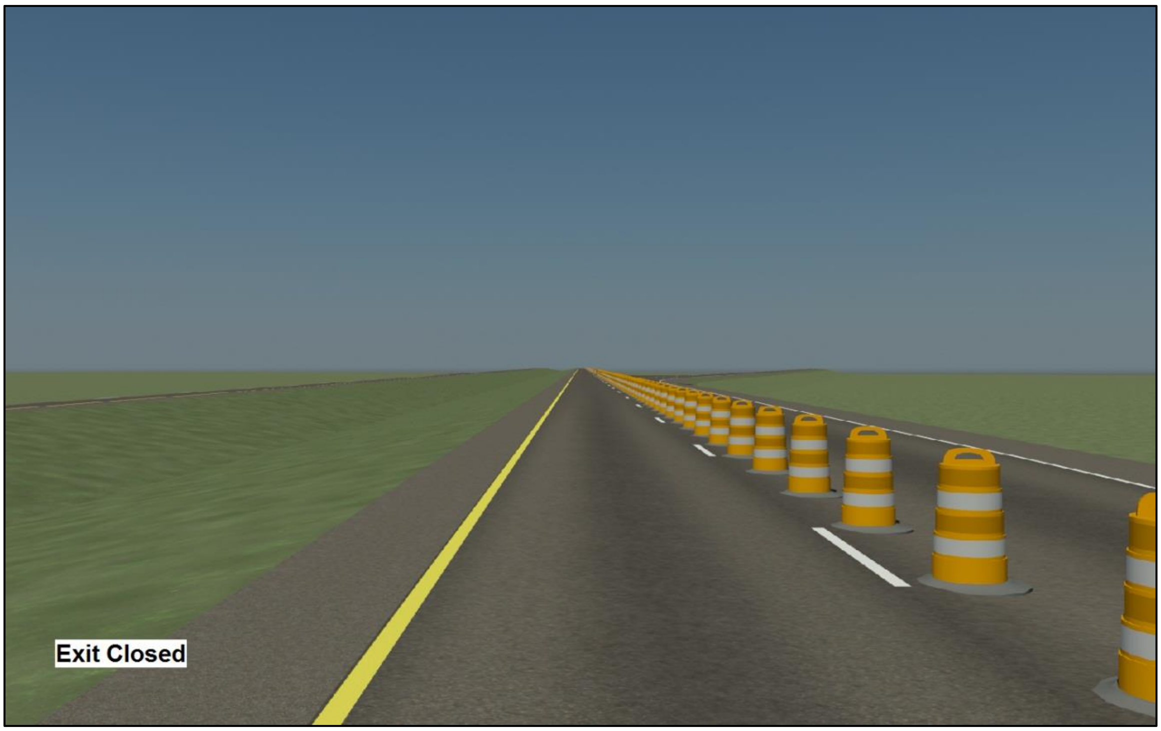

Figure 7: Example slide of a work zone diverge with instructions used in Experiment 1. Figure shows the EXIT CLOSED button to click if the ramp is closed.

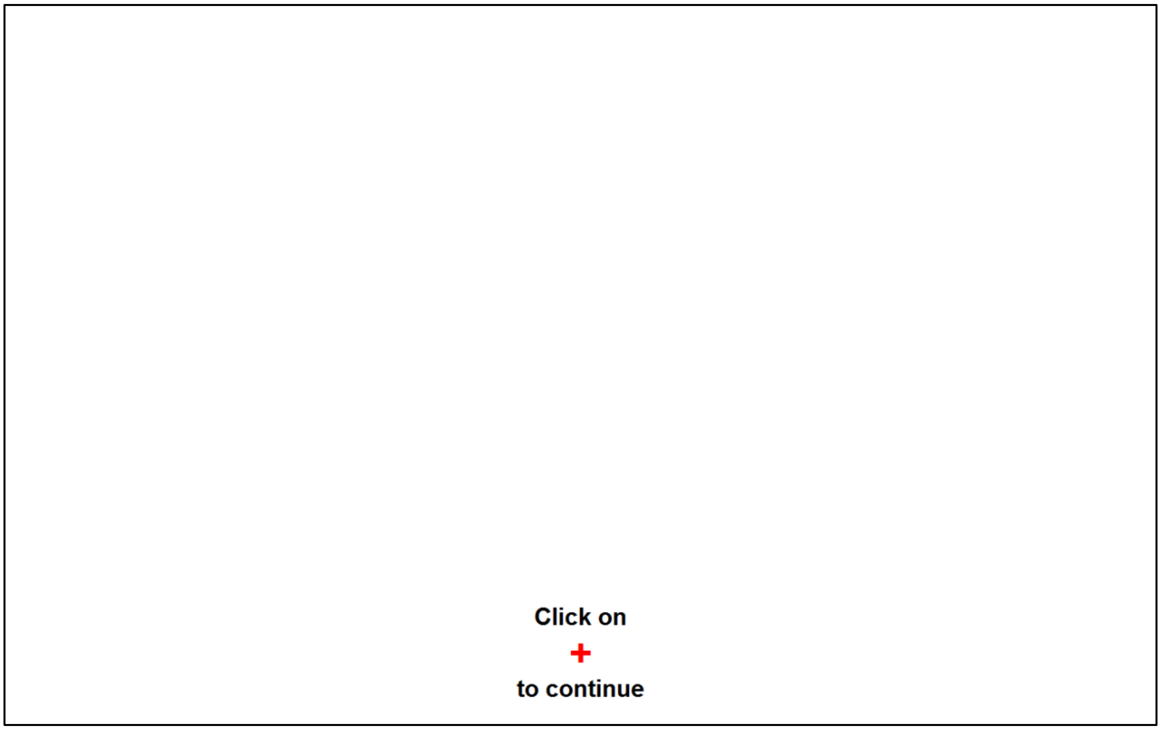

Figure 8: Transition slide used between roadway images. Participants were required to click the target to reset the position of the mouse before each image.

Experiment 1: Existing Channelizing Devices

In Experiment 1, participants were shown rendered static images of ramps configured using various combinations and configurations of existing delineation and channelization devices. This experiment was designed to examine a broad range of existing devices, roadway geometries, and time-to-exit distances. This broad experiment had two principal objectives. The first objective was to provide a preliminary evaluation of the limitations of existing delineation treatments and to explore possible design principles that could be used to develop new devices or methods for overcoming these limitations. The second objective was to evaluate the roadway geometries and time-to-exit distances that could best be used to evaluate more complex conditions later in the project. Experiment 1 explored the following features:

Delineation/Channelizing devices at diverge:

- Drums spaced 40 feet apart
- Drums spaced 10 feet apart
- Drums spaced 40 feet apart with up to 2 feet of random placement error
- Portable Concrete Barriers

Geometries:

- Taper type exit with freeway alignment straight
- Taper type exit with freeway alignment curve to left

Times-to-exit (travel time at 60 mph to the beginning of the diverge taper):

- 5 seconds from the diverge taper
- 4 seconds from the diverge taper
- 3 seconds from the diverge taper
- 2 seconds from the diverge taper
- 1 second from the diverge taper

Ramp Condition:

- Open
- Closed

In addition an Open Ramp Condition for a "No Work" configuration was included as a control. In all work zones with drums a 120 ft. spacing was utilized upstream of the diverge.

A static image was generated for each channelizing device configuration (4 alternatives), geometry (2 alternatives), time-to-exit (5 alternatives), and ramp condition (2 alternatives) combination, for a total of 80 distinct images. Additionally, a "No

Work" static image was generated for the open ramp condition for each time-to-exit and road geometry for a total of 10 additional separate images. Ten replications of each static image were generated resulting in a total of 900 images shown to each participant. For each participant, five replicates (450 images) of each image were shown such that each image was shown at least once before moving to the next replication, followed by a rest period, followed by an additional five replicates (450 images). Each set of 450 images was presented in a different random order to each of the participants. The rest periods were of variable duration, from a few minutes to 10 minutes. The maximum duration of the experiment was one hour. Most participants completed the experiment within 45 minutes.

Experiment 2: New Channelizing Device

Similar to Experiment 1, in Experiment 2, participants were shown rendered static images of ramps configured using various configurations of existing delineation/channelization devices with addition a new linear channelizing device (LCD). As described earlier (section 3.3), the LCD was developed based on the results of Experiment 1. The primary purposes of Experiment 2 were to: 1) evaluate the linear channelizing device (LCD) and 2) to further examine the design principles evaluated in Experiment 1 in a more focused setting. Experiment 2 explored the following features: Delineation/Channelizing devices at diverge:

- Drums spaced 40 feet apart
- Drums spaced 40 feet apart with up to 2 feet of random placement error

- Drums spaced 40 feet apart missing 10% with up to 2 feet of random placement error (To ensure that a single random configuration was not disproportionately impacting data, two random variations were included.)
- Drums spaced 10 feet apart
- Drums spaced 10 feet apart with up to 2 feet of random placement error
- Drums spaced 10 feet apart missing 10% with up to 2 feet of random placement error (To ensure that a single random configuration was not disproportionately impacting data, two random variations were included.)
- Portable Concrete Barriers
- Linear Channelizing Device
- Linear Channelizing Device missing 10% of posts

Geometries:

- Taper type exit with straight freeway alignment
- Taper type exit with freeway aligned with curve to the right

Times-to-exit (travel time at 60 mph to the beginning of the diverge taper):

- 5 seconds from the diverge taper (straight geometry only)
- 3 seconds from the diverge taper
- 2 seconds from the diverge taper (curved geometry only)
- 1 second from the diverge taper

Ramp Condition:

- Open
- Closed

In addition an Open Ramp Condition No Work configuration was included as a control. As with Experiment 1, in all work zones with drums a 120 ft. spacing was utilized upstream of the diverge

As with Experiment 1, images were generated for each delineation/channelization device, geometry, time-to-exit, and ramp condition combination, except as noted (e.g. 5 second time-to-exit only applied to the freeway straight alignment). Additionally, a “No work” image was generated for the open ramp condition for each time-to-exit value. In total, 138 separate still images were created.

For the experiment, the participants were shown six replications of each image resulting in a total of 828 images for which responses were recorded. For the channelizing device alternatives with missing devices the six images were composed of three replications for each of two sub-alternatives. The images were presented in a random order for each participant.

Similar to experiment 1, a rest period was provided at the midpoint. Again, most participants completed the experiment within 45 minutes.

Experiment 3: Varying Roadside Environment and Construction Equipment

To verify and expand the results from Experiment 2, various roadside vegetation and equipment combinations were added to the scenes to evaluate the impact of increasing the overall visual complexity of the scenes for a subset of conditions. Experiment 3 explored the following conditions:

Delineation/Channelizing devices at diverge:

- Drums spaced 40 feet apart
- Drums spaced 40 feet apart with up to 2 feet of random placement error

- Portable Concrete Barriers
- Linear Channelizing Device
- Linear Channelizing Device missing 10% of posts

Geometries:

- Taper type exit with straight freeway alignment

Times-to-exit (travel time at 60 mph to the beginning of the diverge taper):

- 3 seconds from the diverge taper
- 1 second from the diverge taper

Ramp Condition:

- Open
- Closed

Vegetation:

- No vegetation (not presented with equipment)
- Trees along the right edge of the corridor
- Trees along the left edge of the corridor
- Trees along both edges of the corridor
- Trees along the right edge of the corridor and in the median
- Light vegetation on both edges of the corridor

Equipment:

- No equipment
- Three pieces of construction equipment (Configuration A, Figure 9)
- Three pieces of construction equipment (Configuration B, Figure 10)

As with previous experiments in all work zones with drums a 120 ft. spacing was utilized upstream of the diverge

Figure 9: Illustration of Experiment 3 Equipment Configuration A

Figure 10: Illustration of Experiment 3 Equipment Configuration B

As with previous experiments, an image was generated for each combination of the listed features. These combinations generated 320 separate static images. Three replications of each image were generated resulting in a total of 960 images that were shown to each participant. The images were presented to each participant with two rest periods occurring after each set of 320 images. Images within the set of 320 images were presented in a different random order for each participant. Participant rest periods were of variable duration and the maximum duration of the study was one hour. Most participants completed the experiment within 45 minutes.

Data Processing

The data collected from each participant were the X, Y coordinates of their mouse click locations within the various images as well as the time from the instance the image was displayed to the time of the mouse click. To assess the accuracy with which each participant was able to correctly identify the ramp condition (open/closed) and the ramp location (for the ramp open condition), each image was divided into zones for classifying each participant's responses according to the location they clicked on the screen. Participant responses were classified as Ramp Closed, Exit Open, Work Zone, Don't Know, and Indeterminate as described below. Figure 11 and Figure 12 illustrate an overlay of the zoning system on a rendered image for Experiment 1 and Experiments 2 and 3, respectively.

- *Ramp Closed*. A participant response indicating the ramp was closed was recorded if the participant clicked on the zone located in the bottom left of the screen. On all images an EXIT CLOSED text box was shown in this area.
- *Exit Open*. An exit open response was registered if the participant clicked on the ramp diverge location. This response indicates that the participant interpreted the ramp as open and correctly identified the diverge location. This zone is defined based on the judgement of the research team as an area bounded by: 1) a line 2/3 of the distance from the initial cursor position to the ramp opening centroid; 2) a line parallel the horizon including a 50 pixel buffer; 3) lines drawn from the initial cursor position to the outside edges of the ramp opening; 4) lines drawn from the visible portions of the channelizing devices used to delineate the ramp opening.
- *Work Zone*. This zone includes the construction zone and the adjacent area above the horizon, to the right of the exit. This participant response indicates the participant interpreted the ramp as open however incorrectly identified the diverge location as being in the construction area.

- *Don't Know*. In Experiment 2 and Experiment 3, a zone labeled "Don't Know", as indicated by the white "Don't Know" button in Figure and Figure , was included in the upper left section of the screen to allow the participant to indicate they are unable to determine the status or location of the diverge.
- *Indeterminate*. The remaining area in the image was zoned indeterminate. If the participant's response was recorded in this area, it is not known if the participant intended to indicate the ramp diverge as open or closed.

To operationalize these definitions and to associate particular participant responses with a zone, the data were imported into "R" statistical software (R Core Team, 2014). The "R" software package is an open source implementation of the "S" statistical programming language originally developed by the Bell Telephone Laboratories in the 1970's. A set of R scripts using the "point.in.polygon" command was developed to first overlay x, y coordinates of each participant's responses on to the still images, and then to process the graphical data into spreadsheets containing binary information indicating the zone in which each response was located.

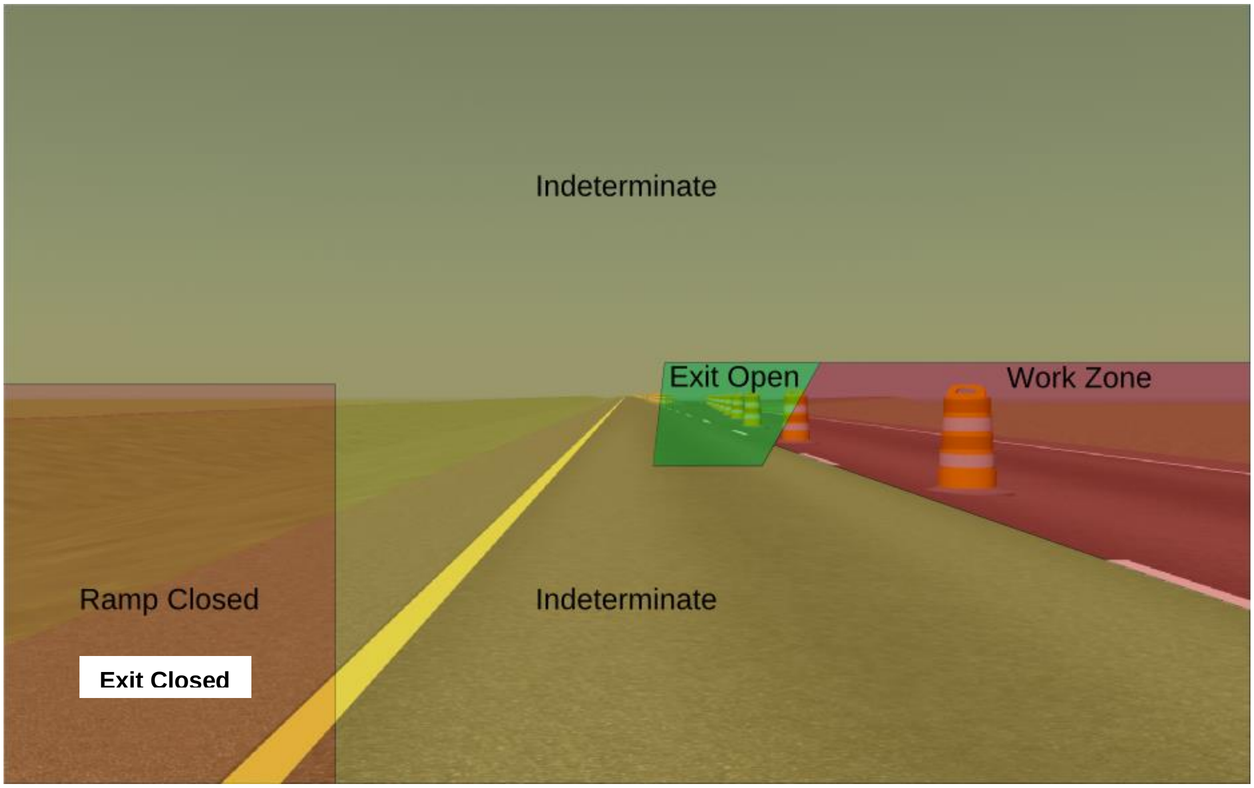

Figure 11: Zoning System for Classifying Responses in Experiment 1

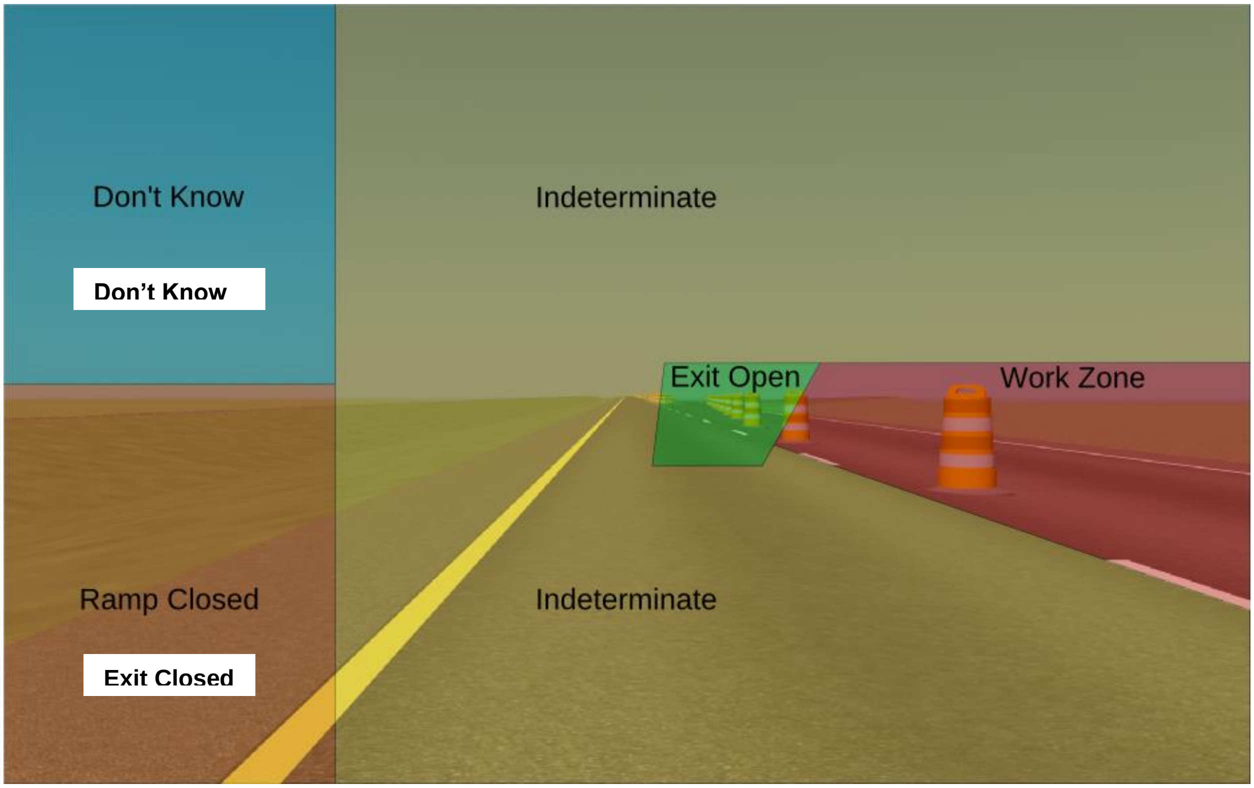

Figure 12: Zoning System for Classifying Responses in Experiments 2 and 3

Results

For each of the experiments, the results from the individual participants for all replicates of a particular combination of conditions were combined to produce three different types of descriptive results related to the speed and accuracy that the participant could identify the ramp position and location. In turn, the individual results could also be further aggregated to produce results for the entire experimental cohort. These results were:

***Percent Correct*:** A response is correct if for an open ramp a participant correctly identified the ramp as open and correctly identified the ramp location. For a closed ramp, the response is correct if the participant correctly identified the ramp as closed. For example, for an open ramp a response was considered correct when the participant's click was within the zone indicated by "Exit Open" in Figure 11 for experiment 1 or Figure 12 for experiments 2 and 3. Thus, for an open ramp, 80% correct indicates that 20% of a participant's responses were either clicks outside of this zone, or non-response due to time-out. Likewise, for a closed ramp an 80% correct indicates that 20% of a participant's responses were clicks outside of the "Ramp Closed" zone in Figure 11 for experiment 1 or Figure 12 for experiments 2 and 3, or non-response due to time-out.

Error Analysis: Two types of errors are analyzed. The first, referred to as an *identification error*, occurs when a participant incorrectly identifies the ramp condition (i.e. as open when closed or as closed when open). The second type of error, referred to as a *diverge location error*, occurs when a participant clicks the active work zone. This latter error can arise in two ways. For an open ramp, a diverge location error occurs when the participant incorrectly identifies the location of the diverge as being within the work zone. In the case of a closed ramp, a diverge location error occurs when the participant incorrectly identifies the ramp as open and indicates a diverge location in the active work zone and not at the intended diverge point.

Latency: Latency is the measure of the time between when the image is displayed and a click response is recorded. Correct response latencies measure the time to react, process, and perform an appropriate action regarding the scene.

Experiment 1: Existing Channelizing Devices

Experiment 1 focused on examining human performance resulting from existing channelizing devices and configurations at varying times-to-exit and geometries. Data were collected for 41 participants, two of whom were excluded for excessive non-responses (fewer than 25% of responses were outside the Indeterminate Zone). The remaining 39 participants were included in the subsequent analyses.

<u>Percent Correct</u>

The overall percent correct across all responses was 82.7%. The "No work" alternative averaged 73.5% for correct responses. Consistent with earlier studies, the portable concrete barriers (PCBs) resulted in the highest overall percent correct, averaging 91.5 % across participants. The second highest overall correct response rates were for aligned Drums spaced either 10 feet or 40 feet apart at the diverge, both alternatives with 82.8% percent correct. The slightly misaligned Drum alternative (40 ft. +/- 2 ft.) had a slightly lower overall correct response rate at 78.4% correct. While overall correct rates (average correct over all time-to-exit, geometry, and ramp open/closed conditions) for each delineation device tended to differ by a small percentage it will be seen that correct rates for certain conditions (e.g. higher time-to-exit locations) could differ dramatically.

Table 1: Percent Correct Responses for Experiment 1 – Curved Geometry

Condition	Time-to-Exit	10 ft. Drums	40 ft. Drums	40+/-2 ft. Drums	PCB
Open	5	84.62%	84.87%	85.13%	81.03%
	4	86.92%	87.69%	87.95%	85.13%
	3	90.26%	86.67%	86.41%	88.97%
	2	88.97%	89.23%	89.23%	84.10%
	1	91.28%	88.21%	91.28%	90.51%
Closed	5	80.00%	81.28%	85.64%	96.92%
	4	81.28%	84.87%	84.10%	95.13%
	3	84.87%	85.13%	87.69%	96.15%
	2	95.64%	94.87%	90.26%	96.67%
	1	96.92%	96.41%	95.13%	97.69%

Table 2: Percent Correct Responses for Experiment 1 – Straight Geometry

Condition	Time-to-Exit	10 ft. Drums	40 ft. Drums	40+/-2 ft. Drums	PCB
Open	5	75.13%	76.92%	34.10%	76.92%
	4	78.72%	79.74%	45.38%	83.33%
	3	87.18%	86.67%	77.44%	87.95%
	2	92.05%	88.97%	88.72%	91.79%
	1	93.33%	93.59%	93.33%	94.36%
Closed	5	62.31%	63.85%	60.77%	97.69%
	4	66.15%	67.44%	70.77%	96.15%
	3	65.90%	63.85%	70.26%	96.41%
	2	61.79%	64.10%	69.74%	95.90%
	1	92.82%	91.54%	74.36%	96.15%

Types of Errors

Figure 13 summarizes the error analysis associated with the alternatives examined in Experiment 1. These error types were generated using the categorization method presented earlier, although there is an implicit assumption that a participant's click location indicated their intended response. In this figure, in the straight geometry and open condition, errors increased as the time-to-exit increased across all channelization alternatives. For Drums spaced 40 feet apart, there were few errors at the 1, 2, and 3 second times-to-exit, with most incorrect responses being categorized as Indeterminate. At 4 and 5 seconds, error rates exceeded 20%, mostly due to Indeterminate responses but also due to an increase in both Identification errors (stating the work zone was closed when it was open) and Diverge Location errors (identifying the diverge as in the construction area). Drums 10 ft. apart had a similar pattern of participant error. A distinctly different pattern was observed for Drums 40 ft. +/- 2 ft. For the misaligned drums Identification errors increased as the time-to-exit increased, from 0.51% at 1 second away from the diverge to 50% at 5 seconds away from the diverge, making the primary error at larger distances identifying the diverge as closed when it is open. For this misaligned drum alternative, Diverge Location errors also increased with distance, from zero at 1 second away to 6.92% at 5 seconds away from the diverge. At distances of 4 to 5 seconds from the diverge the Drums 40ft +/- 2ft alternative also began to see an increase in participant time-out conditions, a potential additional indication that the participants had difficulty in interpreting these scenes. PCB resulted in the best participant performance with Indeterminate responses dominating the recorded errors and almost no identification errors.

In the straight geometry and closed condition, Figure 14, the dominant error type across drum alternatives was the Diverge Location error. At 1 second from the diverge, Drums 40 ft. +/- 2 ft. showed a Diverge Location error rate of 13.33% and an Identification error rate of 7.69%. At 2 to 5 seconds from the diverge, all drum alternatives showed high Diverge Location errors, ranging from 15.64% to 29.23%. Identification errors resulting from drum alternatives at 2 to 5 seconds away ranged from 2.31% to 7.95%. In contrast, portable concrete barriers resulted in very few Diverge

Location or Identification errors across all distances. The highest PCB Diverge Location error rate was 0.51% at 4 seconds away from the diverge, and the greatest percent of Identification errors was 1.28% at 2 seconds.

Patterns of error rates were more difficult to identify from the curved geometry when ramp was open as the errors were generally smaller than for the straight alignment. Figure 15shows the error type distribution for the curved geometry when ramp was open. Figure 16shows the error type distribution for the curved geometry when the ramp was closed. At 1 second away from the diverge, all channelizing devices resulted in no Diverge location errors. Drums 40 ft. +/- 2 ft. resulted in the greatest percent of Identification errors at 1.54%, among all alternatives of channelizing devices. At 2 seconds away from the diverge, drum alternatives began to result in Diverge Location errors, although all were below 4% with PCB at 0% work zone errors. Identification errors also showed a similar trend for all channelizing devices. At 3, 4, and 5 seconds away, PCB continued to show zero Diverge Location errors, and very low Identification errors, while the drum alternatives showed increasingly greater Diverge Location and Identification errors. Interestingly, three participants consistently made Diverge Location errors when the drum alternatives were used, but not for the portable concrete barrier alternative, suggesting that the gaps between drums may have a more pronounced effect on some individuals than for others. Even though the percent of errors was low, Diverge Location and Identification errors increased for drum alternatives as time-to-exit increased.

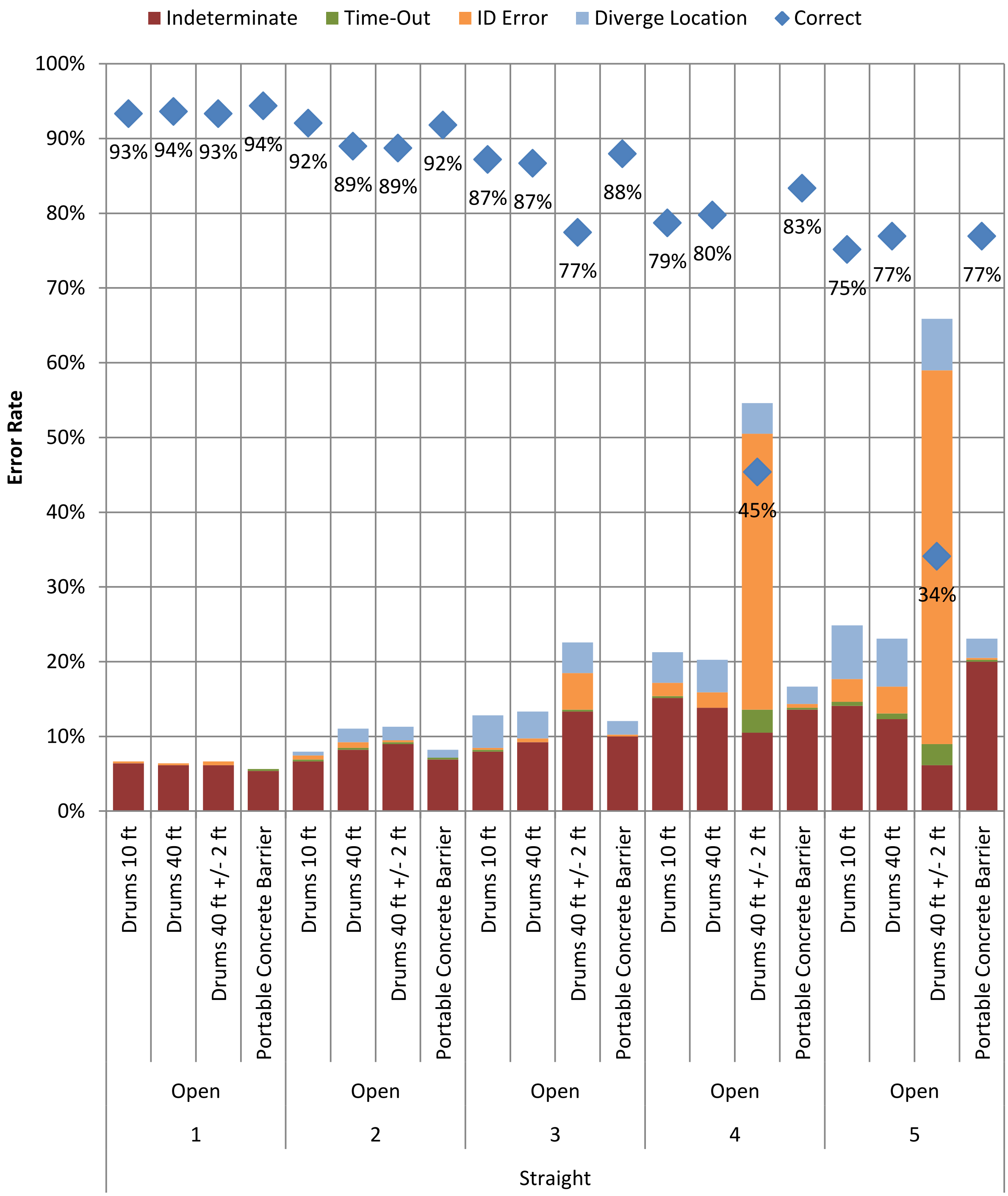

Figure 13: Experiment 1 – Percent Errors in the Straight Geometry and Open Condition
Numbers below the blue dots indicate the percent of correct responses.

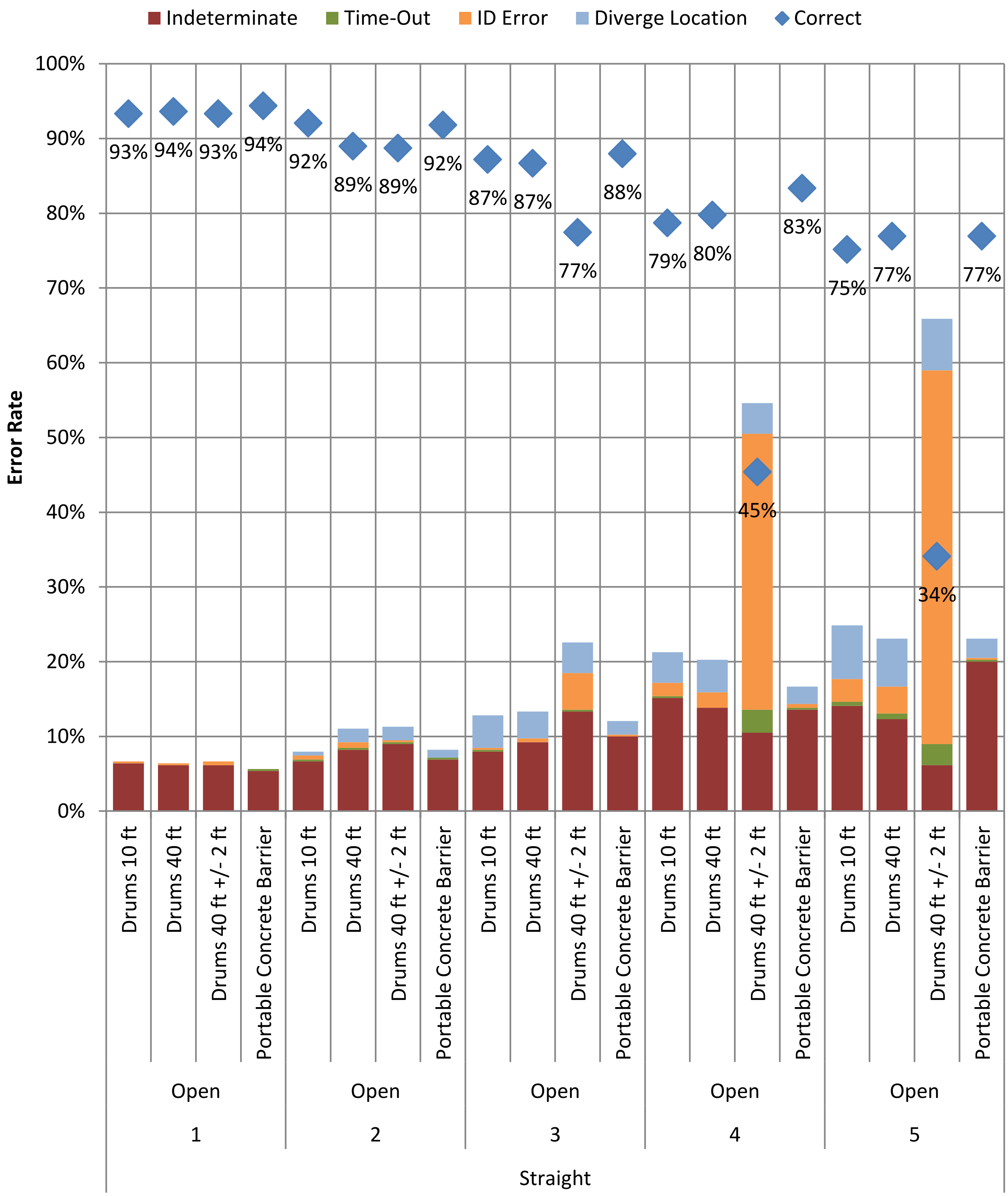

Figure 14 Experiment 1 – Percent Errors for the Straight Geometry and Closed Condition
Numbers below the blue dots indicate the percent of correct responses.

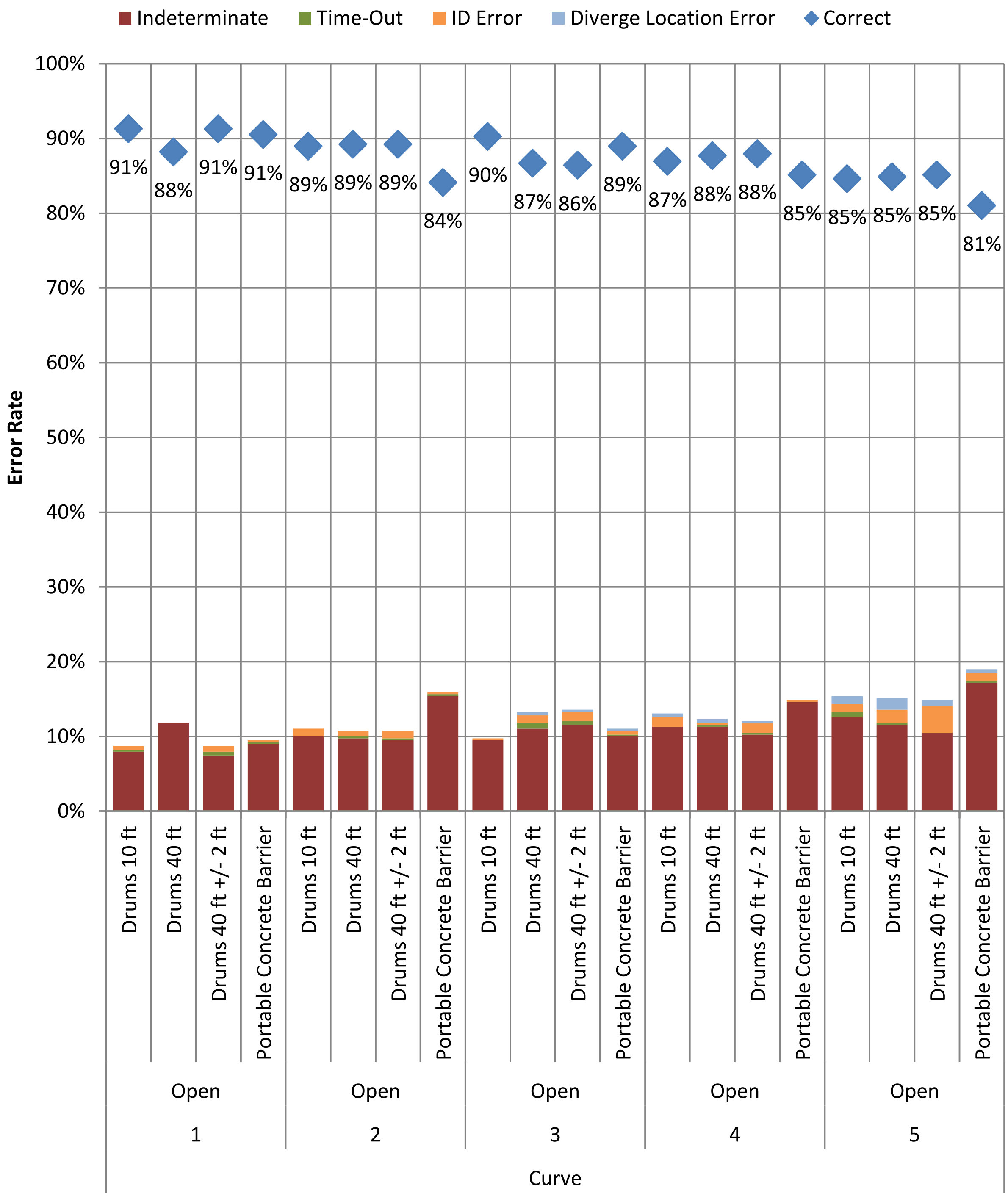

Figure 15 Experiment 1 – Percent Errors for the Curved Geometry and Open Condition
Numbers below the blue dots indicate the percent of correct responses.

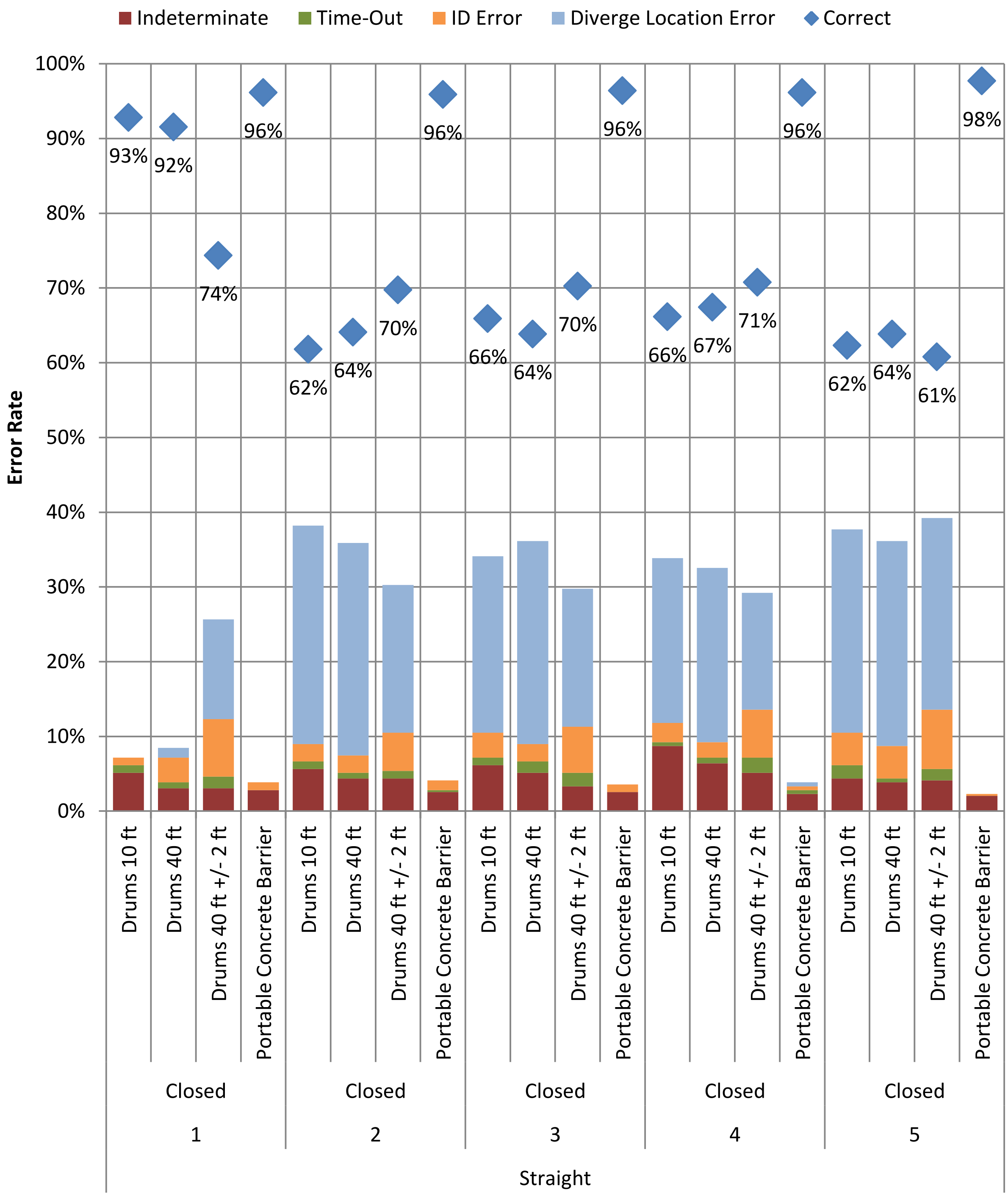

Figure 16 Experiment 1 – Percent Errors for the Curved Geometry and Closed Condition
Numbers below the blue dots indicate the percent of correct responses.

Experiment 2: Novel Channelizing Device

Based on the results of Experiment 1, the linearity and continuity of the discrete devices (delineators) used for channelization were critical elements to high accuracy in identification of the ramp diverge. These observations were the principal influence for the development of the linear channelizing device (LCD) described earlier. Experiment 2 added this linear channelizing device (LCD), to the spectrum of channelizing devices, and also added additional random placement combinations scenarios to the 40 +/- 2 ft. drum alternatives in Experiment 1 to ensure the results were not a product of a specific drum placement configuration. For this experiment, student participants from Morehead State University were used rather than students from Georgia Tech because of difficulty recruiting Georgia Tech students. Among the 51 original participants at Morehead State, data from 4 participants were excluded from analysis due to excessive non-response, resulting in 47 participants in the final dataset.

In addition to the inclusion of the LCD and modifications to the misaligned drum (+/- 2 ft.) alternatives, several other modifications to the Experiment 1 protocol were made in design Experiment 2. Based on the limited information provided by the curved geometry, only the straight roadway geometry was used in Experiment 2. In addition, two random misalignment options for the 10 ft. +/- 2 ft. alternative were included to see if proximity of devices affected performance when continuity was disrupted. Finally, clear trends observed in time-to-exit distances allowed for a reduction in distances included in Experiment 2, to 1, 3, and 5 seconds from the diverge point.

The results of Experiment 2 are similar to those of Experiment 1. The drum alternatives had lower correct response rates than the PCB alternative, especially at the 5s distance; this implies that the issues of closure and continuity observed in Experiment 1 are still relevant. This is reaffirmed with the very similar results between the PCB and LCD alternatives, especially since the LCD was designed explicitly to have both continuity and closure. Also, the issue of proximity was still not seen to be significant, with differences between the 10 ft. drum alternatives and the 40 ft. drum alternatives being very small.

Percent Correct

The overall percent correct for Experiment 2 was 67.7%. The No Work alternative, used for control purposes, resulted in an average of 69.7% correct. As with Experiment 1, PCBs resulted in the highest percent correct averaging 85.1%. The second highest percent correct resulted from the new LCD treatment with all pylons in place at 80.8%, with LCD missing 10% of the pylons slightly lower, averaging 78.6%). Also consistent with Experiment 1, the 10 ft. and 40 ft. spaced properly aligned Drums gave very similar results averaging 67.4% and 68.3% correct, respectively. The two misaligned Drum options were also very similar (61.6% and 59.7% correct) but notably lower than the properly aligned options. As with Experiment 1, it will be seen that correct rates across delineation types for certain conditions will differ dramatically more than the overall average values. For the straight geometry used in Experiment 2, Table 3 lists the percent correct for each alternative at each distance.

Table 3: Percent Correct Responses for Experiment 2

Condition	Time-to-Exit	10 ft. Drums	10 +/-2 ft. Drums	40 ft. Drums	40+/-2 ft. Drums	LCD	LCD-10%	PCB
Open	5	36.42%	16.05%	41.36%	25.51%	62.35%	58.33%	72.22%
	3	87.96%	65.02%	87.04%	66.77%	88.89%	90.12%	89.51%
	1	96.30%	95.68%	96.91%	94.86%	96.60%	95.99%	96.30%
Closed	5	53.40%	48.77%	52.16%	55.56%	78.40%	75.00%	83.64%
	3	56.17%	60.49%	54.32%	50.00%	78.70%	75.31%	83.33%
	1	69.75%	78.91%	72.22%	58.54%	77.16%	75.31%	83.33%

Types of Errors

Figure 17 shows the error type distribution for the open ramp condition. At 1 second away from the diverge, errors were low across all channelizing devices. At 3 seconds, Diverge Location errors start to increase among the drum alternatives and LCD, ranging from 2.5% for LCD to 7.2% for Drums 10 ft. +/- 2 ft. Identification errors resulting from Drums 10 ft. +/- 2 ft. and Drums 40 ft. +/- 2 ft. increase to about 13%, while the Identification errors of all other channelizing devices remained low, at or below 2%. At 5 seconds, Diverge Location and Identification errors increased across all channelizing devices, but the rate of increase is much greater among the drum alternatives than among PCB, LCD, and LCD missing 10% of pylons. For Drums 10 ft. +/- 2 ft. and Drums 40 ft. +/- 2 ft., every participant made at least two errors in identifying the open diverge point at 5 seconds away from the diverge. These results reinforce the Experiment 1 observation that a small amount of variation in drum placement can cause a significant increase in errors. Overall, for the open condition PCBs had the best performance of any alternative in the open condition. The linear channelizing device also resulted in few errors in the ramp open condition.

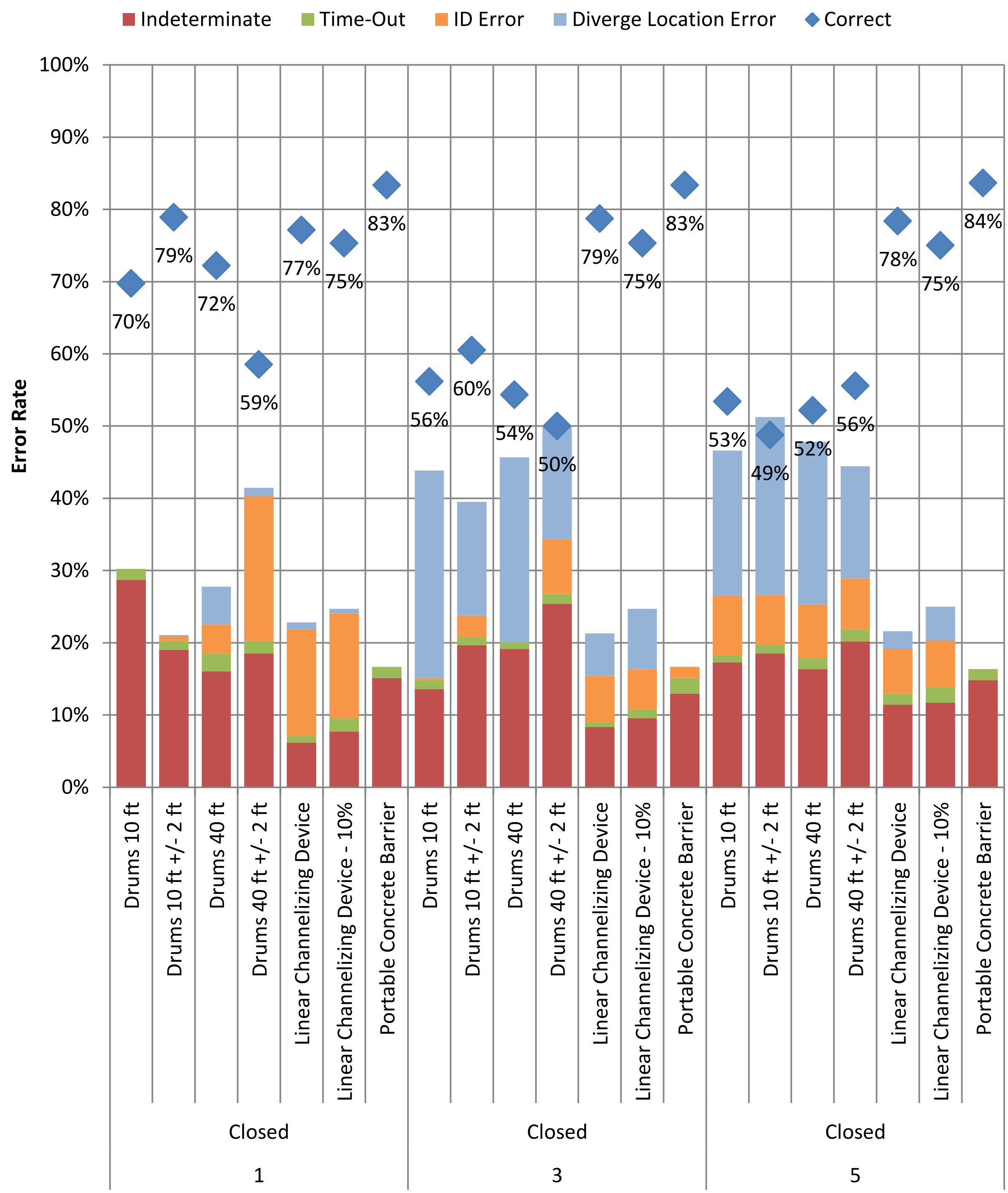

Figure Figure 18 shows the error type distribution when the ramp was closed. At 1 second travel time distance to the diverge, Drums 40 ft. +/- 2 ft., LCD, and LCD missing 10% of pylons had many Identification errors, at 20.1%, 14.8%, and 14.5%, respectively. The other channelizing devices all resulted in Identification errors of less than 5%. With regard to Diverge Location errors, properly aligned Drums at 40 ft. separation resulted in the greatest error rate at 5.25%. The second greatest percentage of Diverge Location errors, 1.1%, was observed with Drums 40 ft. +/- 2 ft. At 3 seconds away from the diverge, there were many Diverge Location errors for the drum alternatives, ranging from 15.6% to 28.7%. Drums 10 ft. +/- 2 ft., Drums 40 ft. +/- 2 ft., LCD, and LCD missing 10% of pylons resulted in greater Identification errors than the other channelizing devices. Similar trends were observed at 5 seconds away from the diverge, with the exception that the Identification errors for Drums 10 ft. and Drums 40 ft. were much greater at 5 seconds away than at 3 seconds away.

Similar to the trends under the open condition, when the ramp was closed, few errors were observed for all participants with the PCB. The LCD also resulted in good performance although with greater Identification errors, most notably at 1 second away from the diverge. These results closely mirrored the results from Experiment 1 and demonstrate the effectiveness of a device designed following the Gestalt principles.

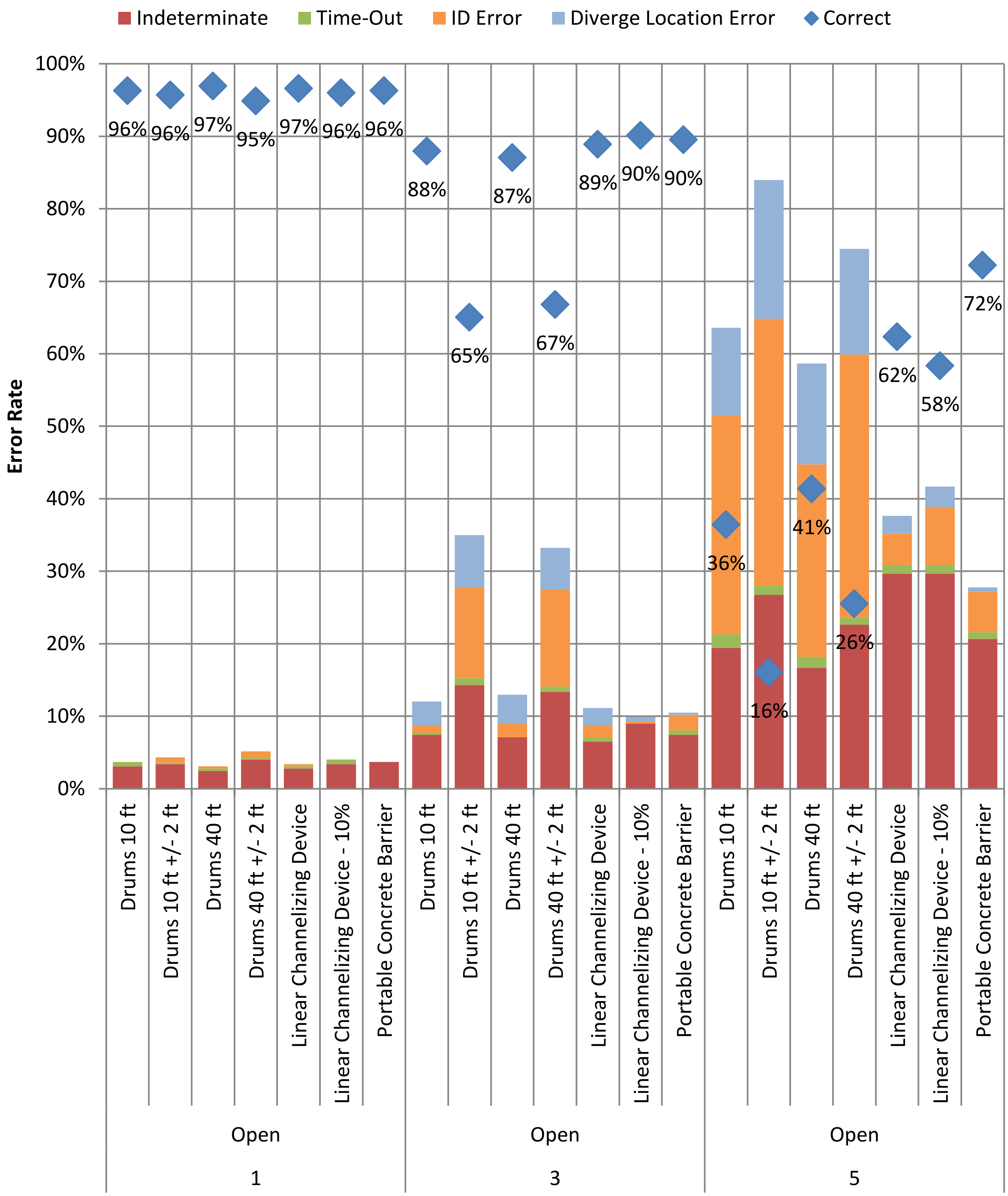

Figure 17 Experiment 2 – Percent Errors for the Straight Geometry and Open Condition
Numbers below the blue dots indicate the percent of correct responses.

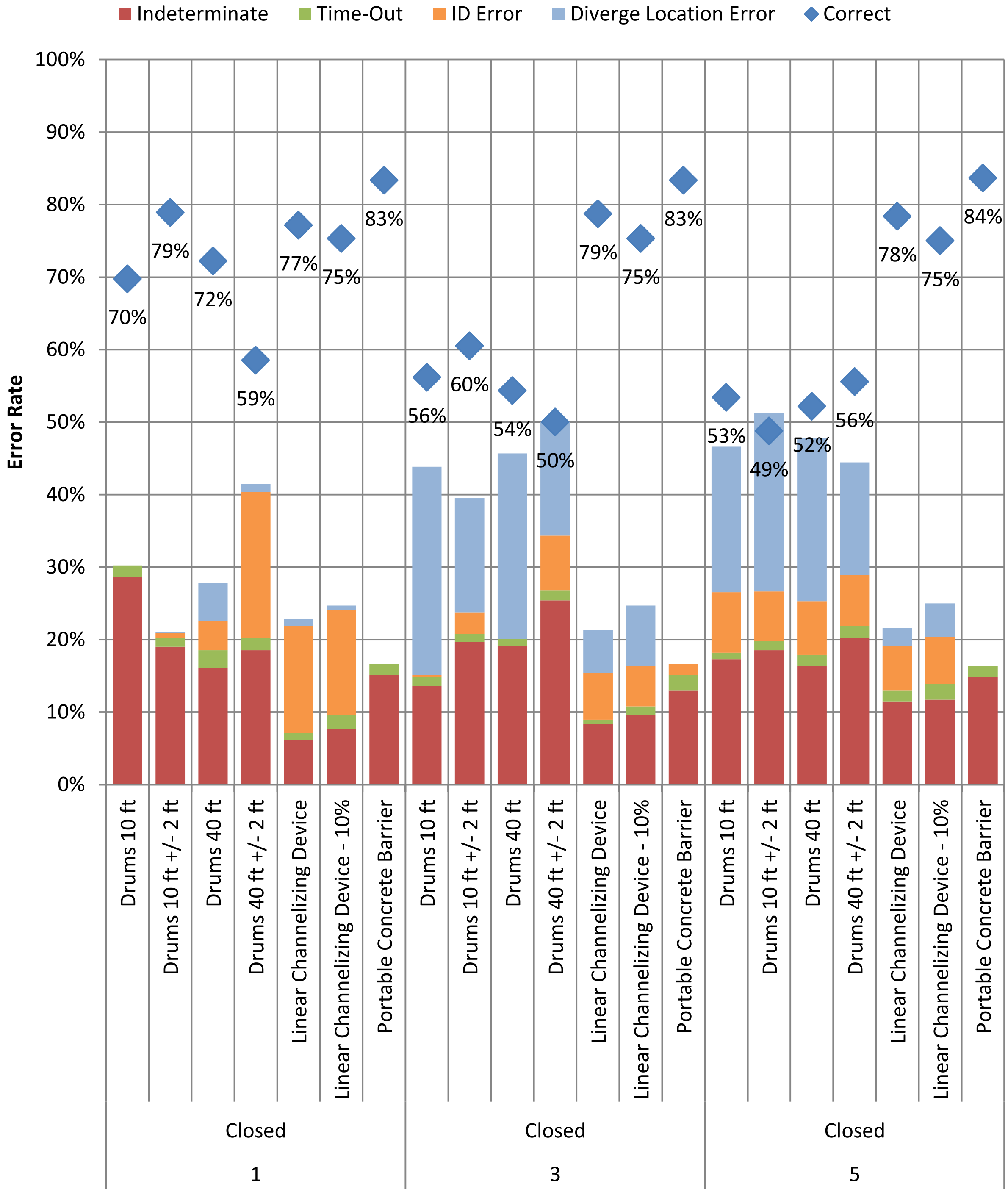

Figure 18 Experiment 2 – Percent Errors for the Straight Geometry and Closed Condition
Numbers below the blue dots indicate the percent of correct responses.

Experiment 3: Varying Roadside Environment and Construction Equipment

Experiments 1 and 2 focused on sparse, straight-line horizon backgrounds to eliminate visual clutter beyond that imposed by the channelizing devices. Experiment 3 introduced varied backgrounds and construction equipment, exploring the transferability of the results from Experiment 3 to more realistic environments. Experiment 3 was conducted at both the Georgia Institute of Technology (18 Participants) and Morehead State University (20 participants).

<u>Percent Correct</u>

The overall percent correct across all responses was 90.5%. The PCB, LCD, and LCD missing 10% of pylons alternatives resulted in similar high correct percentage at 94.2%, 94.7%, and 94.4%, respectively. The Drums 40 ft. alternative resulted in an average percent correct of 90.3%. At 78.9%, Drums 40 ft. +/- 2 ft. had the lowest percent correct. When comparing to Experiments 1 and 2 caution must be exercised as Experiment 3 does not include the 5 second travel distance to the diverge, which had the highest error rates. For instance, when considering only time-to-exit distance of one second and three seconds Experiment 2 straight geometry has percent correct rates of 88.74%, 85.99%, and 85.02% for PCB, LCD, and LCD missing 10%, respectively; the Drums 40 ft. and Drums 40 ft. +/- 2 ft. alternatives each resulted in 78.81% and 68.85% correct responses; the trend is similar to the Experiment 3 results.

Table 4: Percent Correct Responses for Experiment 3

Condition	Time-to-Exit	40 ft. Drums	40+/-2 ft. Drums	LCD	LCD-10%	PCB
Open	3	86.96%	77.63%	96.18%	95.76%	97.53%
	1	96.72%	81.96%	96.68%	95.14%	97.38%
Closed	3	86.96%	77.63%	96.18%	95.76%	97.53%
	1	96.72%	81.96%	96.68%	95.14%	97.38%

Types of Errors

Figure 19shows the error type distributions for the open ramp condition in Experiment 3. When the ramp was open, the percentage of errors at 1 second travel time from the diverge for all channelizing devices was very small. At 3 seconds from the diverge, LCD and LCD missing 10% of pylons resulted in very few Identification errors, both less than 1%. PCB resulted in slightly greater Identification errors, at 4.5%. Drums 40 ft. and Drums 40 ft. +/- 2 ft. resulted in the highest level of Identification errors at 8.5% and 24.3%, respectively. The percentage of Diverge Location errors were similar across channelizing devices, ranging from 2.3% for PCB to 4.4% for Drums with 40 ft. separation.

The error type distributions for the ramp closed condition in Experiment 3 are given in Figure 20. When the ramp was closed, error rates were generally small at both 1 second and 3 seconds time-to-exit but were more variable across channelizing devices than they were under the open ramp condition. At 1 second away from the diverge, Drums 40 ft. +/- 2 ft. resulted in the greatest percent of Identification errors at 7.5%, compared with LCD missing 10% of pylons at 2.2%, the second greatest percent of Identification errors. At 3 seconds away from the diverge, Drums 40 ft. +/- 2 ft. resulted in the greatest percent of Identification errors at 3.9%, while all other channelizing devices had a negligible Identification errors. When considering Diverge Location errors, Drums 40 ft. resulted in the greatest percent of errors at 6.1%. At 4.4%, Drums 40 ft. +/- 2 ft. resulted in the second most Diverge Location percent of errors.

This experiment also investigated the influence of different vegetation (Figure 21) and roadside equipment configurations (Figure 22) on performance. Generally, across all vegetation types, there were only slight differences in the resulting percentage of errors and error types. Similarly, for different equipment configurations the resulting differences in the percent of errors and error types was not significant. However, one notable exception was an increase in the percent of Diverge Location errors for alternatives without work zone equipment in the closed condition at 3 seconds. Further, over all vegetation and equipment alternatives the LCD and PCB alternatives still demonstrate strong performance advantages over drum alternatives.

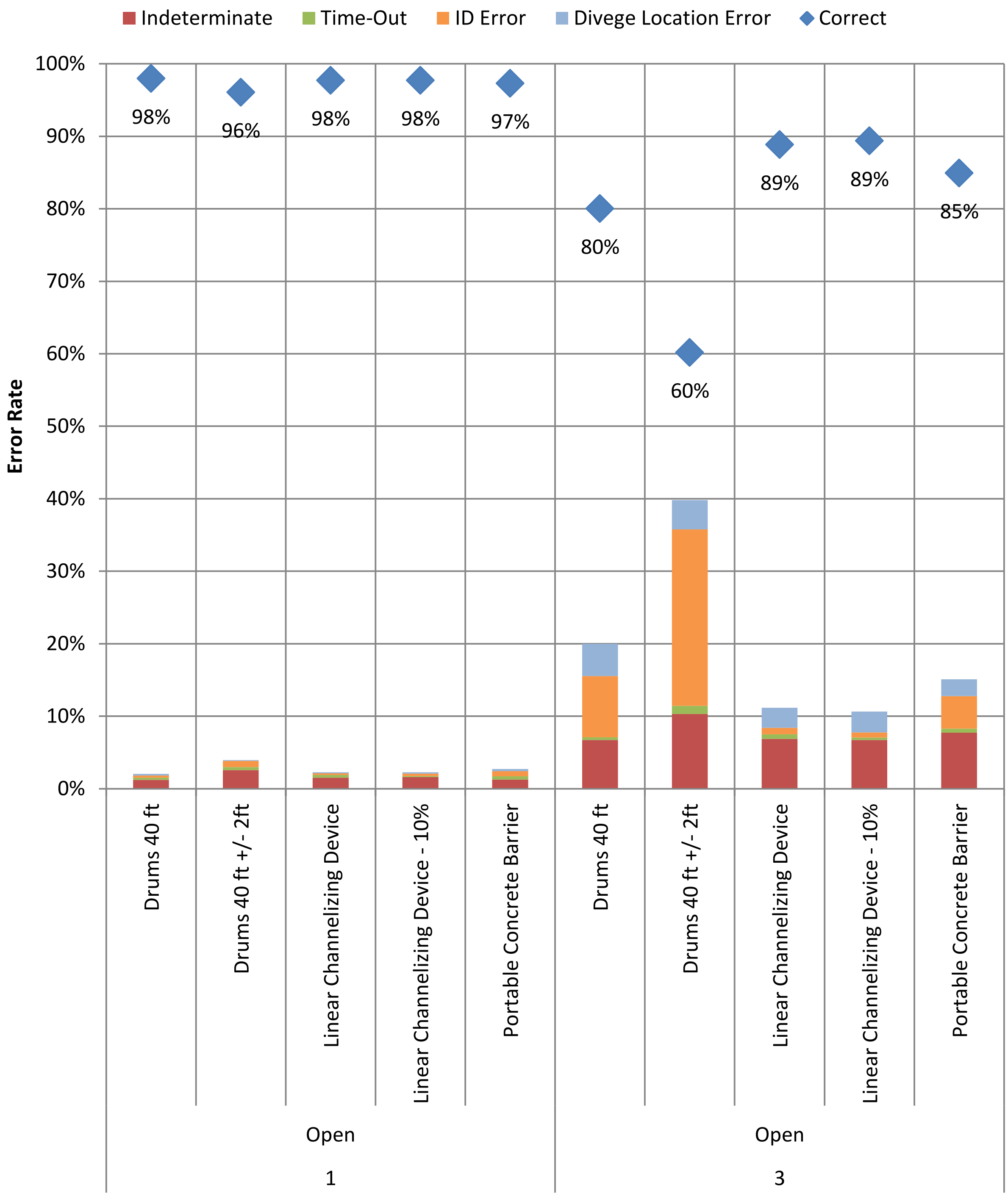

Figure 19 Experiment 3 – Percent Errors for Open Condition
Numbers below the blue dots indicate the percent of correct responses.

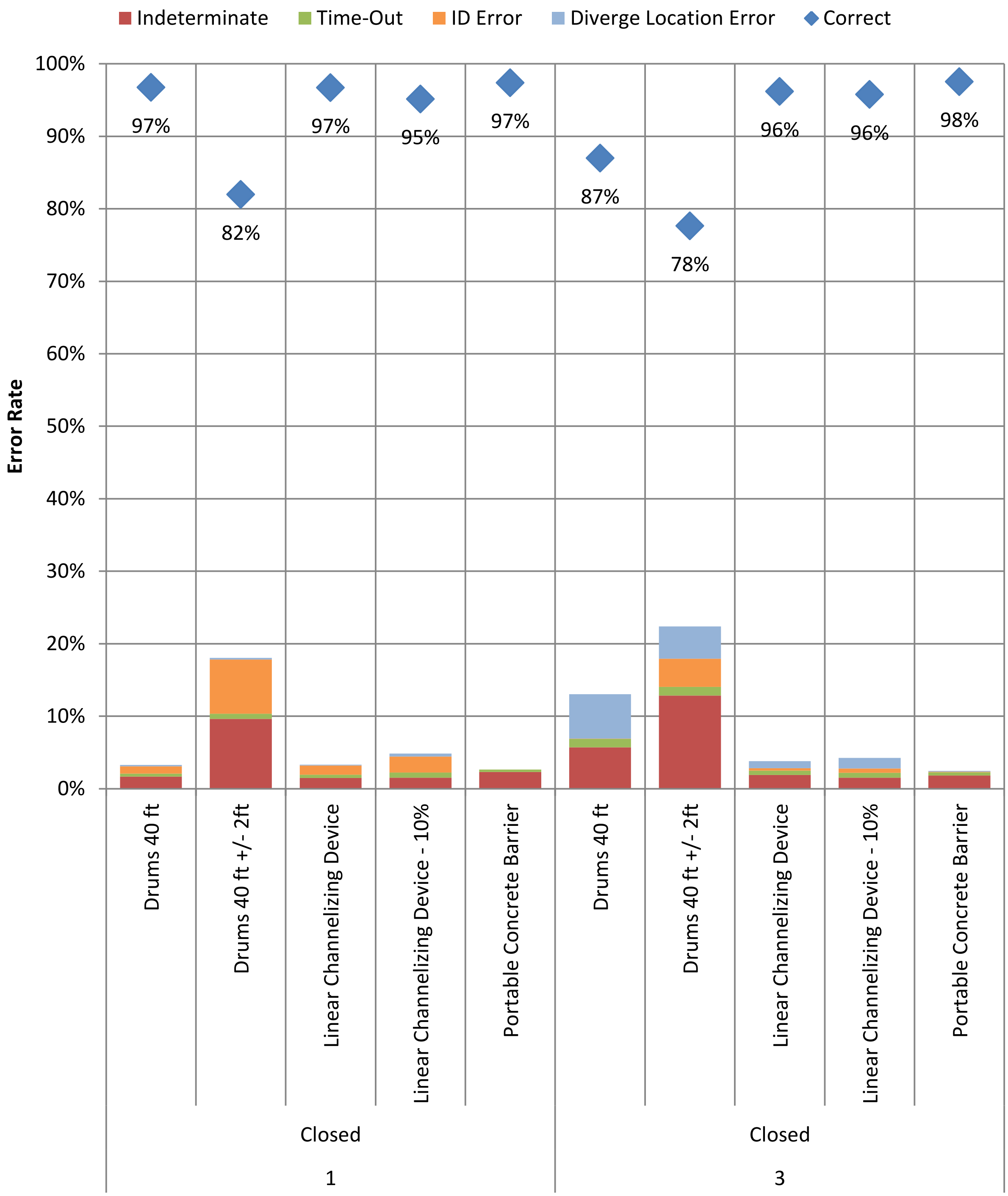

Figure 20 Experiment 3 – Percent Correct and Errors for the Closed Condition
Numbers below the blue dots indicate the percent of correct responses.

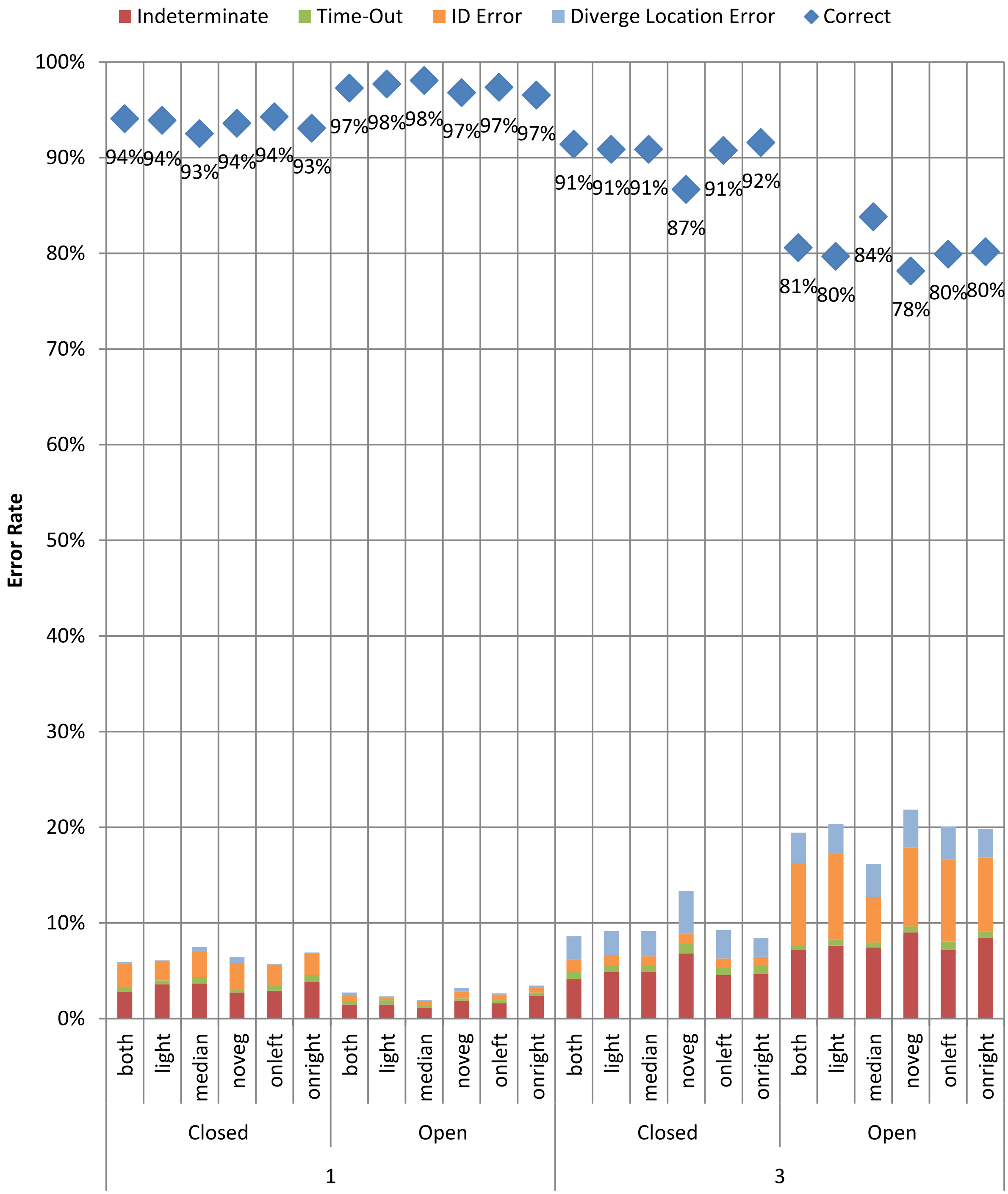

Figure 21 Experiment 3 – Percent Errors by Vegetation
Numbers below the blue dots indicate the percent of correct responses.

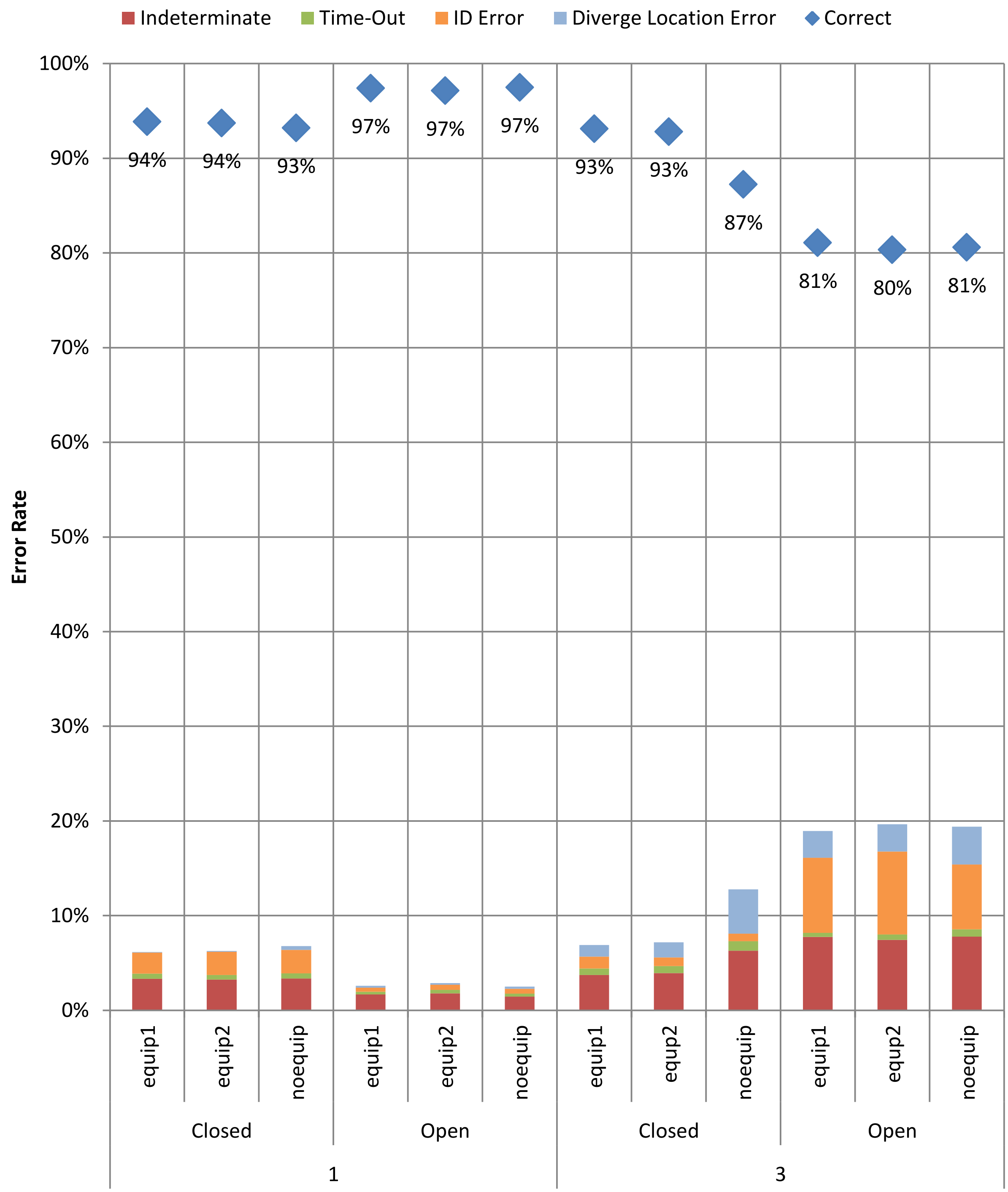

Figure 22 Experiment 3 – Percent Errors by Equipment
Numbers below the blue dots indicate the percent of correct responses.

Findings

While each of the experiments approached the issue of delineation in work zone diverges with varying combinations of devices and configurations, the results regarding each channelizing device were relatively consistent across experiments. In almost all circumstances under open ramp conditions, PCB, LCD, and LCD missing 10% of pylons resulted in better human performance than the drum alternatives. The Drums at 10ft. and 40ft. tended to perform similarly although at a level below that of the PCB and LCD alternatives. This implies that there is likely minimal advantage between the drum spacings considered. The drum alternative with +/- 2 ft. misplacements almost always resulted in significantly lower percent correct than other channelizing devices, indicating that participants found these alternatives most difficult to comprehend. As distance to the diverge increases, the differences between treatments becomes more discernible. Similar results were seen under ramp closed conditions with the exception that under longer time-to-exit distances the LCD and well-aligned Drum options tended to show similar Identification error rates. This trend that both drums and the LCD were difficult to comprehend at further distances in the closed condition may imply that when a construction project requires the full closure of a ramp that the PCB may be the best option. In addition, the impact of roadside vegetation and equipment was not discernible in most situations. However, at a significant distance from the diverge when the ramp was closed, scenarios without equipment showed greater errors. This observation indicates that the presence of equipment may provide additional cues signaling active work zones to drivers. Drivers may find that empty work zones without active construction to be more difficult to interpret than work zones with active work. Interestingly, this finding aligns well with earlier research conducted by Dixon et al. that reviewed crash data at Georgia work zones and found that most crashes occur while the work zones is idle.

Closure and Continuity

The study results follow the Gestalt principles very closely, especially those of closure and continuity. The principle of closure, as it applies to these circumstances, suggests that images that overlap in the visual scene may be perceived as a group. The portable concrete barriers are constructed to appear as a single object and benefits of closure appeared in the data as very low Diverge Location error rates. Similarly, the drums are perceived to overlap each other when they are far down the road, but are not perceived to overlap at shorter times-to-exit. This can even occur when the drums from the taper sections overlap with drums from the tangent section and thus give the impression of a single mass of drums. Finley et al. (2011) reported this feedback when using closely spaced drums.

For Experiment 1, the impact of closure (or lack of closure) can most easily be seen in the closed condition. Here, the PCB alternatives resulted in participants making few errors in the 5, 4, 3, and 2s times-to-exit on the straight geometry. The increased errors resulting from the drum alternatives were dominated by Diverge Location errors, where participants selected within the active work zone as the diverge location. However, these errors were not nearly as prevalent in the open condition, and no statistical differences existed between alternatives. This suggests that the break in closure from nearby drums may have incorrectly cued some participants that the opening between the drums was the ramp location

Results from Experiments 2 and 3 reinforce the impact of closure with comparable results between the PCB and LCD alternatives. Indeed, the LCD was designed following results from Experiment 1 regarding the impact of closure and continuity. By creating a device that could rapidly be grouped as a single unit through the principle of closure, the LCD demonstrates how the results from PCB could potentially be applied with a different device. Results showing no significant differences between PCB and LCD errors demonstrate that the benefits of closure from PCBs can be brought to work zones without the difficulty of transporting and installing heavy portable concrete barriers. However, it is important to clarify that the LCD provides only the visual cues of the PCB, it does not provide a similar physical barrier. The LCD is easily traversable and will not redirect a vehicle encroaching into the work zone. Where the physical barrier attributes of the PCB are needed than the proposed LCD will not suffice.

The elevated number of location errors in areas without solid closure can direct future research and but this finding also raises issues with existing standards. A short review of state standards and of the MUTCD suggest that states have focused on special ramp barriers in the immediate vicinity of a ramp, especially when the ramp is open. But many of the observed errors in the experiments occurred when the ramp was closed, several hundred feet from the start of the ramp treatment. These errors suggest that not only is closure an important issue, but also that a temporary ramp configuration could have an impact on driver understanding at greater times-to-exit than can be accounted for using existing delineation methods.

Continuity is the principle that objects forming a linear pattern will be perceived as a single entity (see Figure 2). In these experiments, channelizing devices in the PCB, LCD, 10 ft. and 40 ft. drum spacing alternatives could be placed in a perfect line with exactly the same spacing between each device. Only the 40 +/- 2 ft. and 10 +/- 2 ft. drum alternatives were not perfectly linear; in those alternative drums deviated by up to two feet in each direction.

The decrease in continuity for the +/- 2 ft. alternatives significantly affected the percent of correct responses in several ways. First, in the open condition, participants were much more likely to make an Identification error (i.e., to say the ramp was closed). This problem of increased Identification errors continued through most time-to-exit distances until very close to the diverge point (2s and 1s). In a driving environment, misunderstanding the state of an exit ramp, even for a short time period, could have a negative impact on safety.

This issue of continuity is important since a number of effects can result in device placement that is not perfectly continuous. Wind and gusts from traffic can shift drums as they are sitting on the road surface, construction equipment can slightly impact drums, etc. The data from this study are not sufficiently comprehensive to draw absolute conclusions, but the findings clearly imply that even a relatively small variation in channelizing device continuity may decrease the ability of drivers to immediately comprehend the condition and location of an exit ramp.

CHAPTER 4

REVISITING THE WORK ZONE CASE STUDY

In the development of the work zone experiments from Chapter 3, there was a vision for a Yes/No detection study; a straightforward study that could be explored using well-developed methods. After participants were presented with images, they would respond either by answering correctly, answering incorrectly, or not answering at all. For the work zone study "correctly" implied clicking on the ramp diverge or EXIT CLOSED, as appropriate, while "incorrect" was any other click. Repetitions of each image (ten, initially) were also used to gain an understanding of the participant response consistency as well as ensuring that any unintended participant clicks would be unlikely to influence analysis results.

In reality, the designed participant task was more complicated than a Yes/No detection, as demonstrated from the data discussed in Chapter 3. The distributions of points suggested that either the response pathway had several possible outcomes or that the participant task could be further decomposed into subtasks for analysis. This chapter will explore how the work zone experiment's design does not lend itself to extracting all of the information about how participants comprehend the traffic control. First, this chapter will present a discussion of simple experiment design and analysis of yes/no detection tasks. Second, it will present a further decomposition of how drivers interact with traffic control devices. Third, issues with the case study in the context of the traffic control interaction subtasks are discussed.

Data Do Not Fit the Pattern of a Yes/No Detection Task

When initially designed, the work zone case presented in Chapter 3 was thought to be a Yes/No detection problem. Within the framework of Signal Detection Theory, a Yes/No detection task is one where a participant is presented with a stimulus and asked to indicate whether or not a target is present. In the context of the work zone case, the stimulus was an image of a work zone, and the target was an open ramp. It was anticipated that the participants would indicate the equivalent of a Yes by clicking the ramp and the equivalent of No by clicking a button saying EXIT CLOSED.

In a Yes/No detection task, each response is classified as a True Positive, a True Negative, a False Positive, or a False Negative, corresponding with the presence of a target and the response given. These are sometimes referred to as “Hits”, “Correct Rejections”, “False Alarms”, and “Misses” to avoid overlapping parts of terms (Macmillan & Creelman, 2005). Table 5 shows the conditions for each in the work zone case when considered as a Yes/No task. Conceptually, a participant selects a criterion value for each image, and then each stimulus greater than such a value is labeled “Yes” and each stimulus less than the criterion is labeled “No”.

Table 5: Response Categories for a Yes/No Detection Task

Category	Alternate name	Case Stimulus	Participant Response
True Positive	Hit	Open Ramp	Ramp
True Negative	Correct Rejection	Closed Ramp	“Exit Closed” button
False Positive	False Alarm	Closed Ramp	Ramp
False Negative	Miss	Open Ramp	“Exit Closed” button

A common analysis for Yes/No detection tasks is to compute the d’, a metric of the ratio of the distance between the means of distributions of negative signals and positive signals and the spread of the distributions. For example, Figure 23 shows the positive and negative distributions on a scale of Z-scores representing signal strength, along with an illustrative criterion and the corresponding classification categories. Thus, in the work zone case study, the left distribution represents the strength of the ramp open signal when the ramp is closed. The “open signal” strength is generated by noise in the image (external) and noise in the participant evaluation (internal). The right distribution likewise represents the ramp open signal when the ramp is open. This distribution represents the signal strength generated by internal and external noise as in the left distributions, as well as internal and external noise related to the open ramp signal.

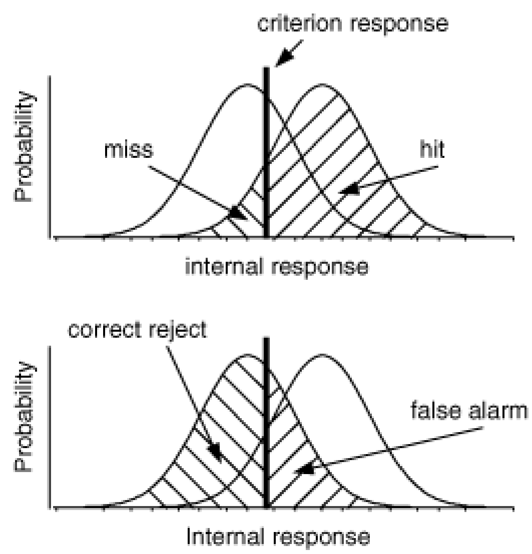

Figure 23: Basic categories for a Yes/No Detection task. All values above the criterion response are responded to as Yes, all below the criterion are responded to as "No." Reproduced from Heeger (1998)

In the Z score distributions d’ is the distance between the distribution means if the assumptions that the signal distributions are Gaussian and that the distributions have equal standard deviations are made. However, it is recognized these equal variance signal detection model assumptions are strong, and they have been criticized elsewhere (Macmillan & Creelman, 2005). However, the equal variance signal detection model is a simple "first stop" on the way to more robust models which use the same inputs. These assumptions allow for a useful illustration of the issues with the data from the work zone case.

The formulation for d' is Z(H)-Z(F) or the difference between the Z-score of the hit rate and the Z-score of the False Alarm rate (Macmillan & Creelman, 2005). However, it may be readily seen that a challenge exists in the application of this metric to the work zone case study. That is, how should responses classified as "Work Zone" be treated, i.e., responses where the participant was correct in stating the ramp was open but identified the ramp in the incorrect location. If the experimenter was blind to the locations of the responses, as they would be if the response method was discrete (e.g. a keyboard Y/N response), and assuming "Indeterminate" values could still be removed, the Hit rate (H) would be:

$$Hit = \frac{Ramp|Open + WZ|Open}{\sum Open}$$

and the False Alarm Rate (F) would be:

$$FA = \frac{Ramp|Closed + WZ|Closed}{\sum Closed}$$

As an example using Experiment 1, these formulations would yield the d' values in Table 6. Where a participant is more likely to distinguish correctly between an open ramp and a closed ramp, d' is greater. The values Table 6 fit with previous analysis and with intuition. Where the d' values are greater, a participant is more likely to correctly distinguish an open ramp from a closed ramp; the highest d’ values below are at the 1 second distance across all alternatives (where the ramp was closest to the camera in the stimulus image) and with the portable concrete barrier at all distances. Values of d' were lowest for the misaligned drum alternative D40M, which fits data in Chapter 3 that suggests drivers had a difficult time correctly identifying when D40M ramps were open or closed.

Table 6: Response Rates and d' for Work Zone Experiment 1, considering WZ responses to be "Yes" responses

Alternative	Distance	Participant Average of Hit	Participant Average of Miss	Participant Average of FA	Participant Average of CR	d'
D10A	1	0.996	0.004	0.007	0.993	5.082
	2	0.992	0.008	0.169	0.831	3.368
	3	0.997	0.003	0.205	0.795	3.623
	4	0.982	0.018	0.216	0.784	2.882
	5	0.978	0.022	0.246	0.754	2.708
D40A	1	0.999	0.001	0.026	0.974	4.957
	2	0.988	0.012	0.169	0.831	3.217
	3	0.991	0.009	0.218	0.782	3.142
	4	0.987	0.013	0.200	0.800	3.058
	5	0.971	0.029	0.244	0.756	2.583
D40M	1	0.993	0.007	0.116	0.884	3.645
	2	0.989	0.011	0.160	0.840	3.275
	3	0.963	0.037	0.174	0.826	2.728
	4	0.774	0.226	0.178	0.822	1.673
	5	0.694	0.306	0.233	0.767	1.237
PCB	1	0.998	0.002	0.007	0.993	5.418
	2	0.999	0.001	0.009	0.991	5.342
	3	0.992	0.008	0.008	0.992	4.821
	4	0.996	0.004	0.014	0.986	4.818
	5	0.993	0.007	0.003	0.997	5.199

However, there is an interpretation challenge with these results introduced by the work zone clicks (Figure 24). When the ramp is open, these responses are not accurately described as "Hits" in that they have not correctly located the ramp, and they are not "False Alarms" in the traditional sense as the participant has identified the ramp as open. There needs to be a way of analyzing the data while accounting for these responses.

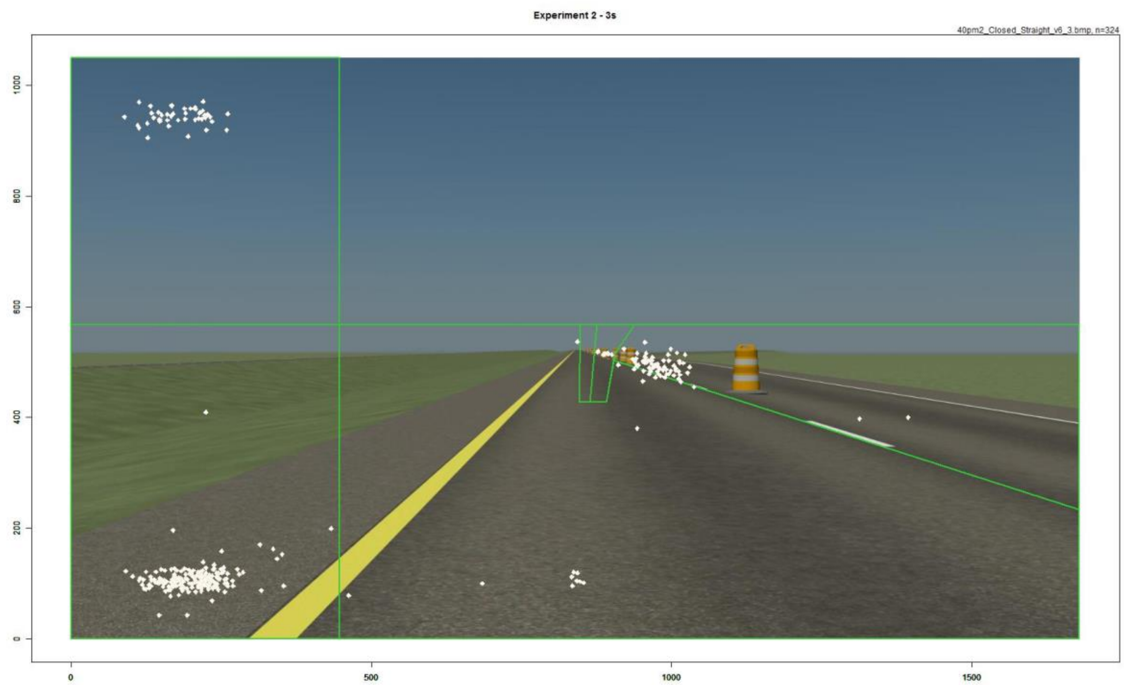

Figure 24: Responses from the D40A alternative at the 3s distance in the closed condition from Experiment 2. Many responses were not clicks on the ramp (a False Alarm in this case), but into the active work zone.

The first method is to simply remove WZ clicks from the data, and analyze d' as

$$Hit = \frac{Ramp|open}{\sum Positive}$$

and

$$FA = \frac{Ramp|closed}{\sum Negative}$$

These formulations yield the following results in Table 7:

Table 7: Response Rates and d' for Work Zone Experiment 1, eliminating WZ responses

Alternative	Distance	Participant Average of Hit	Participant Average of Miss	Participant Average of FA	Participant Average of CR	d'
D10A	1	0.004	0.007	0.993	0.996	5.082
	2	0.008	0.019	0.981	0.992	4.485
	3	0.003	0.037	0.963	0.997	4.571
	4	0.019	0.038	0.962	0.981	3.843
	5	0.022	0.053	0.947	0.978	3.627
D40A	1	0.001	0.020	0.980	0.999	5.076
	2	0.012	0.022	0.978	0.988	4.268
	3	0.009	0.038	0.962	0.991	4.134
	4	0.014	0.033	0.967	0.986	4.042
	5	0.031	0.048	0.952	0.969	3.532
D40M	1	0.007	0.050	0.950	0.993	4.095
	2	0.011	0.041	0.959	0.989	4.015
	3	0.037	0.055	0.945	0.963	3.381
	4	0.230	0.056	0.944	0.770	2.330
	5	0.316	0.080	0.920	0.684	1.884
PCB	1	0.002	0.007	0.993	0.998	5.418
	2	0.001	0.009	0.991	0.999	5.342
	3	0.008	0.008	0.992	0.992	4.821
	4	0.004	0.012	0.988	0.996	4.877
	5	0.008	0.003	0.997	0.992	5.193

In this analysis similar trends arise from earlier analysis; namely, d’ values are highest at the 1 second distance and for the PCB at all distances. However, d’ values are higher also for the drum alternatives and especially for the D40M alternative. This suggests that by throwing out work zone responses, d’ is distorted to make the signal difference seem higher where a sizeable portion of the errors were work zone errors. However, work zone errors represents potential intrusions into the work zone; mistaking an active work zone for an exit ramp is the start the chain of events leading to a motor vehicle crash. It is not reasonable to discard these data when it could be useful (perhaps the most critical data) to the ultimate goal of the study. As mentioned earlier, these are not true "Hits", but they are not "Misses" either because by clicking on the work zone they are indicating a positive detection; however, in the wrong location. These responses in the work zone could most reasonably called "False Positives"; the response, like a False Positive defined above, incorrectly indicates that the ramp is open. The "False" issue though is not that the ramp is open if it is not, but that the ramp is open where it is not. Signal detection theory does not afford us a way of addressing this in a Yes/No task because the definitions of Hit, Miss, False Alarm, and Correct Rejection depend on a comparison with the given response to the actual state of the stimulus. In this case, the work zone is never the actual state of the stimulus--it is incorrect both when the slide presented is a closed ramp and when the slide presented is an open ramp. One possible way of accounting for this response would be to consider it a "False Positive" and include these responses in the false alarm as $FA = \frac{Ramp|Closed + WZ|Closed}{\sum Closed}$.

This definition, though, introduces a more complicated problem that the false alarm ratio is set dependent--a different group of participants could not only have a different total number of work zone errors, but the number of work zone errors distorts the impact of the true False Alarms. Further, it also does not fit to simply add a new category and consider this to be a single scale categorization task with three distributions, since the action of selecting an open ramp in the work zone is nested within the detection of an open ramp and not a characteristic of the stimulus like an open or closed ramp. To move forward with analysis using signal detection theory, the three tasks of interacting with traffic control and the pathways to the four categories of Yes/No detection need to be decomposed to expose the nuances of such an experiment. The reminder of this chapter will seek to provide these insights.

Three Tasks of Drivers Interacting with Traffic Control

The participant's task may be broken into three parts for consideration. A driver must detect that a device or system of devices is present; the driver must localize the device(s) in their field of view; and a driver must identify the device (that is, extract meaning from the device by observing its entirety and assigning a label to its meaning). These three task components do not necessarily happen in sequence, with portions of these tasks occurring in parallel.

An experiment based on an operating situation should mimic that situation being presented in a way that balances the need for data, risk to the participant, and transferability of the results. In a driving task, it is important to first decompose the task into smaller subtasks. For the work zone case study, an initial assumption was made that the task of responding to a traffic control device was primarily a detection task. However, the data suggest that drivers' interactions with traffic control are more complex tasks than initially expected.

Detect

Drivers interpret meaningful messages from traffic control devices and systems, but before drivers can extract that meaning they must first detect or notice such a device. In the model posited by Feature Integration Theory, a stimulus can be detected if its features have enough contrast above some internal threshold that a person detects them. In detection the driver becomes actively aware of the existence of a device or roadway element. In this context the detection task also includes any initial search before detection. However, the detection step does not include any interpretation or identification. For instance, in the work zone example a participant must first detect the presence of drums. The meaning of the drums (i.e. the ramp is open or closed and the ramp diverge location if open) will be determined in the identification task.

Driver detection is influenced by previous experience, potentially resulting in the minimization or elimination of search from the detection task. For instance, when considering the searching process of drivers, Cole and Hughes (1990) discovered that drivers' eye movements while operating a vehicle do not follow that patterns typically associated with serial search, even when told to look for a specific target. Similarly, Chrysler et al. (2004) discovered that participants in a study with still images of road signs always detected a sign, but in a video or simulator environment, they occasionally did not detect the sign. The implication is that it is not necessarily enough to improve a sign by improving only its message. If a traffic control device does not sufficiently stimulate the driver, the later steps of localizing its position and identifying its meaning never occur because the driver does not progress to attentional search (discussed in the next section). Without detection, there is no localization or identification

Localize

Localization is the portion of search where a person attaches spatial position to a target. The process of localization occurs in parallel with detection. Much of the visual search literature suggests a two-step search process (Humphreys, 2015; A. M. Treisman & Gelade, 1980); the initial step of the search process is pre-attentive, where a person sees the entirety of the overall visual field. In this phase, a substantial portion of detection may occur, especially in detecting potential targets (Krupinski, Graham, & Weinstein, 2013; a Treisman, 1988). The second phase is attentional search. In this phase, an individual serially searches through the targets detected in the first phase, making a conscious judgment about each potential target (Krupinski et al., 2013; a Treisman, 1988). Even then, though, some of this search may be attentional but not conscious (Siegel, Han, Cohen, & Anderson, 2013). From this devices or features of interest are localized in the scene. In the work zone study the participants knew *a priori* that there was only one ramp, thus this is likely a self-terminating search.

Localization and detection may happen in parallel. In the pre-attentive phase of search, participants map out potential targets for attentive, serial search. However, under the time constraints created by traveling through a driving environment, the time for attentive, serial search may be limited to selective location pre-attentively mapped. That is, the serial search location is limited based on driver experience.

Identify

Once a driver has detected and localized a device they must recognize the device and interpret its meaning. In this text identification refers to both recognition of a device (e.g. the detected object is a construction drum) and interpreting its meaning (i.e. the ramp is closed). Thus, even after detection there it is still very possible that the message may be lost. The duty of identification is shared by the driver and designer. While the driver performs the identification task the designer's message must be both clear and unambiguous to support proper identification.

.

Difficulties in Interpretation

When a single response has to convey detection, localization, and identification, it can be difficult to extract the information on all three aspects. What is not clear in such a response is when an error occurred what was the potential source, or potentially for a correct response if there was an error or series of errors that led to an appropriate response. Each of these response pathways can lead to different issues in the data and potential to very different conclusions. In the work zone case, the single response also leads to a secondary detection task. If the participant chooses that the ramp is open they may click on the correct ramp diverge location (correct response) or within the active work zone.

Clear Path to True Positive Result

Broadly, a true positive in reaction to a traffic control device is one in which the observer sees the driver complying with the device or acknowledging its meaning as intended. It is desirable that for a true positive the participant responds in such a way that their detection, localization, and identification task are all correct. However, if each subtask is not explored individually the response may be correct but one or more of the subtasks may have been executed incorrectly. Consider the work zone case study: if a participant was presented with a ramp that was open, a true positive would primarily arise if they detected the ramp, localized the ramp, and identified the ramp as open, as their answer would be a click at the location of the ramp diverge. However, there is some chance that a slip could lead to a true positive if a participant was incorrectly localizing their response and accidentally marked their answer at the correct location (while intending to mark an incorrect location). That potential raises more issues about ambiguities in defining a “positive response” however for now this will be assumed to be a low probability event. (This issue will be further explored in Chapter 6.)

Multiple Paths to False Positive Result

One of the issues with a false positive is that the “positive response” is not clearly defined. In a forced choice detection task (one where the response choices are clearly mapped to positions in the scene), there is no room to misconstrue. In a free detection task (one where there are no defined responses and the participant can respond anywhere within the experiment space), however, ramp localization may impact the response. If a participant, for instance, marks a location that indicates that they have detected an open ramp, but localized it incorrectly, that is essentially a false positive: they detected a device that was not present. It is not clear, here, the cause of the false positive. Did the participant erroneously detect an opening at an incorrect location while perceiving that the true ramp location was closed or did they erroneously detect opening at an incorrect location while incorrectly localizing the ramp? The answer is not discernable from the single response.

Multiple Paths to False Negative Result

A negative response is one in which the participant responds saying that they do not detect the device. Again, the detect-localize and identification tasks can both have errors that lead to a false negative. Several possible paths to the error exist. For instance, if a participant’s search for the diverge location fails then, the response is either to search through the time limit or default indicate the ramp is closed. Alternatively, if the participant incorrectly detects and localizes the ramp and then correctly identifies for that location the control devices indicate closure a false negative is given. Finally, if a participant detects and localizes the ramp location but identifies it incorrectly, the response is again negative.

These separate pathways lead to the same response. The analysts cannot identify if the participant failed to find the ramp location or found the ramp location but incorrectly identified it as closed. The control devices can be redesigned for better detection or redesigned for better identification, but knowing which strategy to employ depends on the reason for error. Identifying the path to the false negatives is difficult when there is only one response. By separating the responses in each subtask, disagreement between responses can serve as a cue for a false negative.

Multiple Paths to True Negative Result

Similarly, it is not clear what pathway a participant used to reach a true negative. Again, a negative response in a detection task is indicating that the participant did not detect the target. If the task is actually more complicated though, it is not clear why they did not detect the target. The first cause might be a failure of detection. If the participant detects no ramps at all, the available answer is to indicate that the exit must be closed. Next may be an issue of localization: perhaps the participant detects a ramp but is unable to locate it. A reasonable course of action then would be to assume that the detection was in error and to indicate that the ramp is closed. Lastly, the participant may detect and localize a target that is not the ramp, but still identify that location as closed, leading to a true negative response, but with errors in the pathway.

Summary

Revisiting the data from Chapter 3, several issues arise. Specifically, there are four areas of focus: first, the experiment design is a Yes/No detection task, but the data indicate that the task of comprehending traffic control is a combination of a detection, localization, and identification (DLI) task. Second, the presence of apparent localization in the work zone case's data necessitated ad hoc categorization of responses used in Chapter 3, because traditional Yes/No detection analysis could not account for the responses where participants clicked into the active work zone. Chapter 5 will explain strategies for developing a DLI experiment with categorization as part of the experiment design. Third, there is a possibility for noise within a participant's responses, potentially leading to a disconnect between intended response and observed response. Chapter 6 will discuss strategies for accounting for that error and reducing uncertainty about intended response. Lastly, a method of modeling comprehension that accounts for variations in individual comprehension skill is presented. Chapter 7 presents an application of the Linear Logistic Test Model as a method for describing comprehension of traffic control as a function of individual characteristics. Chapter 8 summarizes the contributions of this work: analysis of the task, the refining of a DLI experiment with categorization, accounting for response error, and application of the Linear Logistic Test Model in analysis of response. These chapters can be read separately, although understanding the context for the conclusions drawn in each can be helpful.

CHAPTER 5

DESIGNING A DETECT, LOCALIZE, AND IDENTIFY EXPERIMENT

The previous chapter demonstrated the major issues with previous experimental design, specifically that the detect, localize, and identify subtasks of traffic control comprehension were convoluted into a single response. The underlying logic for selecting still images as the medium to conduct a comprehension experiment remains: other methods are too risky or resource intensive. Field testing novel treatments may offer risk to drivers while revealing only crashes, not errors; instrumented vehicle studies require expensive equipment and again drivers are potentially exposed to increased risk with novel treatments; simulator studies are similarly expensive to build virtual environments, can be time consuming to a recruit a sufficient sample (which must each be trained on the simulator and run one-at-a-time). Further, each of these methods generates overwhelming amounts of data that can be difficult to interpret. While a still image experiment may not seem on face to be as sophisticated as other methods, using a still image method addresses the questions of interest with quickly analyzed response data, and can collect many participants' data using a lab of personal computers with mice and keyboards, which are readily and cheaply available equipment requiring almost no training for participants. Thus, this chapter will explore the potential to redesign the experiment using the still image medium; however, adapting the design elements to account for the detect, localize, and identify subtasks.

As discussed in Chapter 4 when considering the work zone study, the expectation was that the ability to correctly identify the ramp condition would fit a Yes/No task experimental design. Yes/No detection may be used to allow participants to indicate whether they perceive a stimulus is present or not present. Yes/No detection is commonly applied to search, noticeability, and attention tests. However, as seen in Chapter 4, the experiment design did not cleanly fit into this yes/no paradigm. For instance, when a participant correctly identified the ramp as open but clicked in the work zone in a Yes/No experiment as described in Chapter 4 that would be a true positive, even though it is an error in the identification of the ramp location.

In analyzing this information it became clear that the Yes/No assumption did not fit the data: this was instead three subtasks - detection, localization, and identification. While participants had a single objective: “Did you detect an open ramp?” they had multiple options for localizing their response. In such a task errors may result from any of the pathways outlined in Chapter 4. The source of the error may also differ depending on the ramp condition and devices being tested. For instance, participants were more likely to say a ramp was open when it was closed than vice versa, suggesting that some process was skewed toward an open ramp. This indicates that the task was more intricate than a simple yes/no detection or even a multi-category detection, but is actually three subtasks: detection, localization, and identification.

This chapter will propose a new design of a still image experiment that separates out the Detect, Localize, and Identify subtasks by asking separate questions which yield dichotomous responses. It will be demonstrated how these responses untangle the pathways convoluted by a single response variable which were identified in Chapter 4.

Next, this chapter will demonstrate a process for improving the categorization of localization responses through pilot testing, specifically using think-aloud and response-review techniques. Lastly, this chapter will describe a process for selecting a categorization method which reduces systemic error using hypothetical alternatives as an example.

A New Methodology

The issues of ambiguous causes behind responses make analysis and interpretation of results in the work zone case difficult. A redesigned experiment may make analysis clearer by separating out the subtasks and building an experiment around each. The initial redesign does not require significant changes to the structure of the work zone experiment or to similar studies. The redesigned experiment uses a similar procedure and stimuli, but make changes to the question structure and potential cueing from the stimuli presented.

Separate Tasks

As previously stated, if a traffic control comprehension experiment is structured as a yes/no detection task then there are several pathways to different responses, each with their own underlying cause. The first step in a redesigned traffic control experiment is to untangle these pathways. Rather than a response being the product of several smaller responses, that data is a tuple with each item a separate subtask.

The first subtask is to detect-and-localize. Because these functions happen pre-attentively and attentively, and are considered to occur in parallel (Wolfe & Van Wert, 2010), they can be combined and be considered a single data point. Where the case study asked participants to “click on the ramp if it is open, and click the exit closed sign if it is closed” and all images had a ramp, a proper detect-localize-identify task should begin with images that may or may not have the traffic control system of interest at all, and the ramp may or may not be open. Participants should first be asked (in the case of the case study), “If you see a ramp, mark the location of the ramp. If you do not see a ramp, mark ‘no ramp’.” This separates out the detect-and-localize task from the identification of the device, which is presented separately in the second part of the experiment. The outcome of this subtask is a determination of whether the traffic control devices cue drivers to correct or incorrect ramp locations.

In a second part of the experiment, participants would be shown the same set of images with only the task of identification. A point on the image would be highlighted and participants would be given the question directly; in the case of the work zone experiment, that question would be “Can you drive here?” though it should be tailored to the specific traffic control in question. The highlighted points would include what the experimenter had previously identified as locations where detection and localization are both correct and incorrect to see how participants respond to an identification question with the full range of possibilities from the detect-and-localize experiment.

The two step experimental method outlined here has a distinct advantage over the experiment from the work zone case in that each of the combinations of detect, localize, and identify may be distinguished from the question responses. Being able to separate out problems with detection, localization, and identification each have implications for the designer. Knowing that a device is difficult to detect could simply lead to the designer making the system more salient. For difficult to localize systems, a designer may consider a system with less spatial ambiguity through the use of arrows or stronger lines. For a difficult to identify system, a designer may consider altering or clarifying the message by reducing the demonstrated ambiguity. Each of these cases has different design responses, but knowing where the difficulty lies is key to a proper design change.

Untangling the Response Pathways

While the old pathway to a Hit or open ramp True Positive is the same (correctly detect, localize, and identify a ramp when it is open) some of the other mixed pathways are separated.

Former False Positives

In the work zone case, there were two paths to a false positive, where a participant indicated that a ramp was open when it was closed. First, a participant could have incorrectly detected that a ramp was open, localized the response at the ramp location, and identified the ramp location as open. Previously, this would only be classified as an incorrect response. In this new method, this would be classified a correct detection (i.e. there is a ramp present in the image), a correct localization (the participant correctly located the ramp), and an incorrect identification (the participant incorrectly identified the traffic control as permitting crossing at the localized point).

Second, the participant could have incorrectly detected that a ramp was open, localized the response at the work zone location, and identified the work zone location as open. This would yield a correct detection, an incorrect localization, and an incorrect identification. In this experiment, the false positive pathways are now identified as separate combinations. As a result, the signal strength for detection, localization, and identification can be identified separately.

Former False Negatives

In the work zone case, all false negatives were recorded as "Exit Closed", although the reason for that response was lost. In this new methodology, each error path can be followed. The first path is that a participant did not detect or localize the ramp, terminating search. In this experiment, this pathway would register as an incorrect to detection and an incorrect to localization, but the identification question would still yield an answer when presented with proper cueing in the second experiment. This would allow designers to see that a redesign of the device needs should focus on improving detection along with any appropriate changes to identification.

The second path is that a participant detected a ramp, but localized at an incorrect position, and then identified that position as a closed ramp. In the initial experiment, a designer would not know if the problem with the traffic control was a lack of detection, localization, or identification. Separating these subtasks registers correct detection, an incorrect localization, and correct identification (through the second experiment where the indicated region would be the work zone, where the traffic control does not permit crossing) showing that designers need to address spatial position issues.

The third path is that the participant detected a ramp and localized the response correctly, but incorrectly identified the ramp as closed. In the redesigned experiment, this would register as a correct detection, a correct to localization, and an incorrect identification. This is much more informative to a designer, who can be sure that the device is visible but not understood.

Former True Negative

The phrase “true” implies that a true negative response was correct; it is true that the response was correct, but the reason is lost on the experimenter without the three subtask questions. The intended objective of this True Negative in the work zone experiment was that a participant detected a ramp, localized the ramp, and then correctly identified it as closed. However, two other paths exist. The participant could have not detected a ramp at all, terminating search and defaulting to “Exit Closed” in the lack of other response options. In the new experiment, failed detection would register as an incorrect detection and incorrect localization, showing designers that an issue is the saliency of the traffic control. The identification question would be addressed as part of the second portion of the experiment allowing for an evaluation of the traffic control closure when the localization is given.

Also, the participant could have detected a ramp, but localized the incorrect location, identifying that location as closed. This would register as a correct detection, an incorrect localization, and a correct identification. While there is likely no harmful impact of these responses in the work zone case, it is still useful to see the process behind a participant’s decision for cases where designers need to be sure that a true negative is recorded as a true negative, such as a “Do Not Enter” sign or traffic control indicating a lane closure.

New Methodology Should Remove Cueing to Measure Detection

Cueing is inherent to any driving task because traffic control devices are only in a few spaces around the road. Vehicles will only be present on the pavement or shoulders, signs will only be mounted high and to the left of, right of, or above the road, etc.

However, cueing in the experiment, caused by the stimulus always being in a small portion of the screen, can lead to a diminished pre-attentive activation map for attentive serial search (Hout, Walenchok, Goldinger, & Wolfe, 2015; Schwark, MacDonald, Sandry, & Dolgov, 2013). Every effort should be made to reduce this cueing in the detection and localization portion of the experiment, since it is a result of the experiment and not a result of the nature of driving. Cueing in the experiment should be reduced by not having a ramp in all images, and by changing the location of the ramp in within the images. This can be performed by changing the vantage point and orientation of the camera, so the ramp appears in physically different spaces on the screen.

Collect Data

In the work zone case presented in Chapter 3, the data collection was performed as a visual-manual search task where a participant would indicate their response on the screen using a standard mouse as a pointing device. Because computer mice are widely available and require little training, this method works well for collecting the response coordinates on the stimulus (although, as will be discussed in Chapter 6, it is useful to record the path and speed of mouse movements as well), but the (x,y) coordinates do not themselves indicate the participant's intended response. The next section shows how to extract information from a participant's coordinate response through categorization.

Categorizing Responses

It is possible to only consider the point cloud distributions, latency, and qualitative measures to determine better design and comprehension, but categorizing responses aids in analysis and closely aligns with the way errors are described and modeled. Errors can certainly have descriptors of time, intensity, seriousness, cost, etc. but if they don't trigger a threshold as being an "error" than they are not recorded as errors. Errors are considered to be discrete events that either happened or did not happen, and with descriptors then assigned to the event. Categorization of spatial response follows this paradigm. A response is considered to fit in a category if it exceeds some threshold for classification.

Continuous Variables in Error Chains Cross Thresholds

Actors can approach the limits of acceptable behavior (getting close to the lane lines, for instance), but once they have exceeded some threshold it is considered an error. In many error chains, a sum or product of separate variables leads to an error. For instance, in aviation, a stall occurs when the angle of attack is too high. The angle of attack is a complex function of several issues, including airspeed, nose angle, wind, and many others. But when any combination of the inputs exceed a threshold, the stall occurs.

Considering the tasks associated with comprehending traffic control comprehension may be a combination of the binary conditions described in Chapter 4. Traffic control comprehension is a discrete task comprised of detection, localization, and identification; either the process yields a correct or incorrect response, though the intensity of the failure that results can vary. Often there is opportunity for correction, and many drivers reassess and correct mistaken comprehension as they traverse the road, the underlying selection rests in a state of true or false. It is therefore reasonable to believe that these discrete categories are appropriate in data reduction as well. Future research may well indicate that there is some sort of transition zone between detecting and not detecting, localizing and not localizing, and identifying and not identifying, but the method presented can adapt to that too. A categorization system assigns a single point to a response falling into its classification; the method can be extended either breaking the binary condition into more levels or by breaking the categories into subcategories to account for any found transition.

Developing a Categorization Method

Several methods exist for developing a categorization system, including clustering pilot data, dividing the stimulus into equal sized cells, developing target areas using heuristics such as Fitts' law (which relates action time to target width and distance to target (Zhai, Kong, & Ren, 2004)), and using features of the stimulus image to define target areas. Without some sort of algorithm for automatically identifying the "correct" target areas *a priori*, the experimenter is left to use an iterative pilot testing process to record data, check the classification system, and refine with further pilot testing data. Applying similar, well developed methods from user interface design can give the experimenter a strong sense of what participants' responses mean before intense data collection begins (ISO, 2010; Nielson, 1993; Sharp et al., 2011). All of these methods, though, begin with collecting data and observations in pilot tests.

Pilot Testing

Categorization or coding is a process that should start before the experiment begins. Coding responses further is best completed within the context of actual participant responses. Approximating the intentions of a participant may best be deduced using a think-aloud protocol where participants can provide feedback and context for their responses (Boren & Ramey, 2000; Nielson, 1993). Intentions may also be deduced by using a protocol which reviews individual responses, so that participants can explain any incorrect responses. These experiments may be completed during the experiment, but they would be better placed prior to full scale deployment of the experiment as an experiment that is an IRB-approved pilot, because the data are intended to inform the primary experimental design and analysis. These steps should be iterative and include a diverse group of participants.

Think-Aloud Method

In a think aloud protocol, participants are encouraged to speak through their selections (Boren & Ramey, 2000; Nielson, 1993). Each participant would be presented with the slides as developed, but without the time limit in place. Instructions would ask them to talk through the entire process of selecting an answer, taking their time and verbalizing everything they observe and every step of their process. A proctor records their audio and screen recording their mouse movements, potentially also recording eye gaze. In doing so, experimenters can see a representation of the conscious thought process of search and selecting specific locations, which is useful for determining intent when developing generalizable categories.

There are limitations to this process. First, the act of verbalizing the responses changes the process of search so that it is different from the silent, non-verbalized, time-limited process that individuals would use in a driving situation. Second, the process that a driver is explaining verbally may not adequately give enough detail as to the subtle, non-conscious process underlying search. Further, the responses of participants may not be reliable; it may be such that while they say something to the effect of, "I'm looking over the right side of the screen" their eye gaze is darting across the entire screen.

Response Review Method

In future experiments, response review protocols are another way to pilot test the participant responses. In this protocol, a participant would move through the slides as they would be presented during the actual experiment with the same instructions and time limits. After completing a shorter version of the experiment, the participant will then be given a structured interview with questions about each of their responses in combination with a Retrospective Think Aloud protocol (Nielson, 1993; van den Haak, De Jong, & Jan Schellens, 2003). This includes asking questions both about the image itself (e.g. "How difficult did you find this image?") to questions about the responses, (e.g. "How sure are you of your response?", "Do you think you made the right selection?"). These questions help develop a sense of the participant thought process during the experiment.

There are obvious limitations to this process too. First is participant fatigue. Participants will be asked the same questions for dozens of images, which means that their response quality may decline over time. Second, there is an issue of reliable memories. Participants will be required to draw from their memory of an image they saw for only a few seconds. Participants may be providing insight into how they thought they answered rather than through how they actually answered. Regardless, these processes can still be useful in determining response categories, even if the reliability is suspect.

Select a Categorization System

Selecting which classification method to use is not straightforward. While each point can be assigned a meaning, that's an assigned meaning and not necessarily the intended response of a participant. Approximating the intended response of a participant can only be established through stated meaning, and as mentioned earlier even that is not fully reliable. To begin selecting a categorization system, researchers should first maximize the total information from the data. Second, we should minimize systemic error. Mathematically, it is makes sense that the points we classify as "indeterminate" should be statistically similar across all alternatives if they are truly the result of non-systemic error. Any model formulation that accounts for variations within the human responses will account for random error with a separate term, but if those "unknown" responses tend to bias one alternative or another, that's a sign that the model formulation is not accounting for some of the trends.

If the pilot testing were conducted, the designer could use the information gained from a think-aloud or response-review protocol to set the category boundaries using a validated clustering analysis or other method. Without the benefit of pilot testing to determine a categorization strategy the results of the work zone case can only be analyzed post hoc. Thus, as a hypothetical exercise to demonstrate how a system might be selected, consider selection of a categorization system through the comparison of three hypothetical systems: a “Generous” alternative with a wide “Ramp” target, a “Conservative” alternative with a narrow “Ramp” target, and an “Intermediate” alternative with polar sectors similar to those used in the work zone case categorization. Figure 25, Figure 26, and Figure 27 illustrate these hypothetical alternatives.

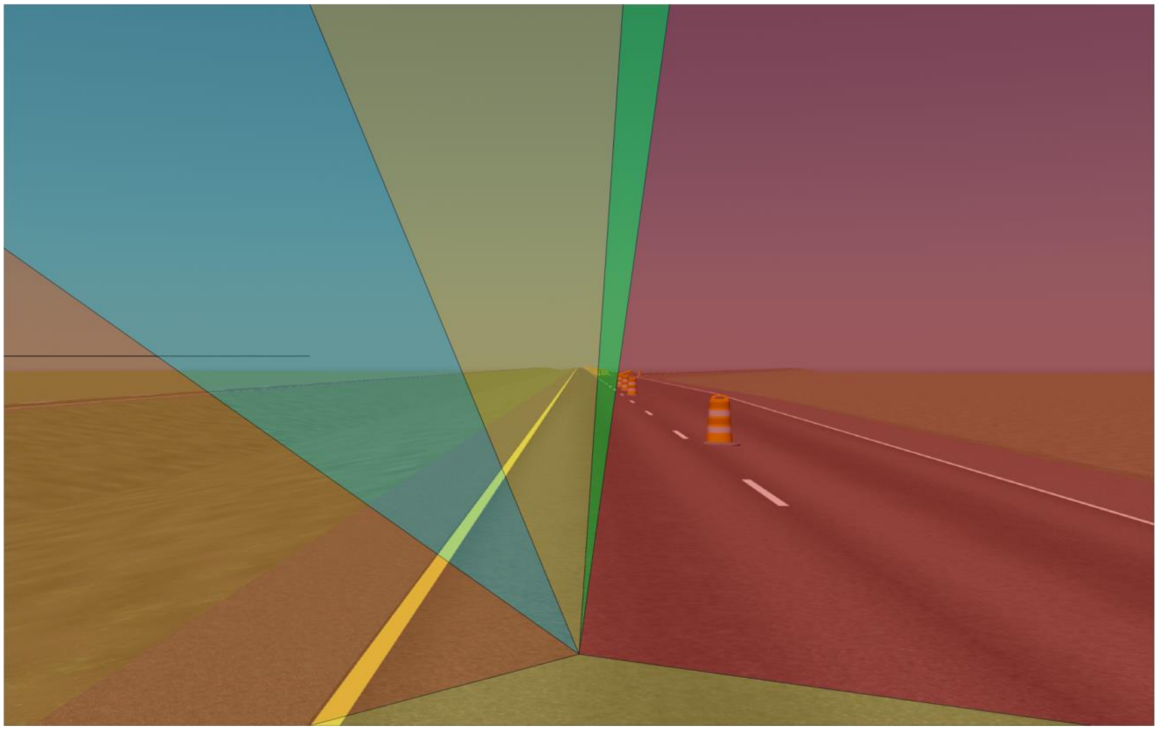

Figure 25: Example of "Intermediate" Categorization (3 second distance, straight geometry, open condition)

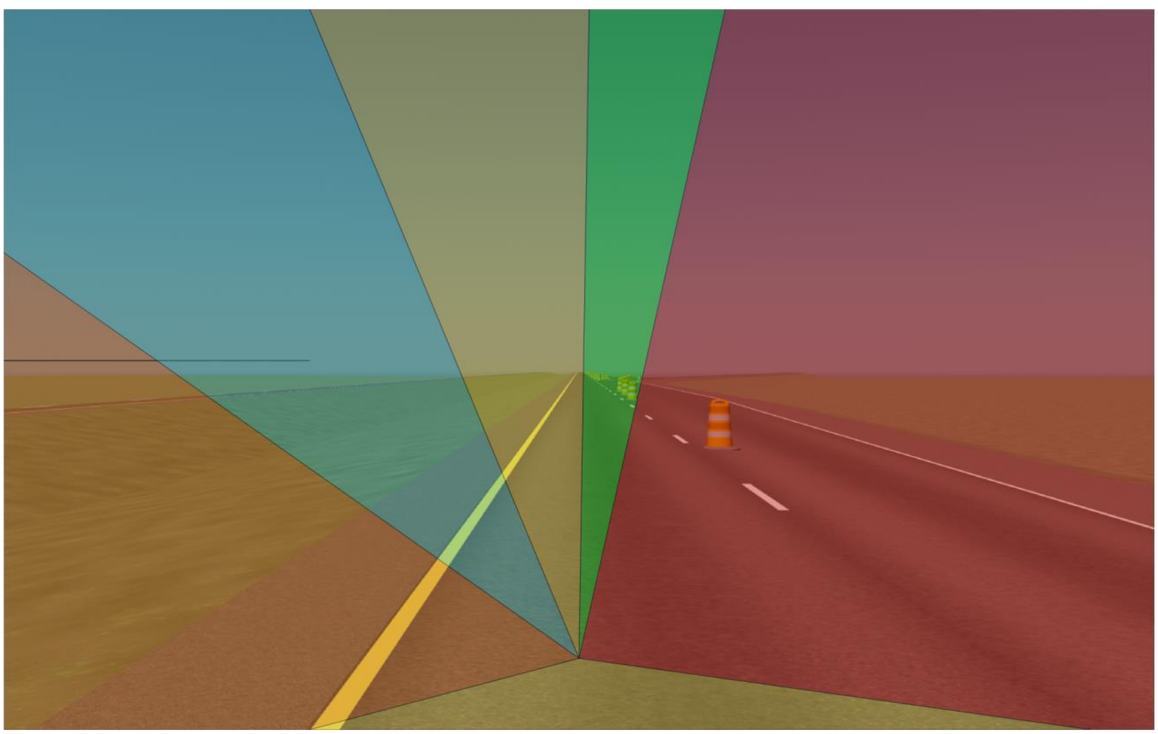

Figure 26: Example of "Generous" Categorization (3 second distance, straight geometry, open condition)

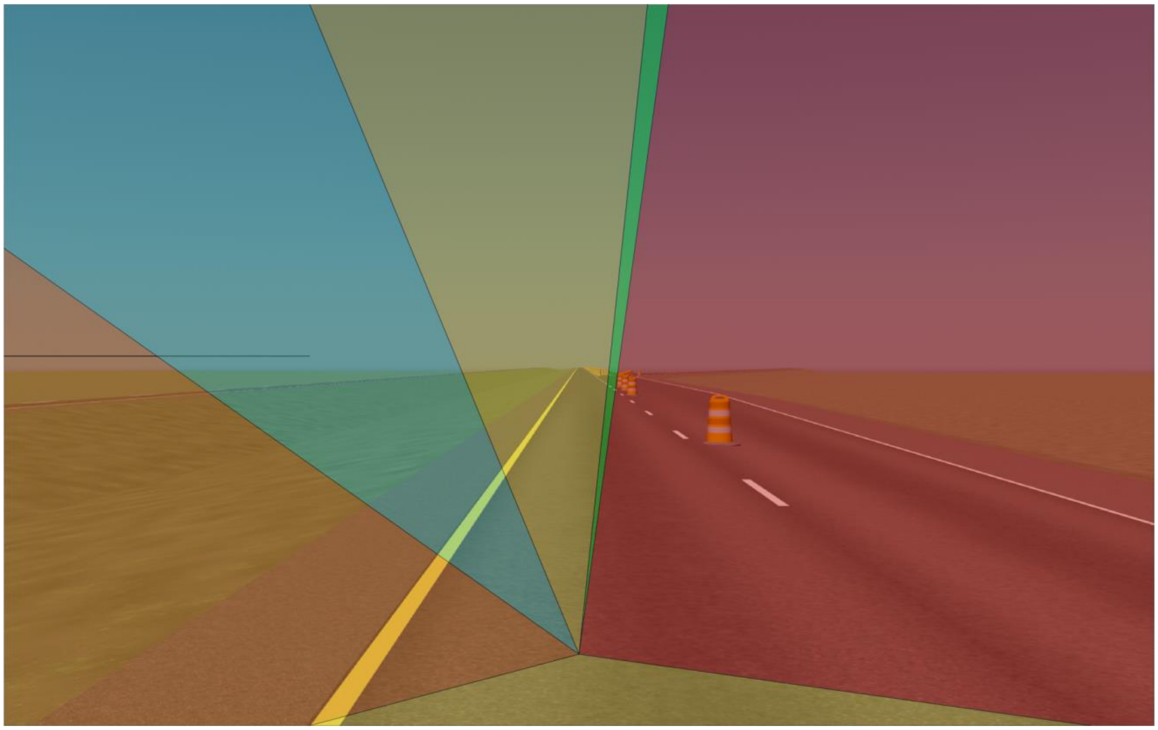

Figure 27: Example of "Conservative" Categorization (3 second distance, straight geometry, open condition)

The zoning strategy for the polar formulations was similar to the original formulation. Data points were first screened as "indeterminate" if the radius from the start point was less than 300 pixels. This placed such points squarely in the pavement and not yet to either edge line, suggesting that they were "misfires." The sectors for "Closed" and "Don't Know" were selected so that points previously classified as such would not change classifications, and that the values would total the same, except where excluded through the previously mentioned 300 pixel rule. The right edge of the work zone covered the right corner of the screen in all alternatives. From there, there were variations.

For the "Generous" formulation, the zone for a ramp was larger. The left edge of the zone was always fixed at 90 degrees, implying that any movement to the right side of the screen was intending to indicate that the ramp was open (though perhaps in the work zone). The right edge of the ramp zone (always concurrent with the left edge of the work zone) was fixed as the outside base of the first right side 40 ft. aligned drum which was not occluded by another drum. That standard was fixed across alternatives.

For the "Conservative" formulation, the zone for a ramp was much smaller. The sector was always defined as the part of the ramp defined by the *last* D40A drums at the end of the ramp. The "Intermediate" formulation sometimes overlapped with these, but considered the center of the *first* D40A drum on the right start-of-taper to be the right edge of the sector. The left edge of the sector was considered to be the right edge of the first D40A drum on the mainline after the ramp opening.

Compare Systematic Error

When comparing the systemic error in the system, experimenters should strive for an equal proportion across alternatives of responses which were categorized as indeterminate (or unclassified--essentially labeled as noise). To evaluate how homogenous the indeterminate responses are between alternatives, a chi-square test of homogeneity was performed. The results of this test on each condition in Experiment 1 can be seen in Table 8 and Table 9; that the conservative formulation does not succeed in this test, especially in the Open condition. The Generous formulation does reclassify a substantial number of the indeterminate responses because it has the lowest total number of indeterminate responses across all distances and conditions, and they tend to balance across alternatives with no significant p-values in the tests of homogeneity. The intermediate formulation also does this fairly well, although it does fail the test of homogeneity twice and has higher total numbers of indeterminate responses. Ultimately, it is reasonable to select the generous formulation because it gives evidence that unclassified responses do not vary with alternative. This is especially true in the "no work" condition, where the target to click is especially ambiguous. Use of pilot testing methods proposed above can greatly reduce this ambiguity by giving researchers insight into participants' understanding of the area around the target which should yield a "correct" classification.

Table 8: Experiment 1 Tests of Indeterminate Response Homogeneity for hypothetical Generous, Intermediate, and Conservative Categorization Methods in the Closed Condition

Closed, 1 second

Alternative	Generous	Intermediate	Conservative
D10A	1	16	17
D40A	0	4	6
D40M	1	4	6
PCB	1	3	4
χ^2 of Homogeneity	0.801	0.001***	0.005***

Closed, 2 second

Alternative	Generous	Intermediate	Conservative
D10A	0	7	7
D40A	1	3	3
D40M	1	3	5
PCB	0	3	5
χ^2 of Homogeneity	0.572	0.392	0.659

Closed, 3 second

Alternative	Generous	Intermediate	Conservative
D10A	2	2	2
D40A	0	1	3
D40M	2	2	5
PCB	1	2	2
χ^2 of Homogeneity	0.532	0.934	0.572

Closed, 4 second

Alternative	Generous	Intermediate	Conservative
D10A	0	3	3
D40A	1	1	3
D40M	1	2	6
PCB	0	1	3
χ2 of Homogeneity	0.572	0.666	0.615

Closed, 5 second

Alternative	Generous	Intermediate	Conservative
D10A	3	3	7
D40A	3	5	5
D40M	2	4	6
PCB	1	1	2
χ^2 of Homogeneity	0.748	0.442	0.423

Significance indicated: $p<0.05$(*), $p<0.01$(**), $p<0.001$(***)

Table 9: Experiment 1 Tests of Indeterminate Response Homogeneity for hypothetical Generous, Intermediate, and Conservative Categorization Methods in the Open Condition

Open, 1 second

Alternative	Generous	Intermediate	Conservative
D10A	5	13	23
D40A	2	5	18
D40M	5	13	27
No Work	4	16	27
PCB	5	10	18
χ^2 of Homogeneity	0.682	0.121	0.299

Open, 2 second

Alternative	Generous	Intermediate	Conservative
D10A	5	15	72
D40A	7	15	51
D40M	8	24	168
No Work	8	17	30
PCB	2	13	72
χ^2 of Homogeneity	0.277	0.269	< 0.001***

Open, 3 second

Alternative	Generous	Intermediate	Conservative
D10A	4	17	195
D40A	8	27	175
D40M	7	35	256
No Work	7	19	32
PCB	5	26	206
χ^2 of Homogeneity	0.857	0.108	< 0.001***

Open, 4 second

Alternative	Generous	Intermediate	Conservative
D10A	8	34	253
D40A	5	32	243
D40M	4	30	142
No Work	9	24	35
PCB	7	40	283
χ^2 of Homogeneity	0.511	0.198	< 0.001***

Open, 5 second

Alternative	Generous	Intermediate	Conservative
D10A	9	40	272
D40A	4	29	261
D40M	3	18	94
No Work	7	16	33
PCB	*7*	*61*	309
χ^2 of Homogeneity	*0.495*	*< 0.001****	< 0.001***

Significance indicated: p<0.05(*), p<0.01(**), p<0.001(***)

Consider a Different Input Method

Categorization of responses provided by a pointing device can be difficult when the zones are expressed explicitly, but an extra layer of complexity is added when the boundaries of response category zones are available to the participants. In the work zone case, target zones were developed *post hoc* using insights from initial experiments, so there was no way to broadcast the zone barriers to the past participants even if the experimenters wished to change the protocol.

An alternate method for recording clear, discrete categorical responses would be to use an input method that had clear, discrete categorical responses. By flashing response categories on the screen, participants could be prompted to select from categories using a keyboard or other categorical input device, rather than depending on a pointing device, with its inherent issues arising from various sized targets. This method would not have even eliminated response latency, because that would still be recorded. In fact, the distance to a discrete selection would be the same, so the latencies between responses would be more comparable.

Participants could also assign a measure certainty to their response. Initially, latency was considered as a way of determining if a participant was guessing or answering with certainty, but this measure proved inconclusive, as the distributions of latency were similar for all responses. A "Don't Know" option similarly proved ineffective as participants were very unlikely to use the option, instead selecting an answer in lieu of "admitting defeat." Another way to have participants indicate their certainty this is through self-report. Each answer essentially comes with two data points: the actual response and a rating by each participant of their level of certainty with the response they have just given, although this measure would be only internally consistent and thus would require within-participant analysis or normalization.

Conclusions

This chapter offers to future researchers a method of testing comprehension with readily available personal computers, and refines the methods used in Chapter 3 to account for issues of measuring a complicated task that were described in Chapter 4. This chapter built on the experimental issues elucidated in Chapter 4 by developing a new methodology that separates out the detection, localization, and identification tasks for clear, defined analysis. Next, this chapter showed a system for categorizing responses. Categorization systems should be constructed using pilot test data, then selected in a way that minimizes uncategorized responses and also minimizes systemic error. Lastly, this chapter showed alternate methods for collecting data to eliminate the need for categorization. With data categorized using a noise-free system, the experimenter could move on to modeling the data using the method in Chapter 7. However, with a physical response system a participant may add noise to their responses, making separating the participant's intended response with the participant's actual response more difficult. The next chapter will show how to account for participant's error.

CHAPTER 6

PHYSICAL ERROR

Just as the process for comprehending traffic control has opportunity for error, so too does the process for indicating comprehension. When researchers analyze data, they see only the participants' responses. From these researchers must infer the participants' comprehension. Those responses, though, may not convey the necessary information to infer a participant's comprehension. Specifically, the response may not match a participant's intent and thus would not represent that participant's comprehension. These cases are a result of physical error in the response process of an experiment.

Uncertainty Lies in Any Participant

The instructions commonly given to participants in response experiments is to "answer as quickly and accurately as possible." How participants interpret that statement and the degree to which they can answer quickly and accurately varies across the participants. Each participant's priority for speed or priority for accuracy will vary, as will their skill and ability to learn to be both precise and accurate in their answers. In an effort to compare participants' responses to a cognitive task, an experimenter should make every effort to gauge the individual patterns of physical response so the participants' responses can be normalized for comparison.

Broadly, this performance can be considered in two umbrella categories which are difficult to separate in practice. First, there is a participant's dexterity or skill at the task. Some participants will simply be more able to exhibit precision or speed than others. Second, there is a participant's value decision regarding accuracy and speed. Some participants may value accuracy more than speed, while others may value speed more than accuracy; these values too may be fluid and changing over the course of the experiment.

Skill

While research suggests that for some tasks even children can in some cases successfully use a mouse with the same dexterity as adults, skill level at using a pointing device varies among people, with variations not only specific to age and physical development (Donker & Reitsma, 2007; Lambert & Bard, 2005). Hand-to-eye coordination related to a specific task may improve for participants as they gain experience during an experiment. However, across participants this will occur at different rates with different skill levels achieved. While skill at using a pointing device sets the baseline for a participant's performance, their response accuracy and precision are also driven by the Speed-Accuracy Tradeoff.

Speed-Accuracy Tradeoff

A significant factor influencing the speed-accuracy tradeoff (which would be better labeled the speed-accuracy-precision tradeoff, as accuracy deals with the central tendency and precision deals with spread and these terms are conflated in discussions of the speed-accuracy tradeoff; this work will continue to use the term, as the literature uses "Speed-Accuracy tradeoff" (Heitz, 2014)), is that in spite of any instructions given, how a participant decides this trade is a value decision by the participant (Zhai et al., 2004). Some participants value accuracy and precision to the extent that they will slow themselves down for the sake of improved precision. Other participants will be as speedy as possible minimizing their precision concern. The common instructions (and the instructions used in these experiments) ask participants to answer "as quickly and accurately as possible" but when speed and accuracy are in conflict, participants will resolve this tradeoff in a manner they believe best represents the instructions.

Accounting for Noise

The advantages to using a computer mouse as an input device are that the equipment (a computer with a mouse) is widely available and participants need little if any training on how to use a mouse. Mice have been shown to be comparably accurate and quick to use as most other pointing devices (Murata & Bullinger, 1991), which explains their near ubiquity. Mice are an obvious choice for use in a study requiring responses on a screen.

Each participant's mouse click location gives their stated response, but there is a chance that they intended to click somewhere else near to that location and could not due to a lack of skill or because of the Speed-Accuracy Tradeoff. By having participants perform a targeting task with known targets at regular intervals throughout an experiment, researchers can fit a probability density function to each participant's clicks that is calibrated to that person's own response pattern. Calibration should take place before, in the middle of, and after the general experiment. Subjects should be instructed to click on point targets scattered around the screen to obtain an estimate of their point target accuracy. Next, participants should click targets of varying area to observe their selection pattern with area targets.

These experiments allow researchers to model the probability density function ellipse around any particular participant click. Each characteristic (size of target, shape of target, time allotted, and location on the screen) can each contribute to a model of the probability density distribution around each point. This model assumes that physical error is independent of the test item, which may not be a valid assumption. This may especially be true if the test target does not offer substantial contrast from the background on which it is displayed. A "blending" of the target and background may lead to difficulty localizing the target border, which could lead participants to either have wider response clouds due to perceiving the target as larger than the researchers have defined when creating categorization zones, or have smaller response clouds due to conservatively perceiving only the clearly delineated portion of the target. These differences need to be considered carefully by researchers testing the identification or selection of ambiguous targets.

Generating a user's accuracy profile requires two steps. First, the participant must make repeated clicks at or around clear targets at various parts of the screen. Both the latency and the point clouds will be recorded to generate probability ellipses with probabilities of selection and of the axis of orientation; if paths are collected, these axes will be oriented to the participant's mouse direction at the point of click. Participants must also click on ambiguous targets, where an entire area would be acceptable, to see if there are any skews to area-drive targets. An effective test for this is the ISO 9241-9 standard test for determining the precision of a pointing device (Soukoreff & MacKenzie, 2004). See Figure 28 for an example of the test pattern.

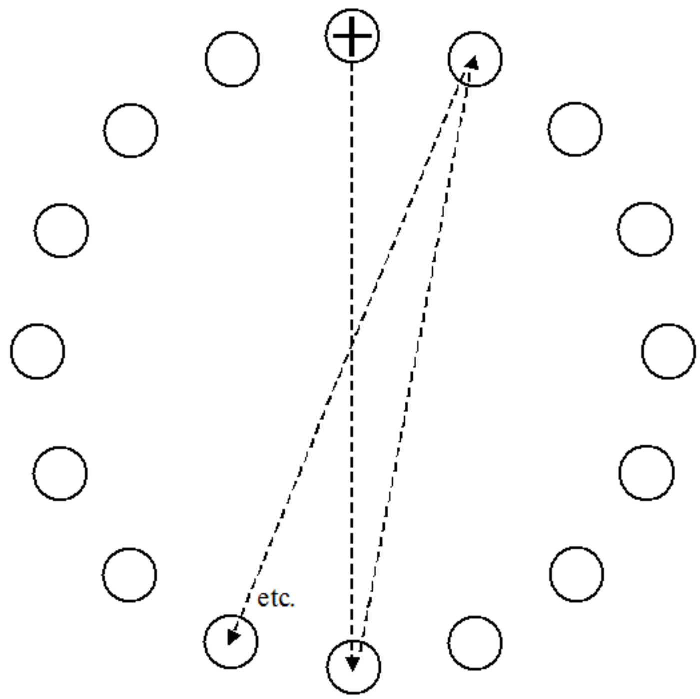

Figure 28: ISO 9241-9 Test Pattern can be used to calibrate participants' performance at using a pointing device.

By presenting each participant with a known target and shape, the comprehension aspect of a test is eliminated. Participants are able immediately to understand where they are supposed to click, and are only constrained by their physical actions. Cycling random shapes of various sizes and locations repeatedly allows the researchers to identify the distributions associated with each shape/size/location configuration, helping them build a model of a participant's shapes. To see how the speed-accuracy tradeoff changes over time or if there are practice or learning effects, a calibration exercise should be performed at regular intervals throughout the experiment. These will inform distributions around those times.

Record Paths

Another way to account for the changes of a participant's pointing device behavior over time is to record more than just the click location and time. MacKenzie (MacKenzie, Kauppinen, & Silfverberg, 2001) recommends recording paths as a way of diagnosing intent and skill. These paths can offer insight into how a participant navigated through the stimulus space while marking their response. The path can show the search process, show issues with the travel and homing phase of the search, and provide the speed and direction of the click at the point when a participant made their mark.

Anecdotally, experiment administrators in the work zone study observed participants consistently moving their pointer up to the location where the diverge would be located, and, then when observing the diverge closed (or not finding the diverge) moving the pointer down to the EXIT CLOSED. Participants appear to conflate the visual and manual aspects of the search task into a combined action. So recording the paths of participants could not only provide the velocity of the mouse at click, but also may provide insight into the search process.

Speed and Direction at Response

When generating an uncertainty ellipse around a participants' response, that ellipse has several parameters that need to be estimated. Having speed and direction at response means being able to fit the probability density function of an elliptical bivariate normal distribution that is both directional and has a calculated axis lengths. Recording the path of the response does not provide this information alone, but recording the path along with the velocity at click allows for these calculations. Further, recording the path could indicate that participants were instead of traveling along a line, travelling along an arc. That would suggest that an elliptical distribution is not necessarily appropriate unless its axis continues along that arc.

Example of Accounting for Physical Error

To illustrate how a researcher may account for physical error in pointing device response tasks, consider an example on the data of three participants from Experiment 1. First, a sample of each participant's data was taken that could be used for calibration of each person specific parameter. While earlier a method using ISO 9241-9 test was outlined, that task was not performed in the work zone case experiments, so as a substitute a "best case" sample was used: the responses to the Portable Concrete Barrier in the open condition of at the three second distance in the straight geometry configuration.

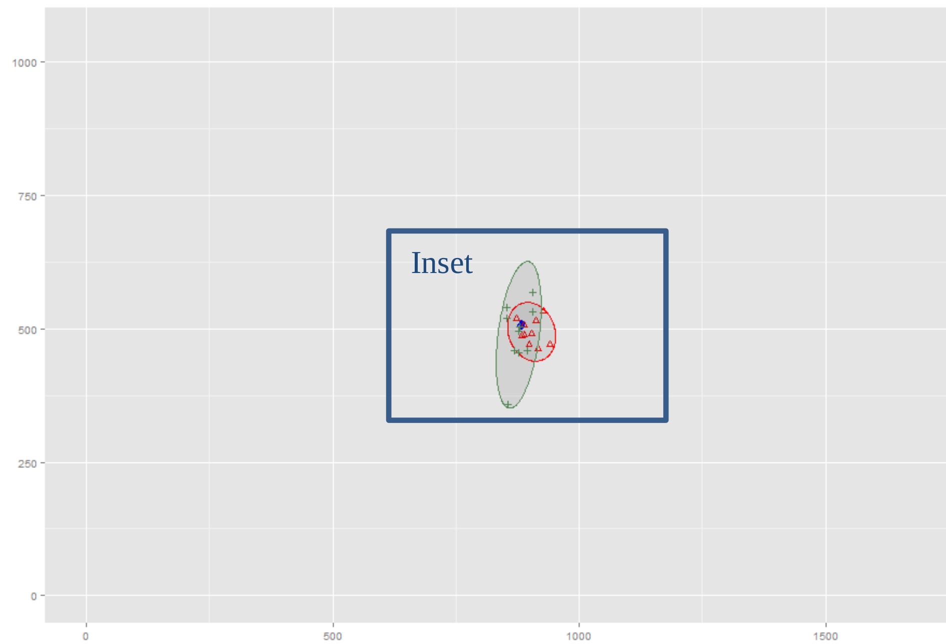

Figure 29 and Figure 30 illustrate that each participant had a different response pattern around what was one of the alternatives with the highest correct response rate to which they had to respond. With the method outlined above calibrating at regular intervals throughout the experiment, a researcher could adjust the data used to model a participant's performance; without such data, this example makes the assumption that the physical error in making a response is constant throughout the experiment for each participant.

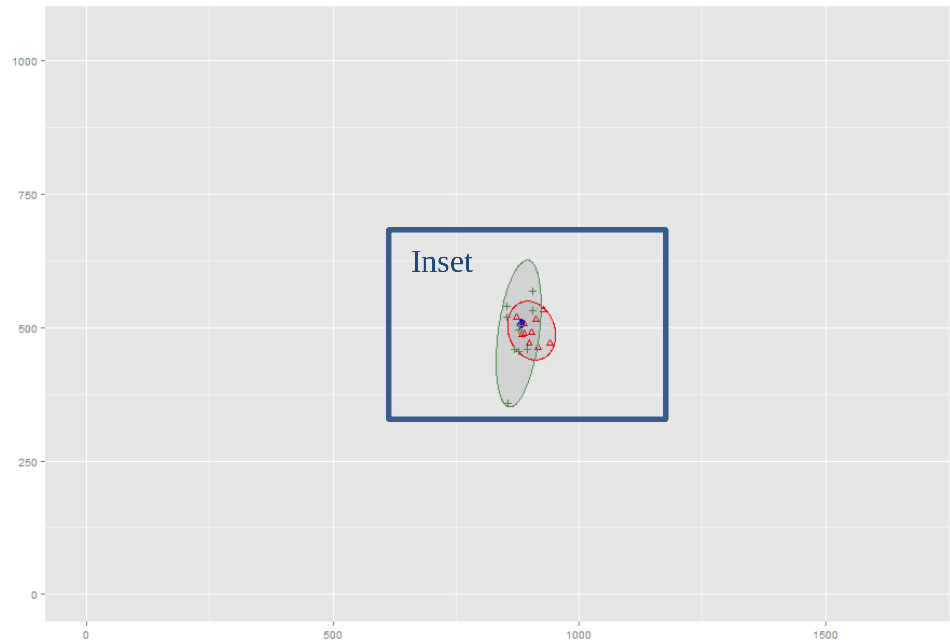

Figure 29: Calibration Points and 95% Bivariate Normal Ellipses for a Participant with a small distribution (blue circles), a medium distribution (red triangles), and a wide distribution (green +'s). Units are pixel coordinates on the stimulus image.

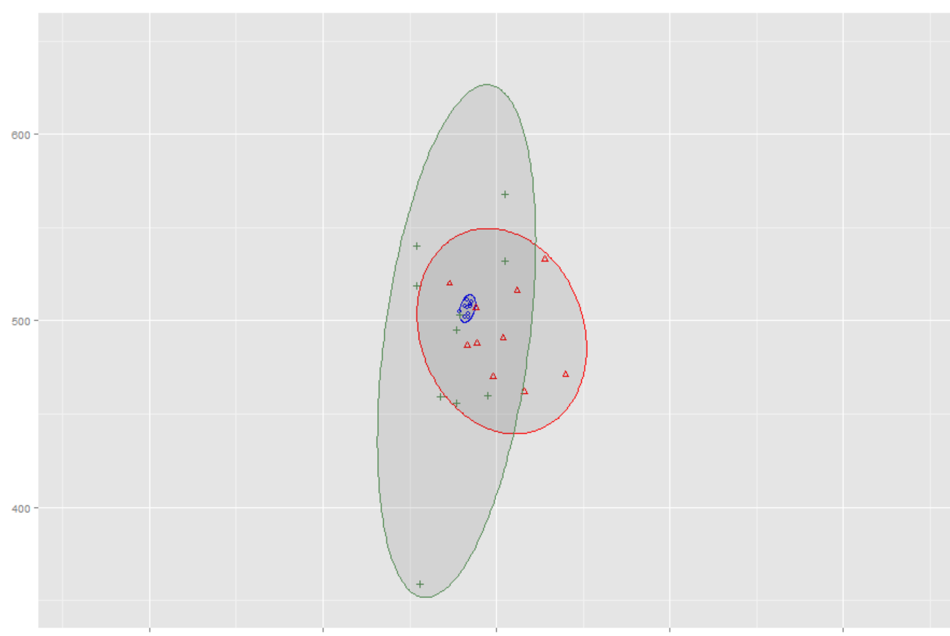

Figure 30: Inset of Calibration Points and 95% Bivariate Normal Ellipses for a Participant with a small distribution (blue circles), a medium distribution (red triangles), and a wide distribution (green +'s). Units are pixel coordinates on the stimulus image.

Next, an elliptical bivariate normal distribution was fitted to the data. If the paths of the mouse movements were recorded and the velocity and direction of the mouse at the click were computed, the distribution could be fit such that the axes were aligned with the path of the mouse and that the velocity could affect the size of the variances, which could be independent such that the axes are not centered in the elliptical distribution. With sufficient calibration data and a recording of the mouse path, modeling these nuances would be possible, but the data from the work zone case also lacks that information and did not record mouse paths. As a result, there are several assumptions: first, the axes for the general calibration data hold true for each point; second, the general shape of the ellipses are such that the axes are centered across the distribution; third, the distributions are assumed to be bivariate normal, centered around the mean of the points. These assumptions are not ideal and as stated could be eliminated with better calibration data, but they are useful for illustrating this example.

Third, a bivariate normal distribution was fit to each participant's points using maximum likelihood estimation. Such a model has an advantage that the probability density function of the overall distribution is the same probability density function of each individual point. Table 10 shows the mean and correlation table for each of the three participants. From these, it is clear to see that while the means were very close, the variance in responses goes from very little with the first participant to quite large with the last participant.

Table 10: Mean and correlation of bivariate normal distributions fitted to the responses of three representative participants' clicks for the PCB alternative at 3 seconds in the open condition from Experiment 1. Results show that mean response is similar across participants, but variance in both the x and y indicates changes in spread and orientation of the distributions.

$$\mu_{tight} = (883.3, 506.6) \quad \sigma_{tight} = \begin{bmatrix} 3.61 & 2.02 \\ 2.02 & 9.24 \end{bmatrix}$$

$$\mu_{medium} = (903.1, 494.5) \quad \sigma_{medium} = \begin{bmatrix} 399.09 & -72.85 \\ -72.85 & 507.05 \end{bmatrix}$$

$$\mu_{wide} = (877, 489.1) \quad \sigma_{wide} = \begin{bmatrix} 347.60 & 405.91 \\ 405.91 & 3155.31 \end{bmatrix}$$

Moving forward, the integral of the probability density function for each point over the polygons defined for the original categorization system was computed. This value represents the probability that each point was intended for each zone. Table 11 shows an example of such a point, where the categorization in the absolute categorization system from Chapter 3 was for Work Zone, though the integral of the PDF suggests more ambiguity. In contrast, other points, the complete set of which can be found in Appendix B, show how most points' categories matched the zone of greatest probability.

Table 11: Example of a point with some ambiguity from fitted probability density function. Integration of the probability density function over categorization zones suggests that the likelihood of intended categorization is low.

x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)	Pr(Indet)
916	493	D40A	Open	WZ	21.84%	23.83%	0.00%	45.33%

The implication of this analysis is that results could vary if participant are not precise in their responses. Analysts can check a participant's variability using this method of integrating the probability density function and use those results to compute how much variability a lack of precision would add to the overall results.

Alternative Answer Marking

With the ubiquity of the personal computer in research laboratories comes the ubiquity of the mouse in research, but they are not the only devices available for recording responses. A computer mouse is easily deployable and almost universally compatible, so they are commonly used for location response recording. The computer mouse has its own issues that can lead to physical error, however, including calibration issues and the likelihood of clicking prematurely. To say that mice are common is not to say that other methods of data collection are not available.

One method for data collection would be to grid the screen for locations and use the keyboard as a data collection method (Zhu, Ma, Feng, & Sears, 2009). This eliminates some of the potential for misfire, but does still leave the opportunity for pressing the wrong keyboard key. This screen-grid method also eliminates the need for categorization (discussed in later chapters), but also provides some inherent cueing for participants; if the participants know that the ramp is going to be in one of, say, 9 parts of the screen, their activation map is pre-cued for those nine parts of the screen.

Another solution to the problem of misfire may simply be to change the input type. Rather than terminating the image and registering a mark at the start of click, the participant may be instructed to draw a shape over top of the point they wish to mark. An example is (Accot & Zhai, 2002), who had participants indicate their responses by drawing an X with the mouse button depressed. This crossing input method removes some speed from the participant’s response, but allows for greater precision.

Conclusions

Computer mice are readily available as pointing devices for localization research, but it is important to understand that participants may have some physical error in using them to mark a response. The categorization method used in Chapter 5 assumes that participants' responses match participants' intent, which may be masked by physical error from marking response. This chapter accounts for that physical error by proposing that future researchers test the precision of participants at regular intervals and fit distributions describing the precision of each participant. This chapter also illustrates with sample data how future researchers can determine the likelihood that a participant's response was intended for a particular category. Using these methods of regular calibration measurements and integrating the likelihood of assigning a point to a category, researchers can account for the uncertainty introduced by physical error from using a pointing device.

CHAPTER 7

EVALUATING COMPREHENSION

This chapter will discuss how the information relevant to informing design can be extracted from the dichotomous categorical data developed in Chapters 5 and 6. Beginning with the reasons for modeling comprehension data, the *Item Response Theory* approach, in particular the Linear Logistic Test Model, will be discussed as a suitable model for evaluating the experiments presented in Chapter 5 and illustrate how these models could be used to analyze detection/localization and identification responses. The chapter will conclude with a discussion as to how these models could be further extended in the future.

Modeling Comprehension

Traffic control devices must be comprehensible by road users from diverse demographic groups with varied skill sets. As designers and engineers work to improve their designs, they must be able to predict both each individual's ability to comprehend and the relative performance of traffic control alternatives. While the analysis performed in the work zone case is certainly useful in developing such a predictive capability (i.e. aggregate error rates can inform practitioners about the relative performance of traffic control alternatives). Aggregate response analyses suffer a similar range of problems as comparing devices with crash data: the results are largely aggregated and impersonal. Earlier work in Hunter, et al. (2014) used ANOVA for comparisons, but some of the literature suggests that ANOVA may not work well with categorical data (Jaeger, 2008; Warton & Hui, 2011), and ANOVA lacks true predictive capability. Here, we consider an alternative, the Linear Logistic Test Model (LLTM), from the Item Response Theory (IRT) group of models.

Item Response Theory

Item response theory is a way of measuring human performance. It was developed as an alternative to Classical Test Theory. Classical test theory uses the total points awarded from many questions to determine a competency score and is the traditional way that most tests in schools are scored. Item response theory considers at least two parameters (e.g. item difficulty and performance) in conjunction with logistic regression to estimate participant skill. This is the way many standardized tests, including the ACT and SAT, are scored. Item response theory can also be used to evaluate the difficulty of questions. This is because the estimation of the parameters depends only on the raw scores from either category (Embretson & Reise, 2000).

The IRT approach is compatible with these data. IRT models use logistic regression, so they use dichotomous data such as the binary (yes/no) data discussed in Chapter 5. These models can used performance to estimate skill level and item difficulty for a population or, conversely, having known or fixed values for either population skill levels or item difficulties allows researchers to predict performance. The basis for this approach is the Rasch model (Masters & Wright, 1984):

$$\Pr(X = 1|\theta, \beta) = \frac{e^{\theta-\beta}}{1 + e^{\theta-\beta}}$$

Where θ is an ability-related parameter and β a difficulty-related parameter. Given the structure of the model, Item Response Theory thus allows simultaneous evaluation of the characteristics of both items and test-takers. By combining these scores, the experimenter can see not only which items are more or less difficult, but also which items are more or less difficult for individuals with a lower ability score (θ). This allows for further investigation into a subset of the population which is generally not accounted for. While the concept of the "Design Driver" exists as a theoretical construct representing the 90th, 95th, or 99th percentile driver, assigning ability scores to a particular driving task allows researchers to directly identify those drivers.

Fundamentals and Assumptions

This model requires a few assumptions, and it is important to acknowledge how they can impact the interpretation of IRT models. These assumptions are a unidimensional ability score, representative and homogeneous sample, positive monotonicity, and local independence (Junker & Sijtsma, 2001).

The assumption of a unidimensional ability score implies that the only reason for getting a question right or wrong is the single skill in focus. For this project, θ could be called that "Traffic Control Device Detection and Localization" and "Traffic Control Device Identification." To determine if the score is unidimensional, it is necessary to decompose the task into separate subtasks and see if some separate measures would better help to describe these scores, which Chapter 4 previously explored. However, if those subtasks are intertwined to the point of being necessary and parallel, it could still be beneficial to model individual skill at each subtask as a single parameter to accurately mimic population variability that can lead to the same response outcome.

The second assumption, specific to forecasting, is sample homogeneity and generalizability. The sample selected is, in unwavering terms, definitely not representative of the demographic diversity common in the driving community. One of the advantages of this model type, however, is that while the θ's are dependent on the sample, the Bs are not. This means that in a new sample, known βs would still provide θ scores on the same scale as the original sample. This characteristic of the model, the ability to estimate the item difficulty or the participant skill level independent of the sample, is known as "specific objectivity". Specific objectivity is very useful for model estimation and forecasting, which will be discussed in a later section.

The third assumption is positive monotonicity of response variables, which states that as ability score increases, the likelihood of correct response also increases. This may be a concern in situations where an increase of the measurement variable may lead to a decrease in likelihood of correct response; for example, increasing overall skill at mathematics may lead to a decline in ability to correctly perform basic arithmetic. This assumption holds for measures of skill because it is intuitive that as skill increases, performance increases.

The last issue is local independence. The point of presentation should not impact the resulting answer. A fully randomized design accounts for some of this potential; however, repeated trials may have either a practice or fatigue effect with seeing the same image multiple times. Indeed, the Rasch model is not equipped to respond to repeated trials, since each stimulus item is estimated with its own difficulty parameter. An extension of the Rasch Model, the Linear Logistic Test Model (LLTM), accounts for this by decomposing the difficulty parameter β into a linear combination, separating out the term for time series effects.

Linear Logistic Test Model

The Linear Logistic Test Model extends the Rasch model by defining the difficulty parameter β as a linear combination of weights (often expressed as dummy variables by being defined to be either 1 or 0) and a characteristic parameter η (Kubinger, 2008). Analysts can decompose the difficulty parameter β into many variables of interest. The formulation of the model is:

$$\Pr(X = 1|\theta, \beta) = \frac{e^{\theta - \Sigma w\eta}}{1 + e^{\theta - \Sigma w\eta}}$$

This extension of the Rasch model is also advantageous in that rather than simply comparing the β values for each alternative, a common η for all alternatives can be estimated as well. An assumption of this formulation is that the weights of each η are known or decided upon prior to the experiment. This is not of much concern to this experiment since the weights are binary dummy variables.

Limitations

There are two limitations to the Linear Logistic Test Model, one specific to Rasch model extensions and one generally about Item Response Theory. First, the shape of the logistic curve used to estimate the probability of correct response to a question, commonly called the Item Characteristic Curve (ICC), is assumed to be the same shape for all items in a Rasch model. Some variations within IRT, such as the 2PL and 3PL model account for varying ICC spread with a 'discrimination parameter', but this removes the specific objectivity of the model and makes it more sample dependent (Embretson & Reise, 2000). The second limitation is generally about all Item Response Theory models: questions must be sufficiently difficult and sufficiently easy that participants do not answer all of one question correctly or all of one question incorrectly. In such a case, estimating the parameters is useless since the parameter of interest exceeds the limits of the model. In such a case the model estimates the parameter as unbounded positively or negatively. This can be managed in processing by removing the question from consideration, but this should be avoided by pilot testing questions with the populations of interest to ensure that questions are comprehensive enough to estimate the range of parameter values.

Application to Sample Data

The application of this model is straightforward and can be accomplished with readily available software packages (P. Mair, 2007). While the data from the experiments in Chapter 3 are not in the dichotomous format of data proposed from the experiments in Chapter 5, a few assumptions can give us sample data permitting the demonstration of sample calculations. Since the pathways for the closed and open conditions response vary, this section will only use the closed stimuli at the 3 second distance in Experiment 1. These had fewer "Work Zone" responses, so the open/closed dichotomy is more closely followed. A further assumption however is that work zone responses count as "incorrect" along with timeouts and indeterminate, though these can be counted as missing data with a sufficient sample size. While Chapter 4 explained how detection, localization, and identification are subtasks of comprehension, these data are not separated in the case experiment. This example modeling requires the assumption of unidimensionality of the overall comprehension task for the purpose of demonstrating the modeling procedure, although in implementation three separate model should be estimated representing the dichotomous data for the three subtasks. A later section will explore how to adapt the model to Chapter 5's methodology.

A Note on the Predictive Power of Data

The predictive power of a model of human performance is only as good as its sample. To effectively extrapolate out to a population, the experimenter must be sure that their test group is representative of that population. Using the data from Chapter 3's experiments does not allow for such predictive power, since the demographics of the participants were not collected. Even the ability to compare between institutions where data was collected is limited due to small sample size in Experiment 3, the only experiment which contained participants from both populations. Thus, it is left to future researchers to compare populations using this test. The process, however, is only a few steps beyond what is shown here: experimenters can use descriptive statistics to highlight differences in the estimated skill level distributions for a group of interest as compared to a control group.

This method can be applied in several ways toward the general population. First, it can be used to determine the population's skill levels. By anchoring the β values through initial conditional maximum likelihood estimation, the experimenter can determine the general population's ability scores by presenting the experiment in front of diverse population samples. A number of educational tests use this method as a way of estimating the ability scores of test takers.

Estimation of the Model

With data parsed as a matrix of all presented stimulus items and all participants, the data can processed into item difficulty scores and ability scores as a two-step process. First, using the characteristic of specific objectivity, it is possible to estimate the item difficulty parameters from the total responses using conditional maximum likelihood estimation (Andersen, 1972; Pa. Mair & Hatzinger, 2007). Estimating item difficulty from Experiment 1 in the 3 second and closed condition for the first three repetitions (to avoid overcomplicating sample calculations), the estimated βs for individual items are found in Table 12. Since presentation of a stimulus was randomized so that each repetition was shown once within a randomized block, the block numbers serve as timepoints to be used to show how many times a participant had been presented with a stimulus. At first glance, these β values are intuitive--Portable Concrete Barrier has the lowest item difficulty score, while misaligned drums have the highest. Note that because some question and timepoint combinations had no incorrect answers, those questions were removed from the model. Specifically, the Portable Concrete Barrier responses at timepoints 2 and 3 received no incorrect responses; as explained in previous sections, a weakness of IRT models is that they cannot estimate parameters for perfect scores, since the parameters of the model at Pr=1 are unbounded.

Table 12: Calculated Beta values for Experiment 1 at the 3 second distance in the Closed Alternative

Timepoint	Alternative	Beta
1	D10A	1.132
	D40A	1.477
	D40M	1.132
	PCB	-3.741
2	D10A	0.346
	D40A	0.691
	D40M	0.346
	PCB	NA
3	D10A	0.392
	D40A	0.047
	D40M	0.392
	PCB	NA

The impact of presentation is lost in the previous table, though. Looking at the η values from the model as well, however, shows that when put into a linear combination, the effects of the alternative are easier to follow (Table 13). These results fit those from the analysis in Chapter 3, with Portable Concrete Barrier being the least difficult alternative, although for this model estimation at this alternative combination, D40A appears to be a more difficult alternative than D40M and D10A.

Table 13: Etas for Linear Logistic Test Model

D10A	D40A	D40M	PCB	Time 2	Time 3
1.132	1.477	1.132	-3.741	-0.786	-1.525

The next step in analysis is estimating the person parameters, the skill level θ for each participant. Solving for the LLTM equation using the previously estimated βs gives θ values that are relevant for the sample calculated. Again, these are not representative of the overall population, but those θ's can be determined with a known sample. Table 14 shows the person parameters and standard error.

Table 14: Estimate Participant Ability Scores

Participant	Estimated Theta	Standard Error
P1	1.791	0.831
P2	2.666	1.084
P3	-1.693	0.984
P6	2.666	1.084
P8	1.791	0.831
P9	-3.054	1.379
P10	1.185	0.739
P12	1.791	0.831
P13	0.181	0.700
P15	1.791	0.831
P16	-0.910	0.807
P17	1.791	0.831
P18	-3.054	1.379
P19	-1.693	0.984
P20	1.791	0.831
P21	-0.910	0.807
P23	0.670	0.703
P24	-0.910	0.807
P25	0.181	0.700
P26	1.185	0.739
P27	-3.054	1.379
P28	2.666	1.084
P29	1.791	0.831
P30	-3.054	1.379
P32	0.670	0.703
P33	1.791	0.831
P35	-0.910	0.807
P36	-1.693	0.984
P38	0.181	0.700
P39	1.791	0.831

The last step of model estimation is computing the goodness of fit. A common metric for goodness of fit is the Andersen Likelihood ratio test (Andersen, 1972), which computes both the log-likelihood that the data can be estimated using the model and the p-value associated with that. The likelihood ratio statistic for this model is 5.671, yielding a p-value of 0.684. Clearly, these results do not indicate excellent model fit, but that can be expected for two reasons: first, the sample size is relatively small for an Item Response Theory calculation for a rather homogenous population. The more important reason though is that these data are not unidimensional--as demonstrated in Chapter 4, the subtasks of detection require separation in the analysis of an experiment that tests these questions separately.

Extending to Proposed Methods

While these sample calculations serve as an example of what types of outputs to expect and how to approach them, application to the proposed methodology will require some revisions to the process. First, recording data which is categorized as correct or incorrect from experimental administration software is very helpful for input into processing programs. Most importantly though, the skill levels and item difficulties for the detect/localize phase and the identification phase of the experiment must be estimated separately. These values are each unidimensional, but in combination it is quite possible that a device or combination of devices is, for instance, easy to detect but difficult to identify. Similarly, a participant may easily identify a device, but struggle to detect it. Comparing these values, estimated separately, will offer further insight into the limitations and strengths of traffic control alternatives.

Model Uses

There are three main uses of this model. This chapter's sample calculations showed how the model can be used to describe a sample's performance and compare alternatives and participants within that sample. There are two further extensions of this model, however: comparing alternatives against known baseline traffic control and identifying populations which require extra consideration during design.

First, this model can be used to develop baseline difficulty scores for common traffic control for use as a baseline comparison. By taking images that are commonly used and banking them to be used in all tests, the estimated difficulty levels of these items can be fixed to better inform comparisons with new traffic control. This would require extensive testing of those images with a large sample size and a sample that is known to be representative of the design audience.

The second further application of this model is to identify target groups. By comparing a control group's θ's to a target group's θ's, the experimenter can identify if any group characteristics, such as health or demographics, vary with θ as opposed to a control group. Identifying these groups with more difficulty detecting, localizing, or identifying traffic control can inform designers as to the vulnerable groups who should be the focus of design.

Conclusions

This chapter first outlined the issues with using aggregated results to measure the differences between alternatives. In order to develop a predictive model that can measure both alternative differences and participant differences, the Item Response Theory group of models are introduced for this application. Specifically, the Linear Logistic Test Model could be used with data from the proposed experiments because it decomposes the difficulty score into a linear combination of variables which can be used to measure time series effects in the data. Sample calculations illustrate how researchers can use these models to interpret data. This model can be used immediately to measure differences between alternatives being tested, but a long term effort could also be used to measure the difficulty of known stimuli to bank responses as a reference point for future research. This test can also be used in conjunction with demographic information to identify traits of participants who do poorly in these comprehension tests. These efforts are outlined in the next chapter.

CHAPTER 8

CONCLUSION

The objective of this dissertation was to explore issues related to the testing of human perception of systems of traffic control devices and to develop guidance for such testing through the use of an extensive work zone related case study.

Chapters 1 and 2 discussed the safety problem in work zones and difficulty of evaluating driver comprehension of work zone safety devices compared to that of its physical performance, for example its crashworthiness that can be tested by well-established standard procedures. In the absence of such procedures, Chapter 3 discussed a series of efforts to evaluate how drivers comprehended work zone traffic control at diverges on freeways. This study had several notable results, including the evaluation of the Gestalt principles of grouping as design guidance for practitioners that lead to the subsequent development and testing of a novel Linear Channelizing Device.

The second portion of the dissertation focused on evaluating and improving upon these initial methods. Chapter 4 explored issues with the analysis of data from the work zone case study discussed in Chapter 3, ultimately leading to a proposal to decompose the task of comprehending traffic control into the three subtasks of detection, localization, and identification. By separating out these subtasks for evaluation, researchers can gain a more nuanced view of a participant's underlying reason for errors, when they occur. An experimental procedure for measuring performance related to those three subtasks through a computer-based test was introduced in Chapter 5. This chapter also described methods for pilot testing and classifying participant responses for the localization subtask. These methods can be applied quickly using readily-available equipment, thus allowing researchers to employ testing in a user-centered design framework.

Chapter 6 explored additional issues related to this testing including potential noise in computer mouse responses. This discussion centered on methods to allow researchers to account for differences between participants' physical ability to indicate their intended response to better separate cognitive and physical errors. Chapter 7 introduced *Item Response Theory* as a framework for both measuring participant comprehension and comparing understanding of traffic control alternatives..

After analysis of the three experiments described in Chapter 3, there were several main findings. First, the correct response rate was lower for drum alternatives than for the Portable Concrete Barrier or Linear Channelizing Device. This implies that the Gestalt principle of closure can be employed to reduce error rates in temporary traffic control. Second, the correct response rate was lower for misaligned drums than aligned drums or PCB and LCD alternatives, implying that the Gestalt continuity can be used to reduce error rates. Third, there were no significant differences in the performance of participants observing 10 ft. or 40 ft. spacing between drums, implying that proximity does not have an impact in error rates, at least at those distances. These findings were employed in the design of a Linear Channelizing Device, which was developed to explicitly to demonstrate continuity and closure and had similar correct response rates to the Portable Concrete Barrier alternatives with a smaller physical profile.

Another finding was the result of difficulties with analyzing the data from the experiments in Chapter 3. While the experiments were envisioned as Yes/No detection, the data suggested that there were actually three sub tasks to comprehension: detection, localization, and identification. These three sub tasks created multiple pathways to each type of response, complicating analysis. Chapters 5 and 6 built off of these findings to develop proposed methodologies for future work. Finally, Chapter 7 outlined the application of the Item Response Theory models to analysis of comprehension data.

This work has impact both immediately and for directing work in the future. The immediate contributions are the results of the work zone case: the decomposition of the comprehension task in relation to the results of the work zone case, the methodology for data collection, and the application of IRT to traffic control comprehension.

Contributions

This dissertation contributes to the knowledge within the field of traffic control design. The first contribution is the work zone case study. This study developed comparable still images of work zone diverges as stimuli in a detection task. The data showed that the diverge location and identification task was more nuanced than expected. The data were categorized for aggregated analysis and comparison of temporary traffic control treatments was accomplished. Included in the work zone case contribution are the development of the Linear Channelizing Device, the application of Gestalt grouping principles to traffic control design, and practical recommendations to agency inspection standards.

The second contribution is the decomposition of comprehension into the three subtasks of detection, localization, and identification based on the multiple pathways to response from Chapter 3's work zone case. Designers and researchers can use this framework to identify the underlying cause of comprehension problems with current and novel traffic control. The third contribution is the experimental methodology developed in Chapter 5. This method is implementable using widely available personal computers and allows researchers and practitioners to quickly test comprehension among many participants in a low-risk, laboratory setting. The fourth contribution is the application of Item Response Theory to traffic control comprehension. This modeling technique allows for identification of target populations and for comparison of traffic control alternatives through a single test.

These contributions are useful to practitioners now. The design principles from the work zone case are employable in temporary traffic control planning today. Further, practitioners can use the method and modeling strategies from Chapter 5, 6, and 7 to test devices and alternatives against each other for comprehension. Expansion on this work both through research and a large practice to develop a difficulty score standard for devices, though, are exciting opportunities for future work.

Future Research

The work presented here is a step forward in the evaluation of the design of traffic control, but there remains needed research to reach the goal of a safe, comprehensible traffic control system. First, using the methods described in this dissertation, researchers can directly compare traffic control devices and systems. With a known, representative sample of the population, this can contribute to a “stimulus bank” where certain images are known to have fixed difficulty scores for future comparison. This allows for the next category of future research, the refinement of these methods into a common, implementable regulatory standard. An advantage to using the ability scores and difficulty scores from Item Response Theory models is that they give a distribution of the population’s ability to respond correctly to traffic control, temporary or otherwise. As a result, with future work to calibrate these tests to the general population, a minimum difficulty score could be established that novel devices would need to exceed in order to be approved for use on public roads.

A second direction for future research is direct validation of the assumptions of this work. The decomposition of the task of comprehension into detection, localization, and identification was derived from observations and analysis made of the data in the work zone case. The resulting subtasks are derived from understanding of the psychology literature and state of the practice, though theories and understandings of the cognitive process of perception may change quickly, especially as neuroscience works to validate psychological theories. Future work should adapt newer methods from these fields into a validation of this task decomposition.

Another point for validation is the implicit assumption that improving comprehension would lead to improved safety. While it is appears logical that increasing comprehension will decrease errors and improve safety, it is also possible that improved comprehension would have no significant impact on crash rates, or worse that it could unexpectedly lead to unsafe behavior by some drivers and higher crash rates. Future work should empirically establish the link between comprehension, error rates, and crash rates. Several routes are possible for this research. First, researchers could establish the behavioral response to known traffic control devices with high comprehension rates in a driving simulator or instrumented vehicle. Then researchers could compare other metrics of driving performance to traffic control alternatives to establish links between comprehension and drivers' action. Second, researchers could use crash data to conduct a larger statistical analysis in corridors where various traffic control strategies are used.

A large-scale study could compare the crash rates in these areas to the difficulty scores of various devices.

Impact on the Practice

Practitioners have used standards for testing the crashworthiness of devices for decades, but both a lack of comprehension testing standards and performance standards have held back the industry from testing the comprehension of devices. This work offers a testing method to develop a comprehension standard for temporary traffic control devices and systems.

Developing a comprehension standard would be a large undertaking; practitioners would need to test a wide range of images presented as stimuli to calibrate the difficulty scores for a wide range of commonly used stimuli. Because the nature of an IRT model, stimuli would need to cover a wide range of difficulties to be predictive. Also, because of how difficulties are computed, extensive pilot testing would be required to ensure a wide range of difficulties in a standard set of banked stimuli. These banked stimuli difficulties should be established using a representative sample of the population of drivers. Again, though, since comprehension is computed from the difficulty scores, the process of determining a representative sample of participants with a range of ability scores would require pilot testing. As a first step, a subset of participants with demographics representative of the driving population could be used as a starting point for determining the range of observed ability levels.

While the process of developing a bank of stimuli with known difficulties would take many rounds of testing and validation, the results have the potential to greatly improve the quality of messages being conveyed by temporary traffic control.

Manufacturers and inventers of new devices could quickly test the comprehension of new devices in the design stage, rather than after design is complete and ready for production. This rapid, low cost (both in time and in resources) testing could be integrated into user centered design not only as a way of comparing design alternatives to each other but also to the banked stimuli to see where on the spectrum of comprehension these new designs would fall.

Another use for the test developed in this dissertation is to identify groups that may need special consideration in the design of temporary traffic control. With a standard set of stimuli, participants with particularly low comprehension ability scores are identifiable for further investigation. Researchers could investigate links between physical, mental, or cultural characteristics and ability score to identify any traits that would impact safety while perceiving the driving environment. Designers could use such information to ensure that designs are comprehensible to all groups of drivers by focusing design around groups with identifiably poor comprehension.

Lastly, a combination of the method in this dissertation with renderings from roadway design software may allow for evaluating how drivers comprehend not only new devices, but also specific proposed temporary traffic control plans. Every work zone is unique, and temporary traffic control is site specific even though the devices themselves are certified for crashworthiness. By enabling designers to test specific temporary traffic control plans for comprehension, engineers can have a better sense of potential problems and needed design revisions to the temporary traffic control plans on a specific project.

In its current state, temporary traffic control design depends heavily on the judgment of a designer to ensure comprehensibility of the temporary traffic control plan, even though it has extensive testing methods for evaluating crashworthiness of devices. While other fields have moved to a more user-centered design, without a standard testing method temporary traffic control designers have not been able to evaluate their plans. The work in this dissertation offers several steps needed to move temporary traffic control design forward toward user-centered testing, and thus to temporary traffic control that roadway users have a better opportunity for comprehension.

APPENDIX A: IMAGES FROM WORK ZONE EXPERIMENTS

These images represent the characteristics varied in the stimuli presented to participants. This Appendix is divided by experiment and also by the characteristics used in the images. While each combination of image characteristics is not shown, each variable of interest is represented.

Experiment 1

Alternatives

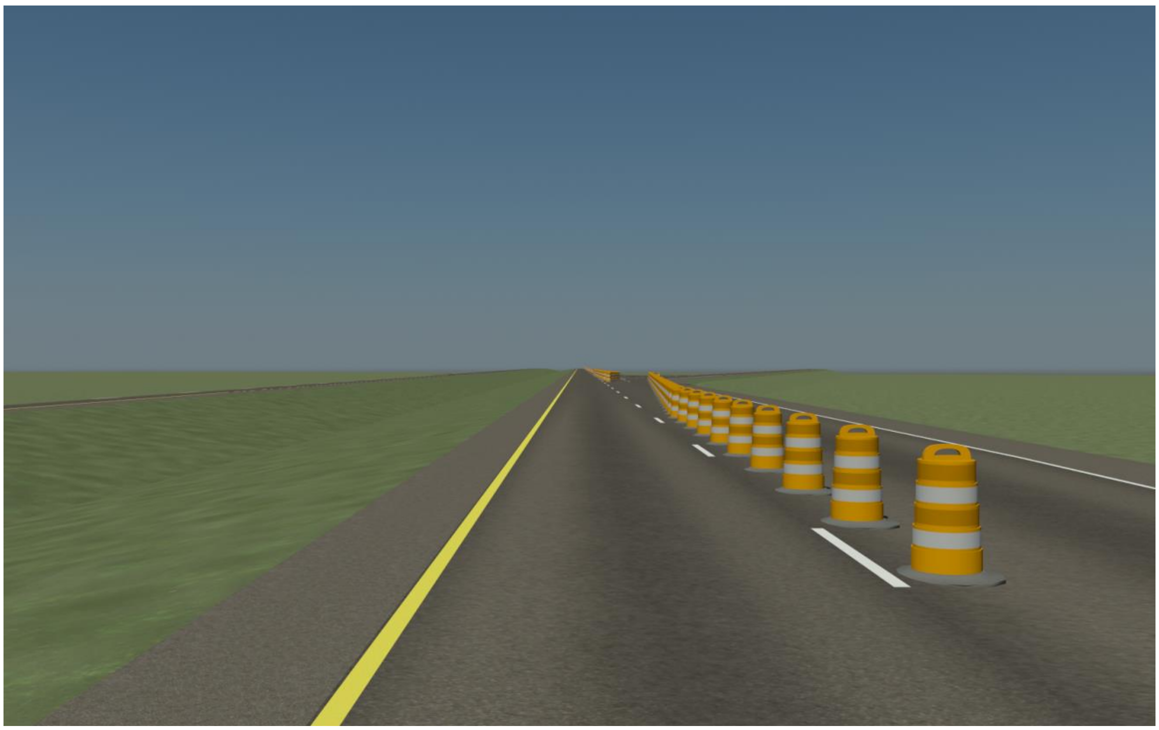

Figure 31: Representative image of the D10A alternative. Image from Experiment 1, D10A alternative at 1 second distance in the straight geometry and the open condition.

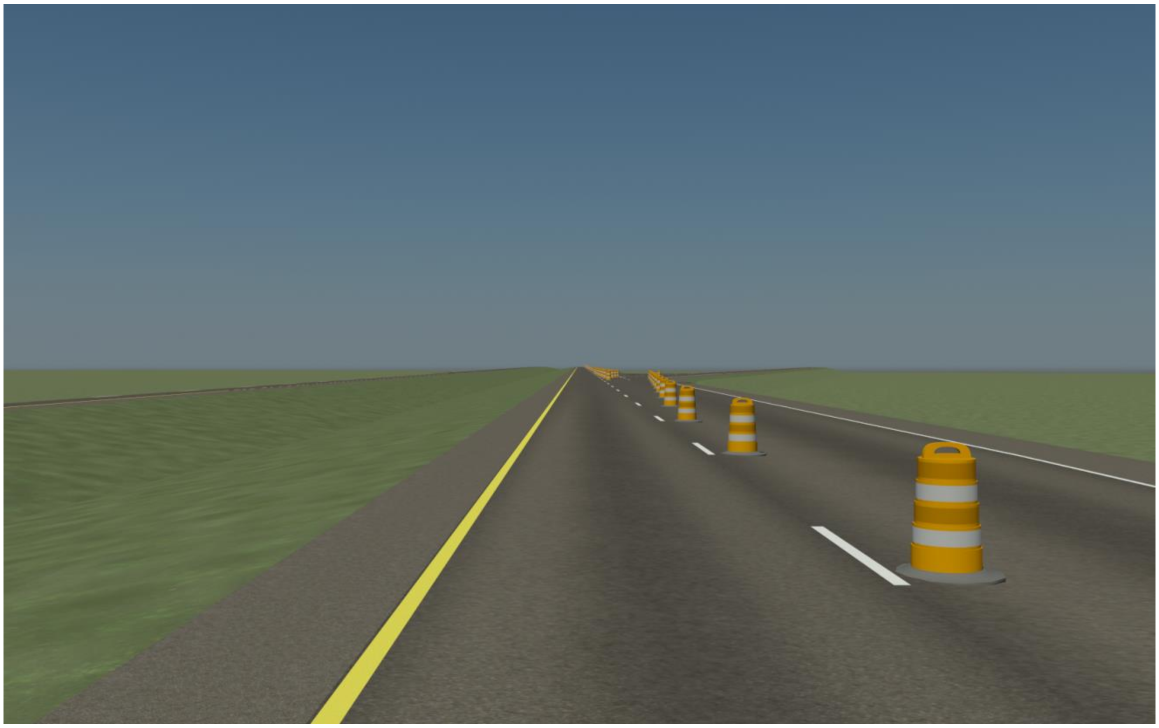

Figure 32: Representative image of the D40A alternative. Image from Experiment 1, D40A alternative at 1 second distance in the straight geometry and the open condition.

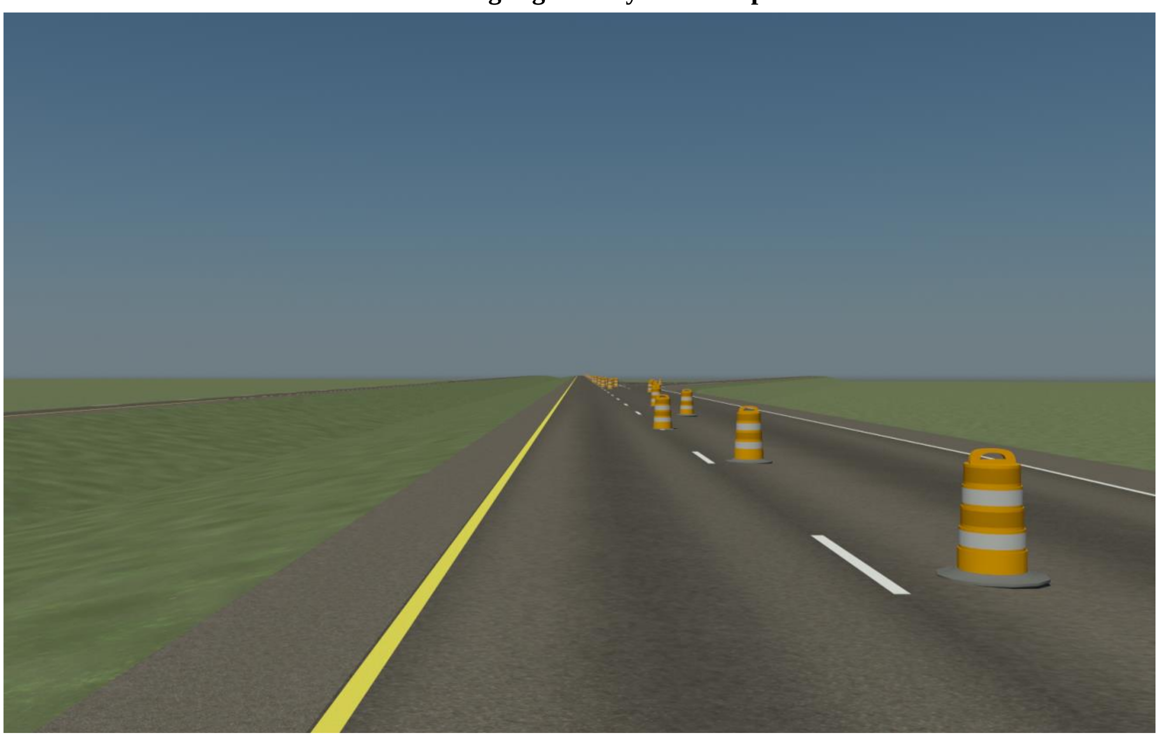

Figure 33: Representative image of the D40M alternative. Image from Experiment 1, D40M alternative at 1 second distance in the straight geometry and the open condition.

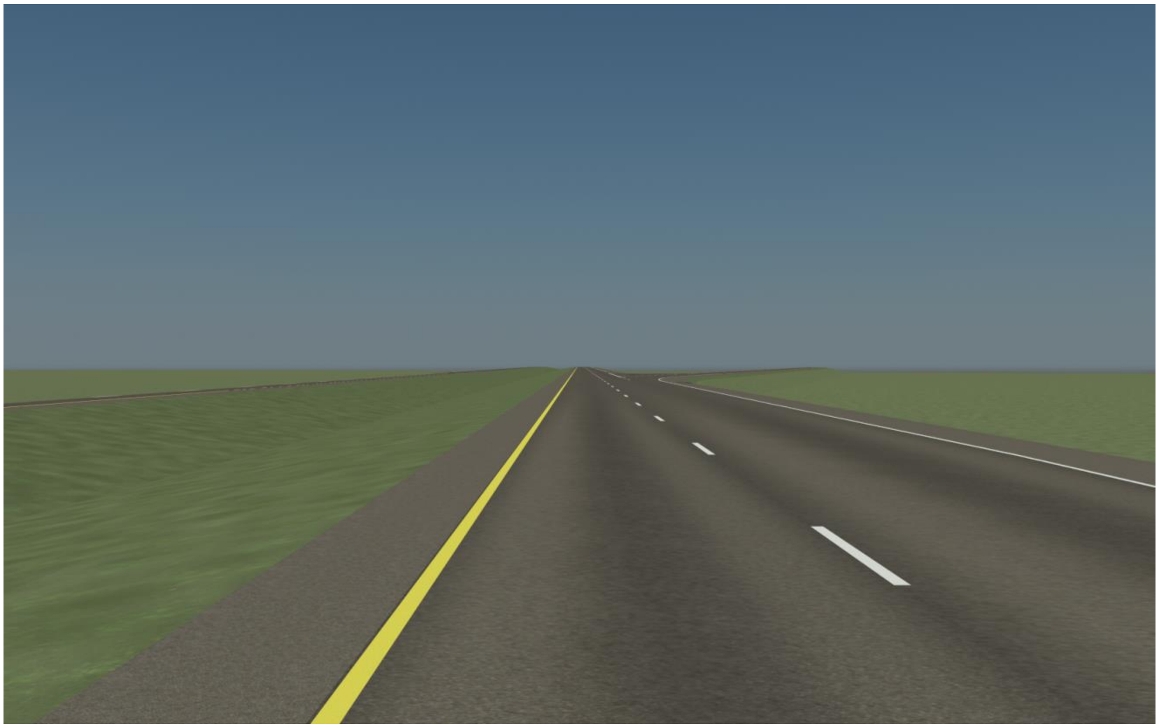

Figure 34: Representative image of the No Work alternative. Image from Experiment 1, No Work alternative at 1 second distance in the straight geometry and the open condition.

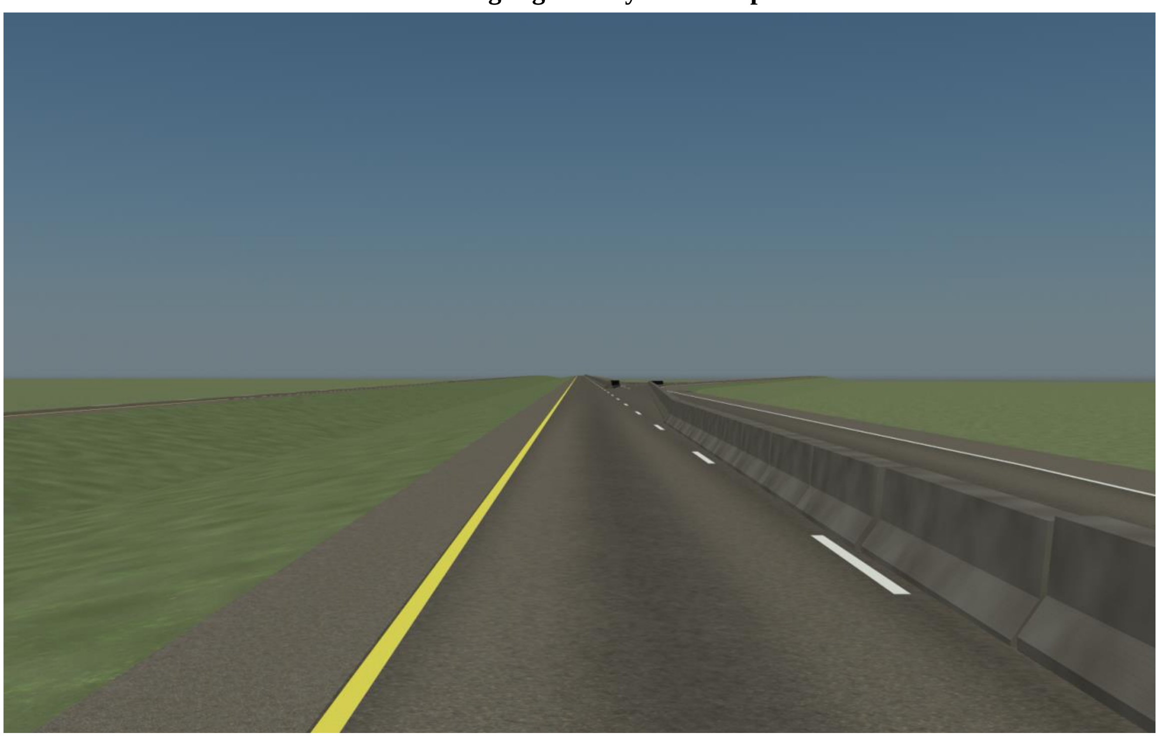

Figure 35: Representative image of the PCB alternative. Image from Experiment 1, PCB alternative at 1 second distance in the straight geometry and the open condition.

Distances

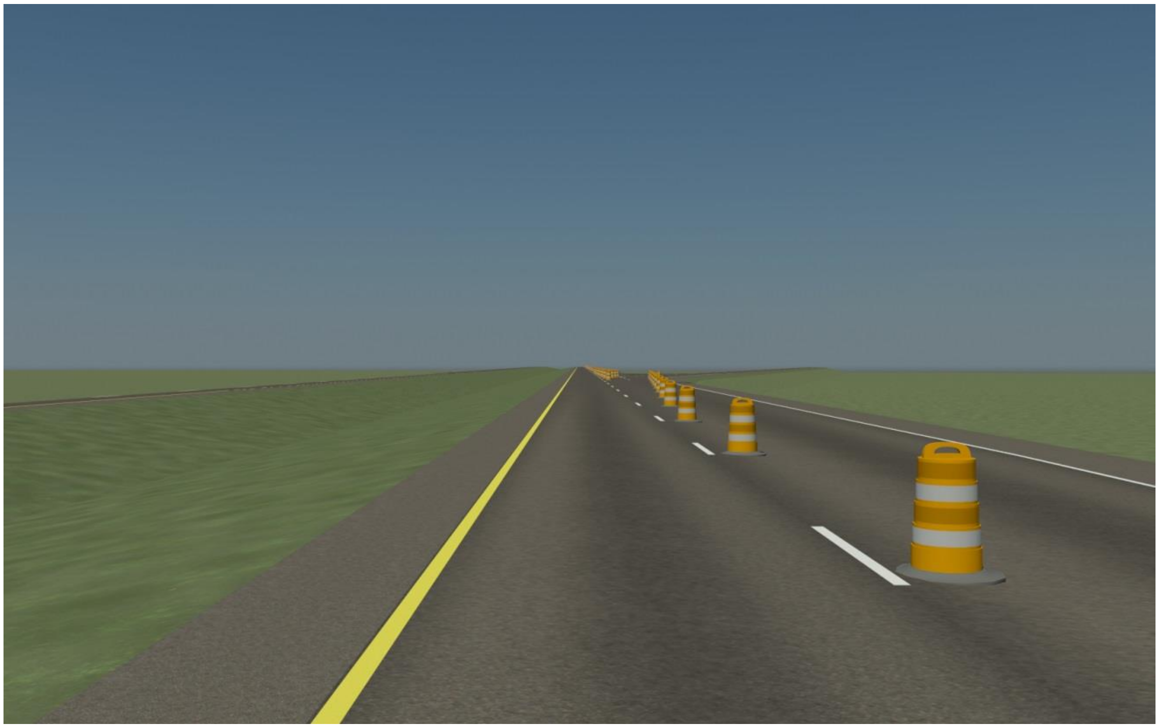

Figure 36: Representative image of the 1s distance. Image from Experiment 1, D40A alternative at 1 second distance in the straight geometry and the open condition.

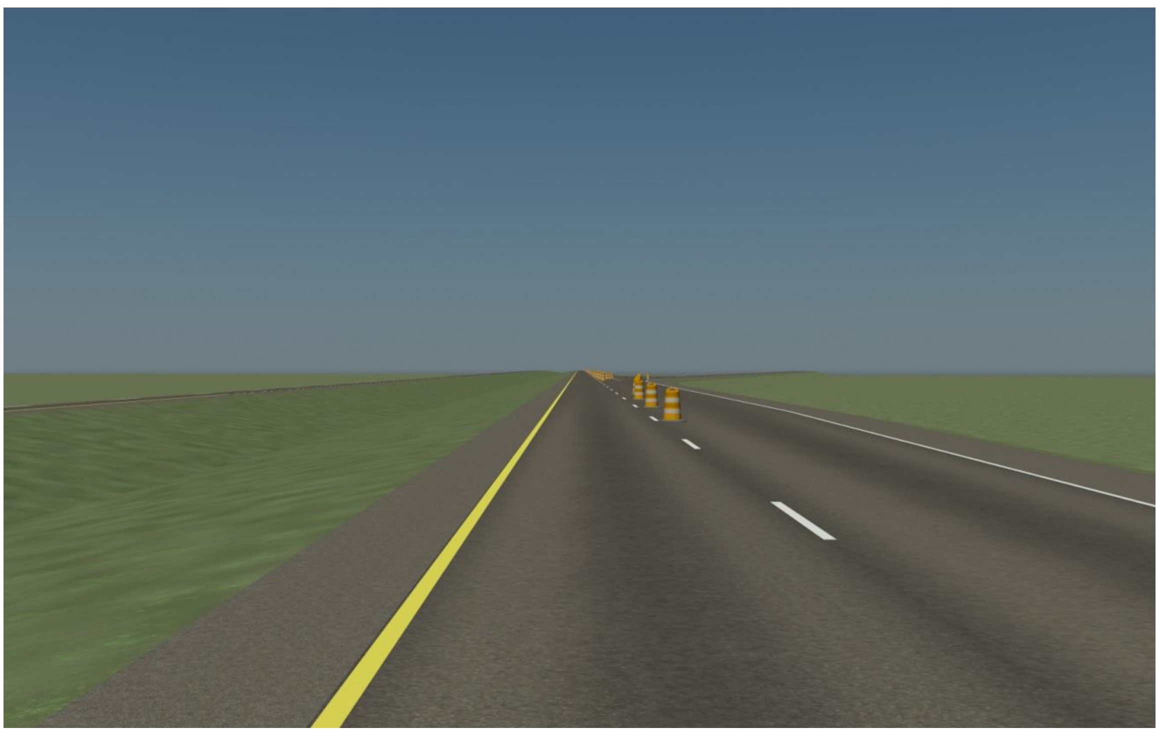

Figure 37: Representative image of the 2s distance. Image from Experiment 1, D40A alternative at 2 second distance in the straight geometry and the open condition.

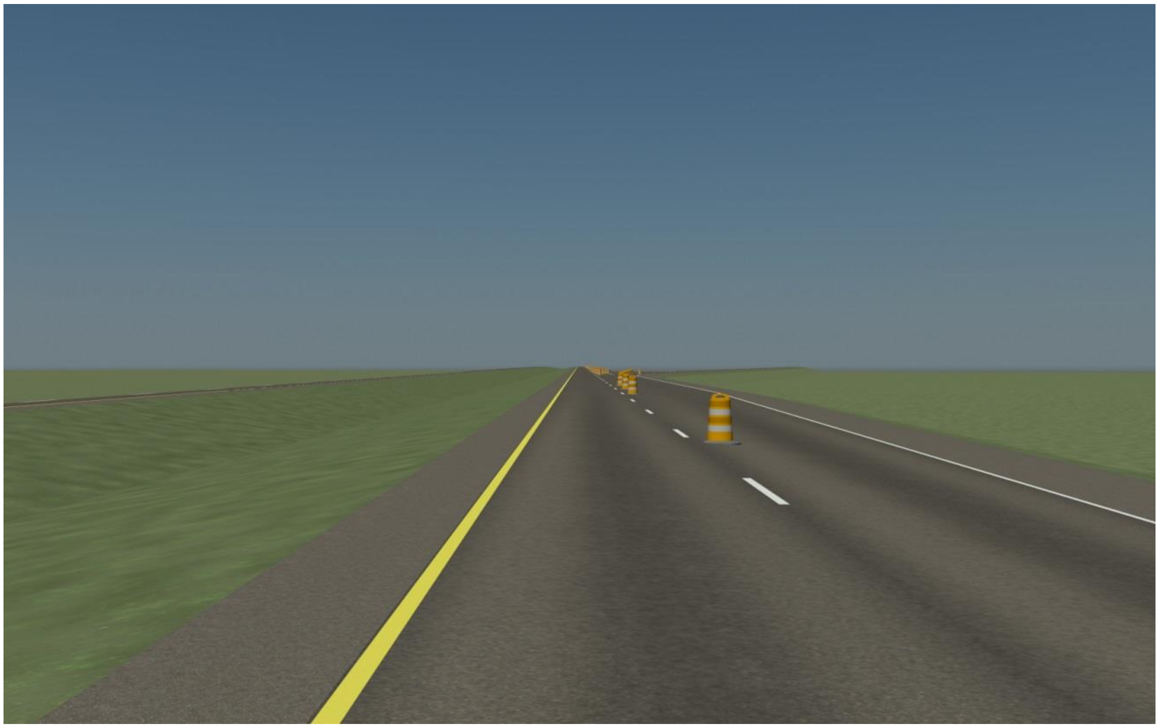

Figure 38: Representative image of the 3 s distance. Image from Experiment 1, D40A alternative at 3 second distance in the straight geometry and the open condition.

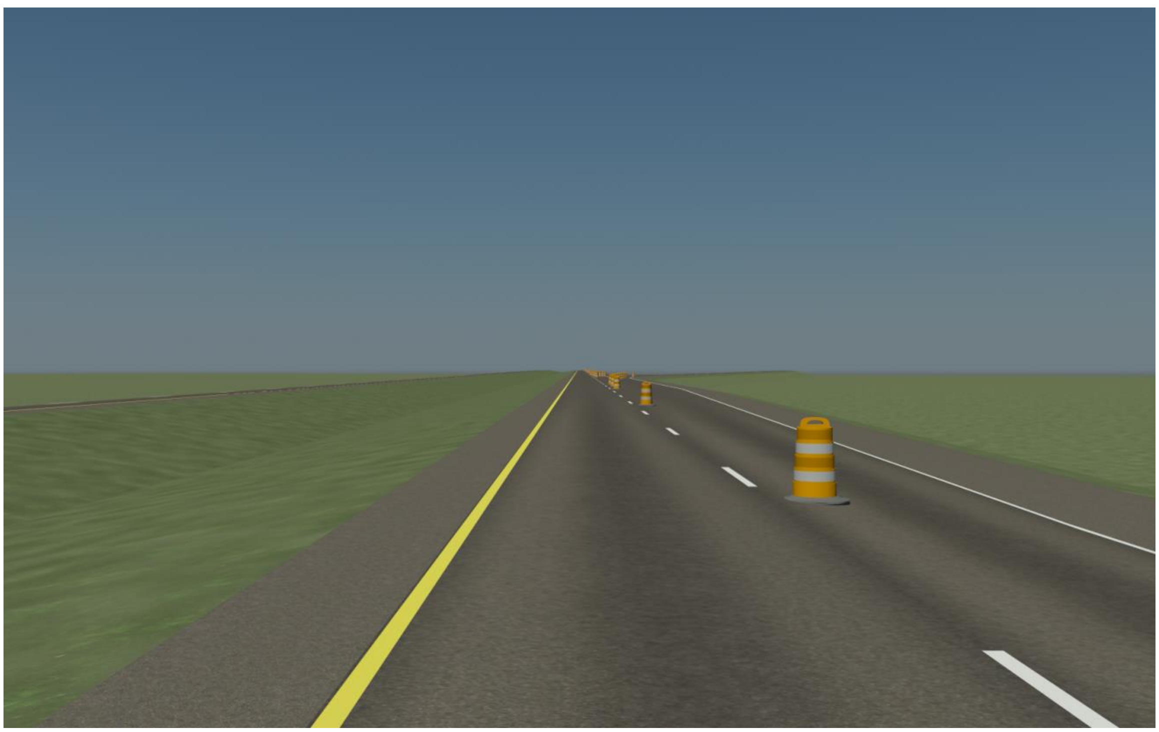

Figure 39: Representative image of the 4 s distance. Image from Experiment 1, D40A alternative at 4 second distance in the straight geometry and the open condition.

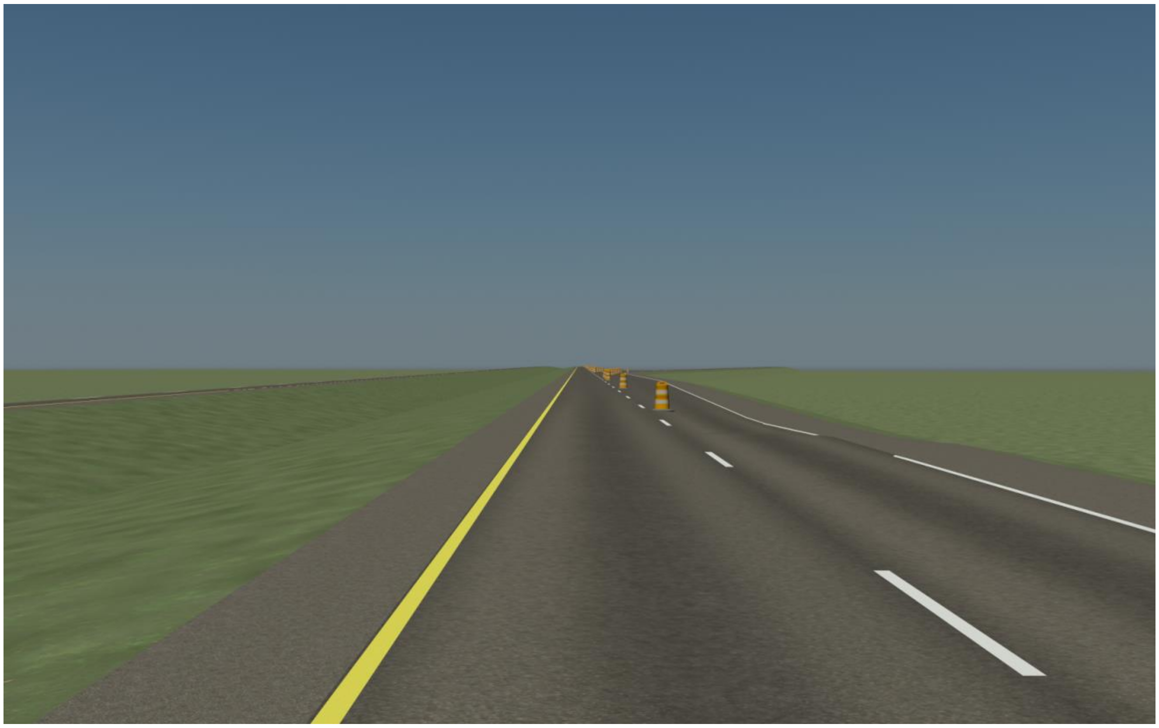

Figure 40: Representative image of the 5 s distance. Image from Experiment 1, D40A alternative at 5 second distance in the straight geometry and the open condition.

Geometries

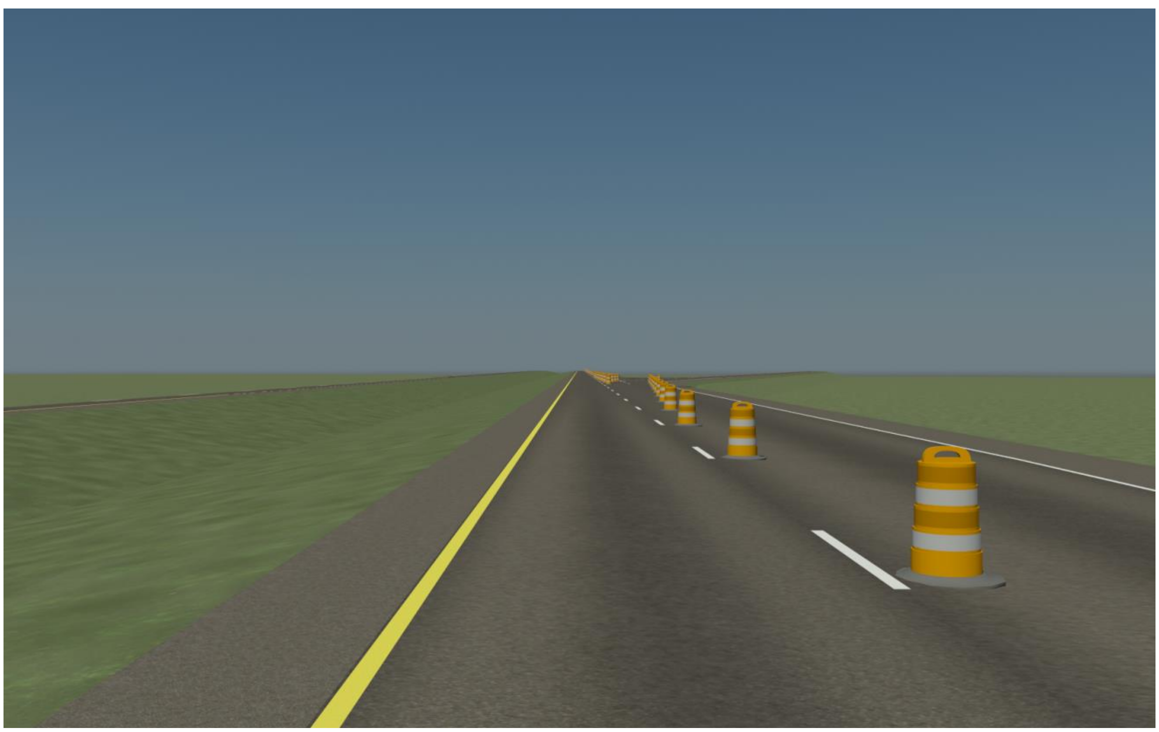

Figure 41: Representative image of the straight geometry. Image from Experiment 1, D40A alternative at 1 second distance in the straight geometry and the open condition.

Figure 42: Representative image of the curved. Image from Experiment 1, D40A alternative at 1 second distance in the curved geometry and the open condition.

Open and Closed Condition

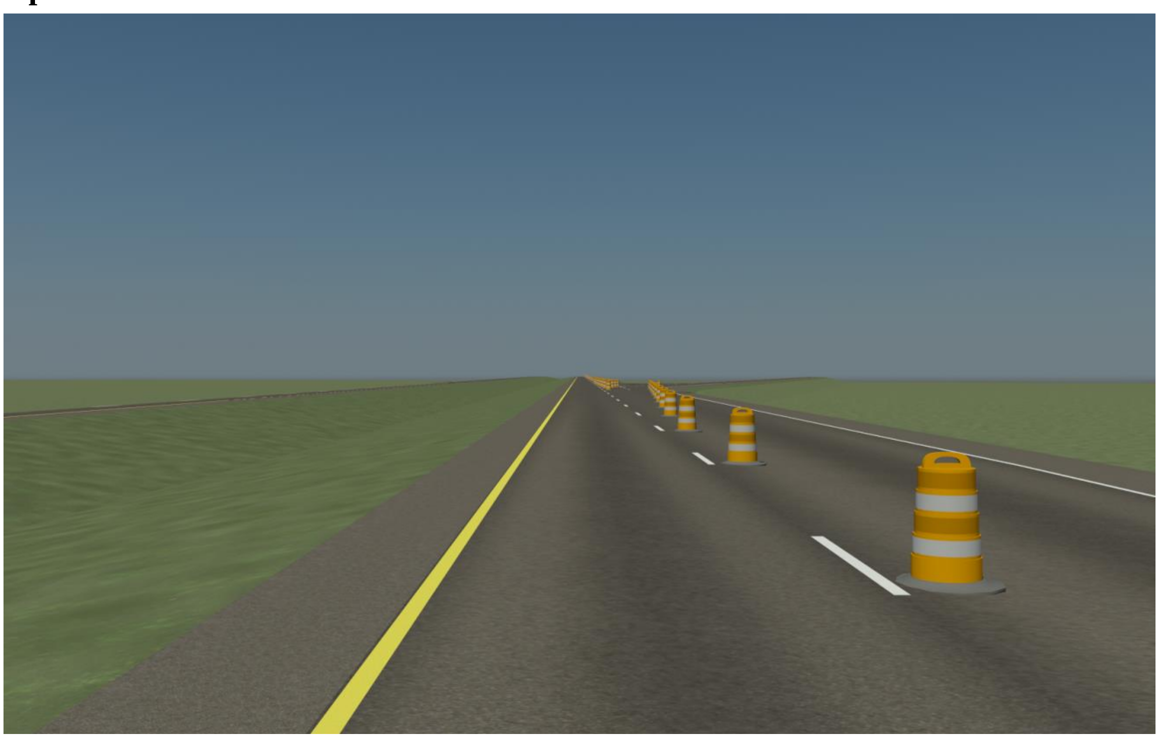

Figure 43: Representative image of the open condition. Image from Experiment 1, D40A alternative at 1 second distance in the straight geometry and the open condition.

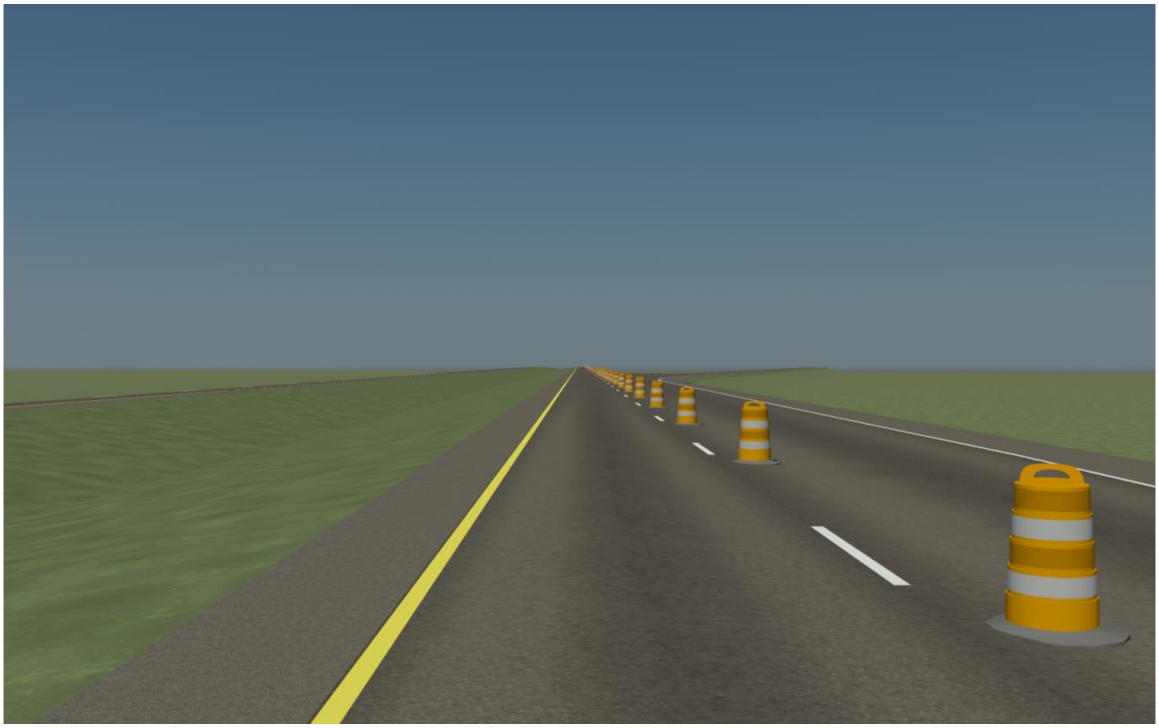

Figure 44: Representative image of the closed. Image from Experiment 1, D40A alternative at 1 second distance in the straight geometry and the closed condition.

Experiment 2

Alternatives

Figure 45: Representative image of the D10A alternative. Image from Experiment 2, D10A alternative at 1 second distance in the straight geometry and the open condition.

Figure 46: Representative image of the D10M alternative. Image from Experiment 2, D40M alternative at 1 second distance in the straight geometry and the open condition.

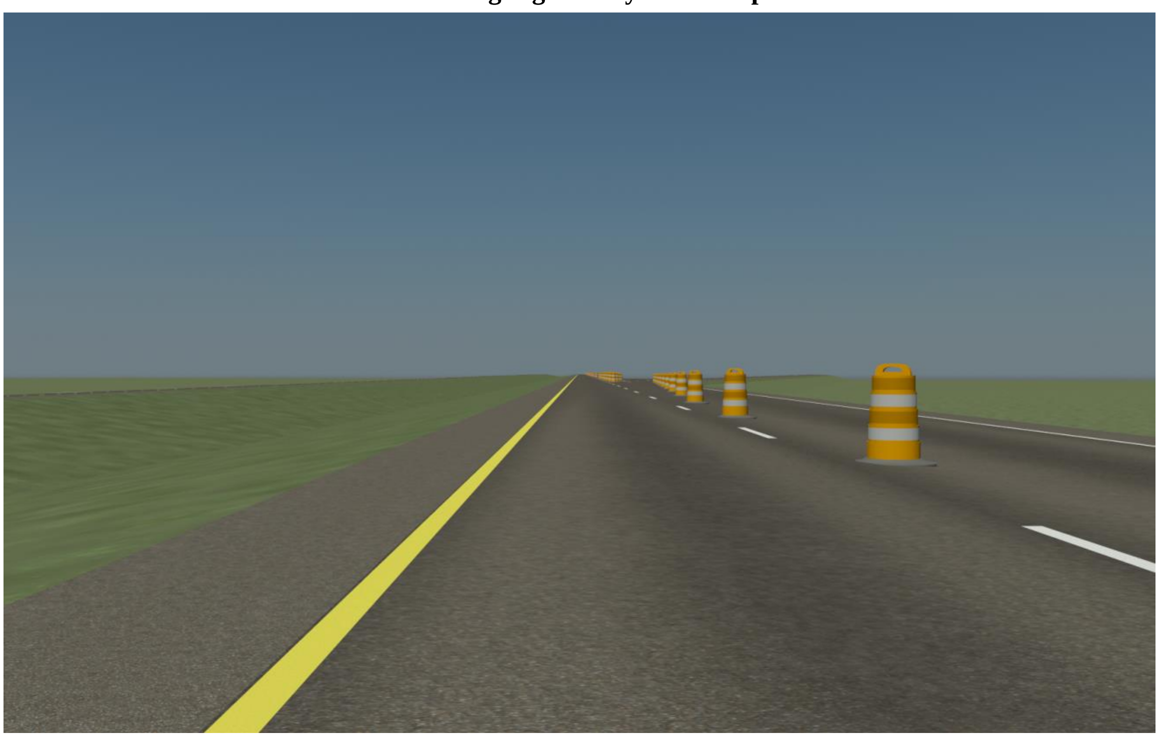

Figure 47: Representative image of the D40A alternative . Image from Experiment 2, D40A alternative at 1 second distance in the straight geometry and the open condition.


Figure 48: Representative image of the D40M alternative . Image from Experiment 2, D40M alternative at 1 second distance in the straight geometry and the open condition.

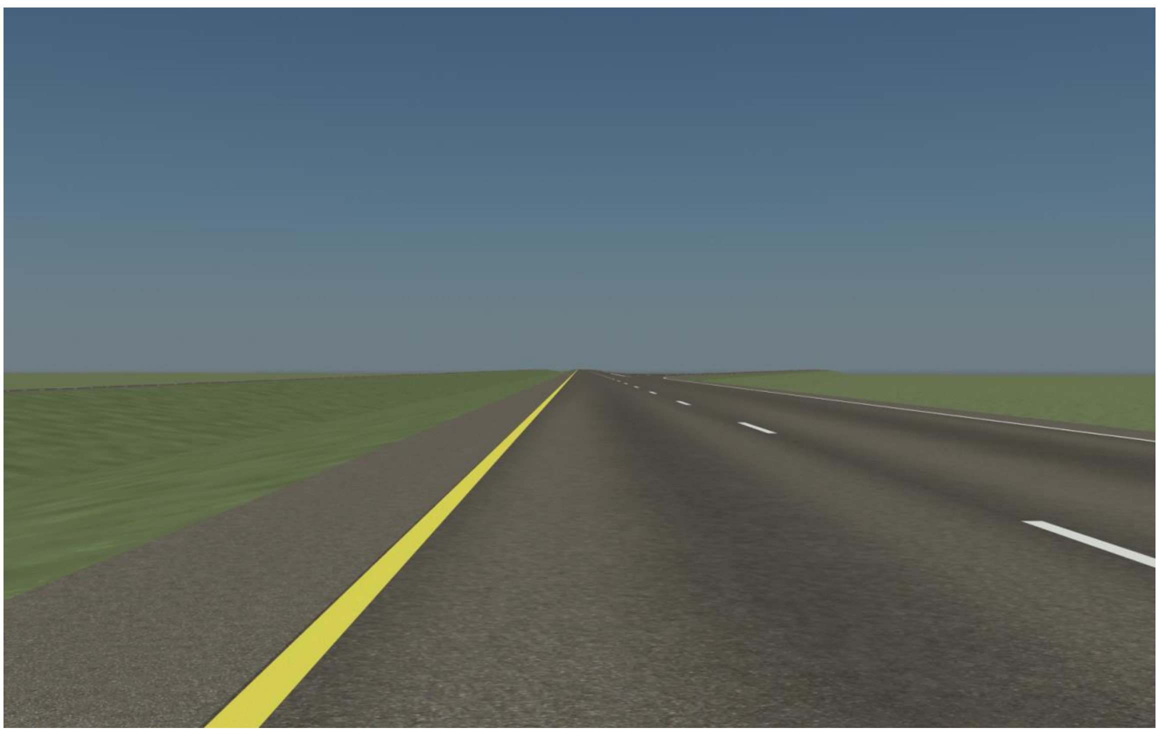
Figure 49: Representative image of the No Work alternative. Image from Experiment 2, No Work alternative at 1 second distance in the straight geometry and the open condition.

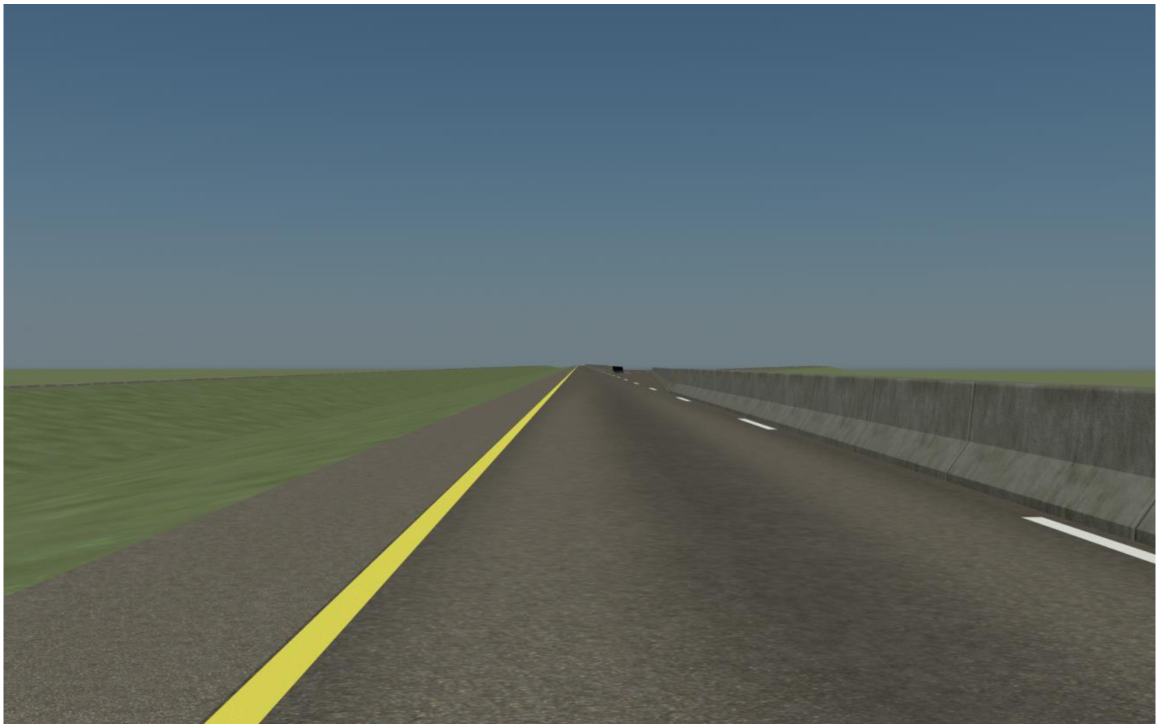

Figure 50: Representative image of the PCB alternative. Image from Experiment 2, PCB alternative at 1 second distance in the straight geometry and the open condition.

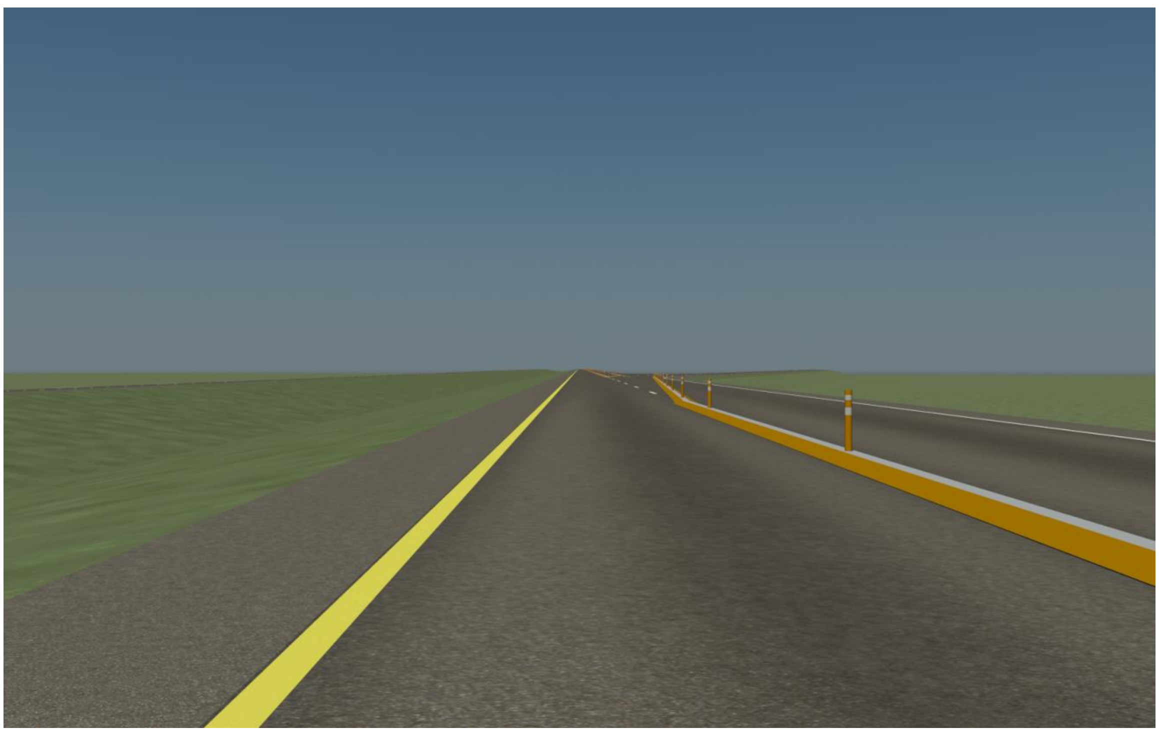

Figure 51: Representative image of the LCD alternative. Image from Experiment 2, LCD alternative at 1 second distance in the straight geometry and the open condition.

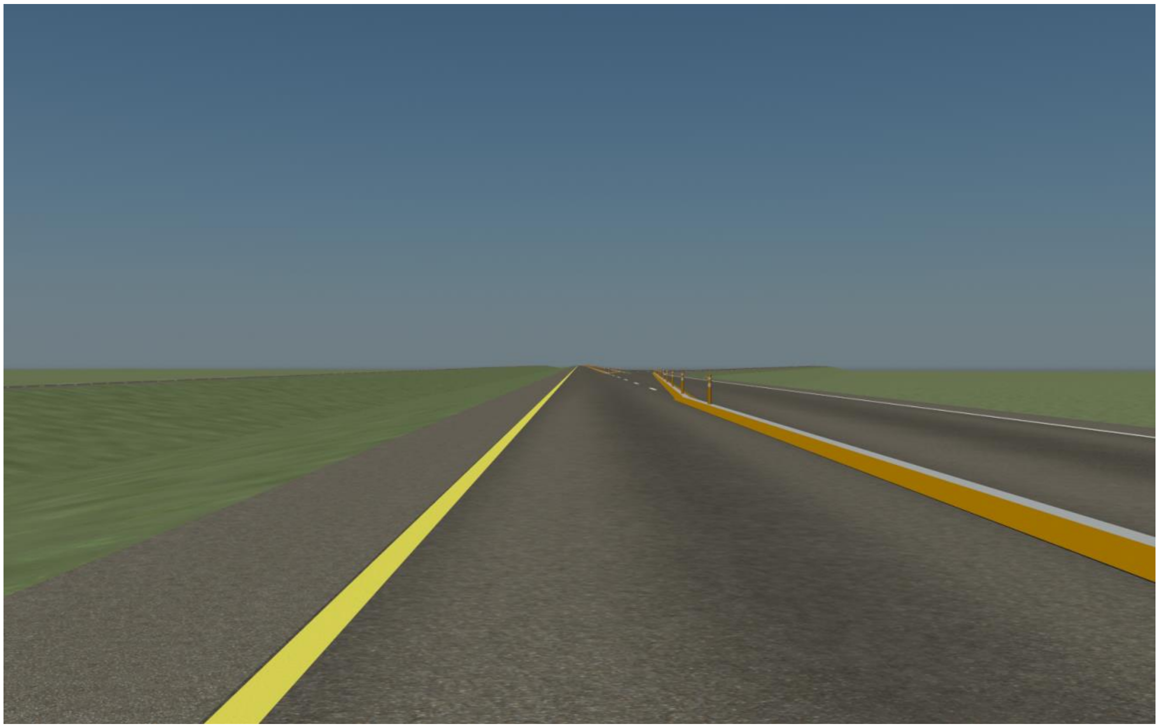

Figure 52: Representative image of the LCD -10% alternative. Image from Experiment 2, LCD -10% alternative at 1 second distance in the straight geometry and the open condition.

Distances & Geometries

Figure 53: Representative image of the 1 s distance for the straight geometry. Image from Experiment 2, D40A alternative at 1 second distance in the straight geometry and the open condition.

Figure 54: Representative image of the 3 s distance for the straight geometry. Image from Experiment 2, D40A alternative at 3 second distance in the straight geometry and the open condition.

Figure 55: Representative image of the 5 s distance for the straight geometry. Image from Experiment 2, D40A alternative at 5 second distance in the straight geometry and the open condition.

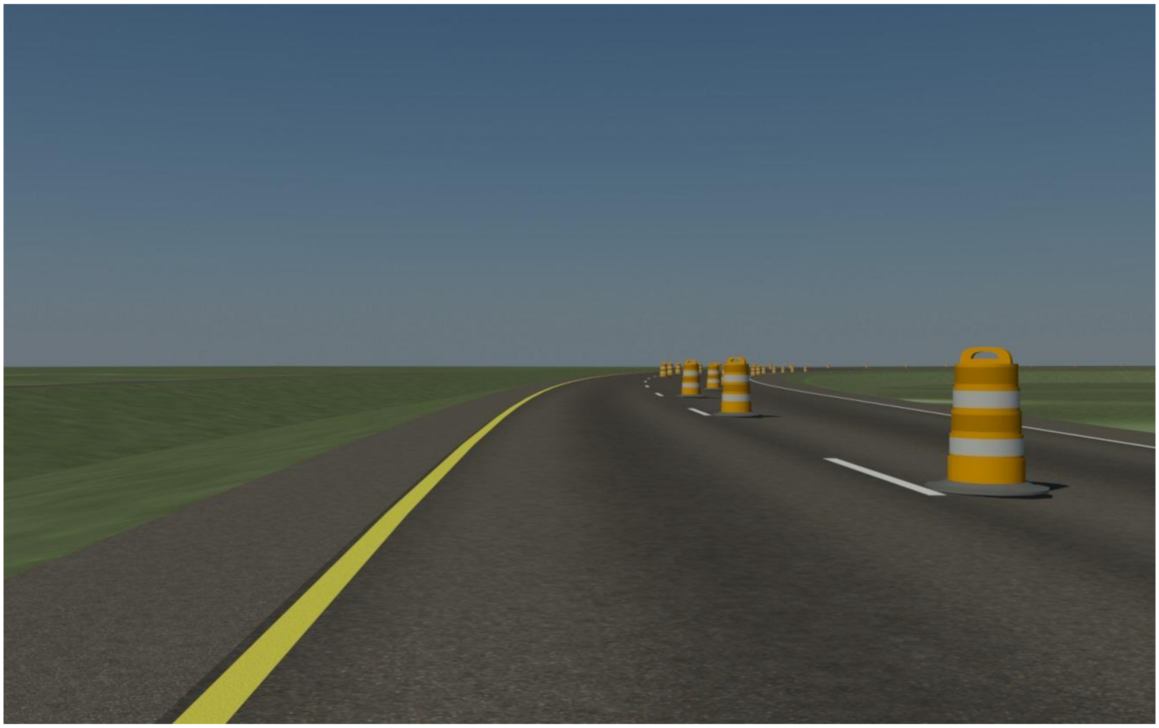

Figure 56: Representative image of the 1 s distance for the curved geometry. Image from Experiment 2, D40A alternative at 1 second distance in the curved geometry and the open condition.

Figure 57: Representative image of the 2 s distance for the curved geometry. Image from Experiment 2, D40A alternative at 2 second distance in the curved geometry and the open condition.

Figure 58: Representative image of the 3 s distance for the curved geometry . Image from Experiment 2, D40A alternative at 3 second distance in the curved geometry and the open condition.

Open and Closed Condition

Figure 59: Representative image of the open condition. Image from Experiment 2, D40A alternative at 1 second distance in the straight geometry and the open condition.

Figure 60: Representative image of the closed condition. Image from Experiment 2, D40A alternative at 1 second distance in the straight geometry and the open condition.

Experiment 3

Alternatives

Figure 61: Representative image of the D40A alternative. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

Figure 62: Representative image of the D40M alternative. Image from Experiment 3, D40M alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

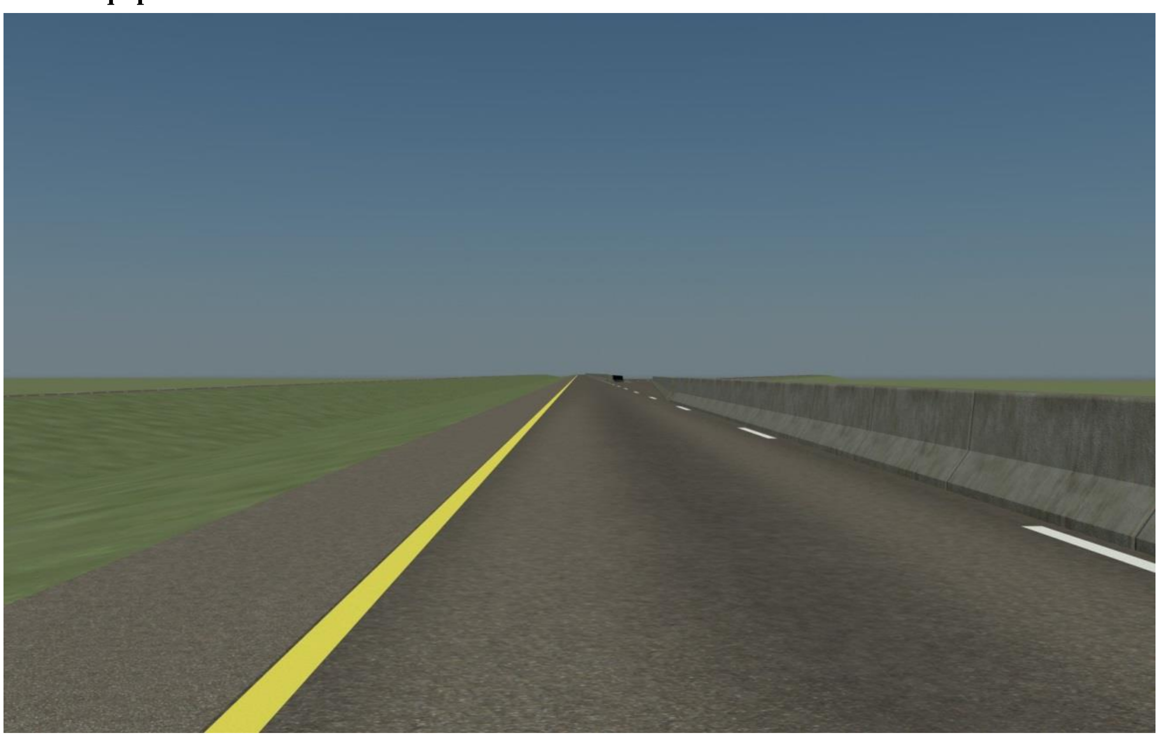

Figure 63: Representative image of the PCB alternative. Image from Experiment 3, PCB alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

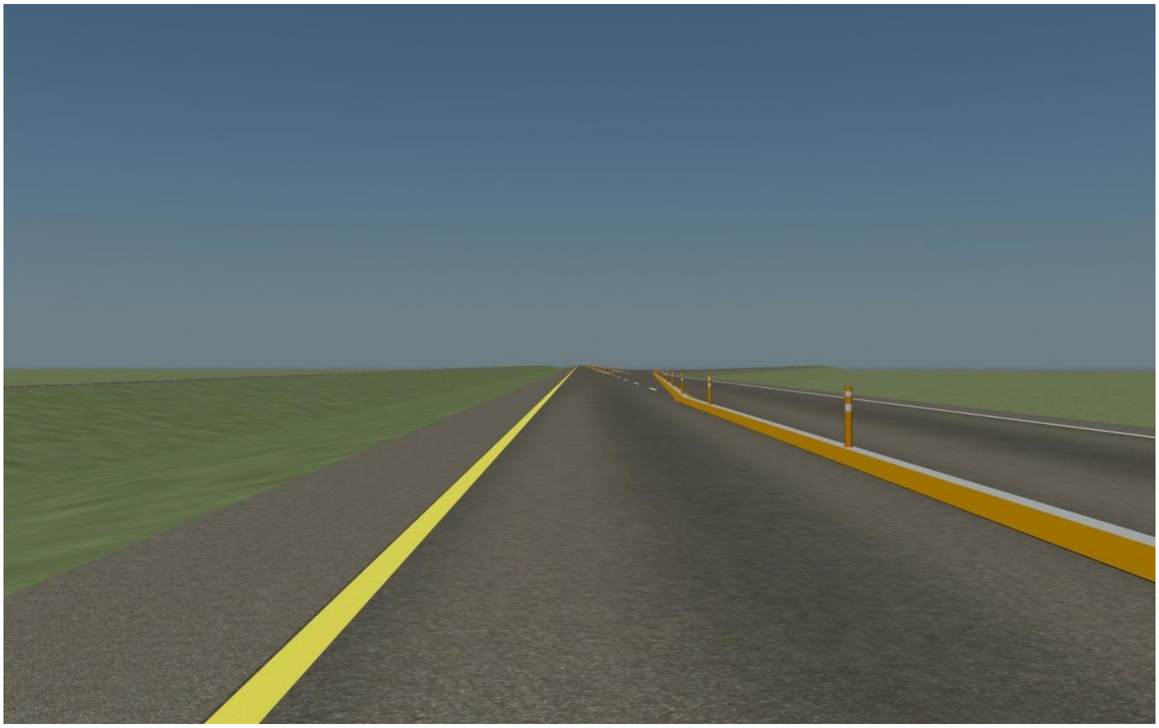

Figure 64: Representative image of the LCD alternative. Image from Experiment 3, LCD alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

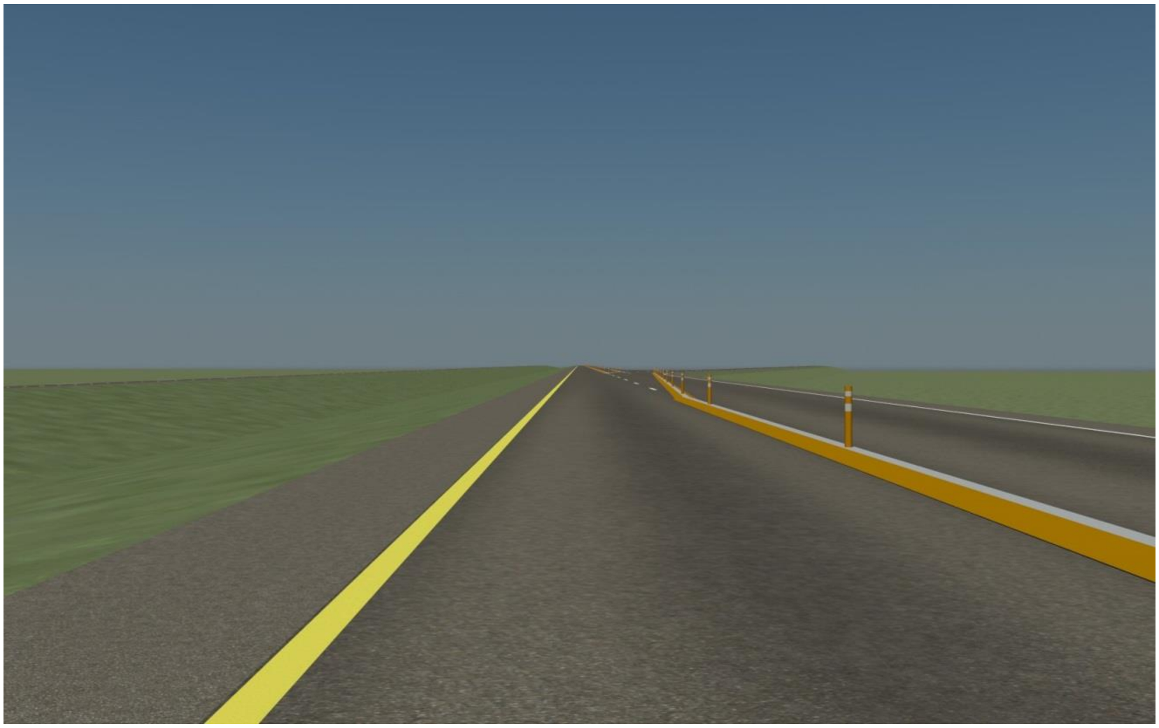

Figure 65: Representative image of the LCD -10% alternative. Image from Experiment 3, LCD-10% alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

Distances

Figure 66: Representative image of the 1 s distance. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

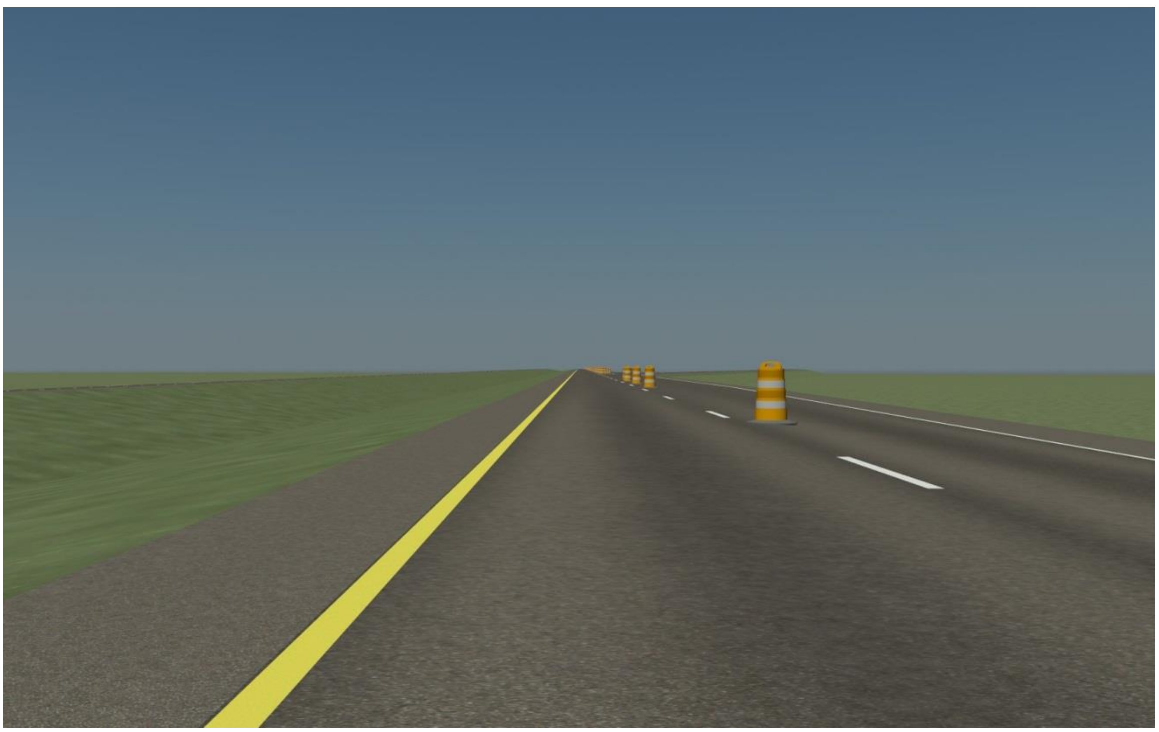

Figure 67: Representative image of the 3 s distance. Image from Experiment 3, D40A alternative at 3 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

Open and Closed Condition

Figure 68: Representative image of the open condition. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

Figure 69: Representative image of the closed condition. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the closed condition with plain vegetation and no equipment.

Equipment

Figure 70: Representative image of equipment configuration A. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and equipment configuration A.

Figure 71: Representative image of equipment configuration B. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and equipment configuration B.

Vegetation

Figure 72: Representative image of plain vegetation. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with plain vegetation and no equipment.

Figure 73: Representative image of trees on both sides. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with trees on both sides and no equipment.

Figure 74: Representative image of light vegetation. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with light vegetation and no equipment.

Figure 75: Representative image of trees in the median. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with trees in the median and no equipment.

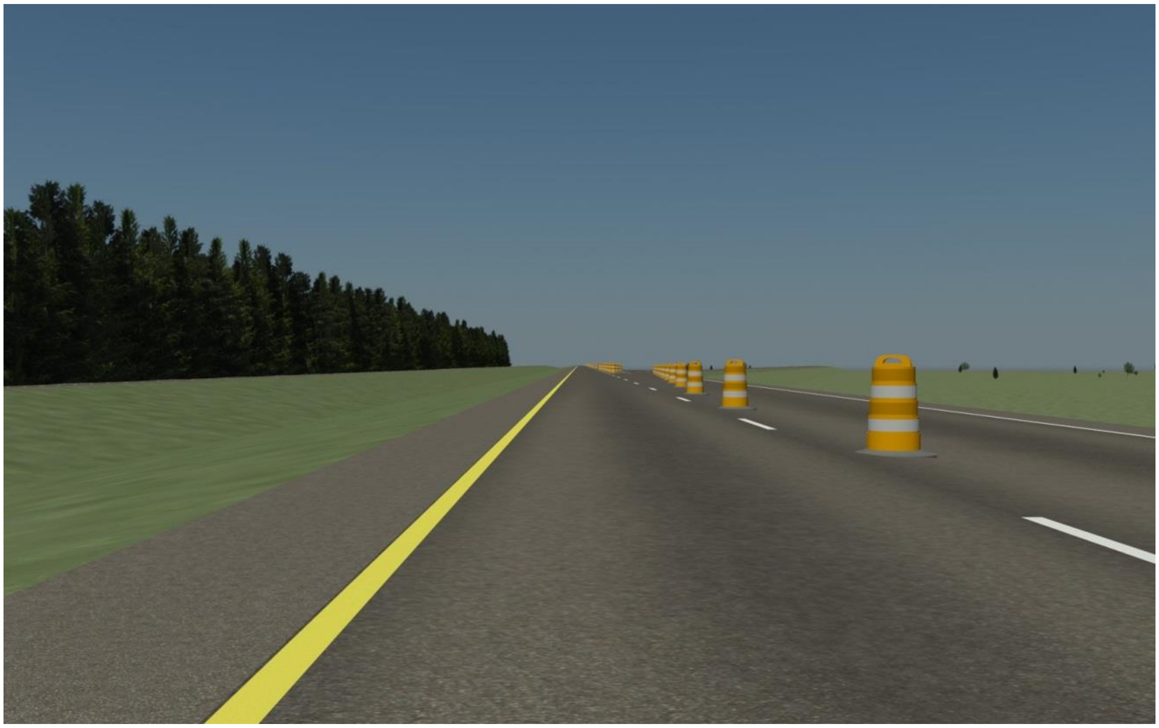

Figure 76: Representative image of trees on the left. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with trees on the leftand no equipment.

Figure 77: Representative image of trees on the right. Image from Experiment 3, D40A alternative at 1 second distance in the straight geometry and the open condition with trees on the right and no equipment.

APPENDIX B: COMPUTED PROBABILITY OF RESPONSE CATEGORY

These tables show the results of integrating the bivariate normal distributions fit to three representative participants' points described in Chapter 6. These distributions were fitted based on response to a stimulus with low error rates, although future researchers should use a calibration test at regular intervals in the experiment. The results can be interpreted as the likelihood that the given response was intended for a particular categorization zone (Closed, Work Zone, or Ramp). Note that due to the mutually exclusive and collectively exhaustive nature of the zones, that while the integral over the Indeterminate zone was not computed, it is implicitly the remaining probability not computed for the other three zones. Also note that in the open condition, the correct response is to click in the ramp zone, and that in the closed condition, the correct response is to lick in the closed zone.

Table 15: Response probabilities for a participant with a wide point distribution for the D10A alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	204	105	D10A	Closed	Closed	0.00%	0.00%	96.92%
Wide	265	71	D10A	Closed	Closed	0.00%	0.00%	89.69%
Wide	320	74	D10A	Closed	Closed	0.00%	0.00%	90.61%
Wide	328	83	D10A	Closed	Closed	0.00%	0.00%	93.02%
Wide	249	110	D10A	Closed	Closed	0.00%	0.00%	97.49%
Wide	472	106	D10A	Closed	Indeterminate	0.00%	0.00%	8.08%
Wide	185	87	D10A	Closed	Closed	0.00%	0.00%	93.93%
Wide	338	105	D10A	Closed	Closed	0.00%	0.00%	96.92%
Wide	332	126	D10A	Closed	Closed	0.00%	0.00%	98.76%
Wide	314	99	D10A	Closed	Closed	0.00%	0.00%	96.10%

Table 16: Response probabilities for a participant with a wide point distribution for the D40A alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	214	112	D40A	Closed	Closed	0.00%	0.00%	97.69%
Wide	324	44	D40A	Closed	Closed	0.00%	0.00%	78.33%
Wide	277	98	D40A	Closed	Closed	0.00%	0.00%	95.95%
Wide	228	151	D40A	Closed	Closed	0.00%	0.00%	99.64%
Wide	169	93	D40A	Closed	Closed	0.00%	0.00%	95.11%
Wide	306	83	D40A	Closed	Closed	0.00%	0.00%	93.02%
Wide	252	96	D40A	Closed	Closed	0.00%	0.00%	95.63%
Wide	152	92	D40A	Closed	Closed	0.00%	0.00%	94.93%
Wide	319	74	D40A	Closed	Closed	0.00%	0.00%	90.61%
Wide	330	131	D40A	Closed	Closed	0.00%	0.00%	99.02%

Table 17: Response probabilities for a participant with a wide point distribution for the D40M alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	217	101	D40M	Closed	Closed	0.00%	0.00%	96.39%
Wide	488	120	D40M	Closed	Indeterminate	0.00%	0.00%	1.24%
Wide	469	108	D40M	Closed	Indeterminate	0.00%	0.00%	10.87%
Wide	288	81	D40M	Closed	Closed	0.00%	0.00%	92.53%
Wide	344	85	D40M	Closed	Closed	0.00%	0.00%	93.49%
Wide	374	108	D40M	Closed	Closed	0.00%	0.00%	97.27%
Wide	444	115	D40M	Closed	Closed	0.00%	0.00%	54.60%
Wide	290	94	D40M	Closed	Closed	0.00%	0.00%	95.29%
Wide	161	82	D40M	Closed	Closed	0.00%	0.00%	92.78%
Wide	311	113	D40M	Closed	Closed	0.00%	0.00%	97.79%

Table 18: Response probabilities for a participant with a wide point distribution for the PCB alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	351	131	PCB	Closed	Closed	0.00%	0.00%	99.02%
Wide	271	93	PCB	Closed	Closed	0.00%	0.00%	95.11%
Wide	417	108	PCB	Closed	Closed	0.00%	0.00%	91.90%
Wide	227	99	PCB	Closed	Closed	0.00%	0.00%	96.10%
Wide	508	104	PCB	Closed	Indeterminate	0.00%	0.00%	0.04%
Wide	358	93	PCB	Closed	Closed	0.00%	0.00%	95.11%
Wide	307	71	PCB	Closed	Closed	0.00%	0.00%	89.69%
Wide	268	101	PCB	Closed	Closed	0.00%	0.00%	96.39%
Wide	182	130	PCB	Closed	Closed	0.00%	0.00%	98.97%
Wide	193	68	PCB	Closed	Closed	0.00%	0.00%	88.70%

Table 19: Response probabilities for a participant with a wide point distribution for the D10A alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	871	505	D10A	Open	Ramp	38.11%	0.50%	0.00%
Wide	874	491	D10A	Open	Ramp	44.42%	0.89%	0.00%
Wide	905	522	D10A	Open	Ramp	44.26%	13.08%	0.00%
Wide	864	505	D10A	Open	Indeterminate	26.90%	0.17%	0.00%
Wide	844	406	D10A	Open	Indeterminate	6.15%	0.01%	0.00%
Wide	834	98	D10A	Open	Indeterminate	0.00%	0.00%	0.00%
Wide	889	529	D10A	Open	Ramp	51.19%	3.14%	0.00%
Wide	883	532	D10A	Open	Ramp	46.07%	1.51%	0.00%
Wide	889	493	D10A	Open	Ramp	53.62%	4.53%	0.00%
Wide	903	523	D10A	Open	Ramp	46.57%	11.34%	0.00%

Table 20: Response probabilities for a participant with a wide point distribution for the D40A alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	886	505	D40A	Open	Ramp	54.12%	3.14%	0.00%
Wide	879	484	D40A	Open	Ramp	49.68%	1.70%	0.00%
Wide	870	403	D40A	Open	Indeterminate	19.79%	0.32%	0.00%
Wide	869	456	D40A	Open	Ramp	35.82%	0.52%	0.00%
Wide	859	458	D40A	Open	Indeterminate	23.14%	0.12%	0.00%
Wide	883	464	D40A	Open	Ramp	46.69%	2.59%	0.00%
Wide	887	477	D40A	Open	Ramp	50.43%	3.87%	0.00%
Wide	877	482	D40A	Open	Ramp	47.80%	1.36%	0.00%
Wide	870	440	D40A	Open	Ramp	32.91%	0.55%	0.00%
Wide	916	493	D40A	Open	WZ	21.84%	23.83%	0.00%

Table 21: Response probabilities for a participant with a wide point distribution for the D40M alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	895	498	D40M	Open	Ramp	51.38%	7.39%	0.00%
Wide	877	455	D40M	Open	Ramp	42.18%	1.35%	0.00%
Wide	868	479	D40M	Open	Ramp	36.37%	0.43%	0.00%
Wide	862	463	D40M	Open	Indeterminate	27.43%	0.19%	0.00%
Wide	864	426	D40M	Open	Indeterminate	24.06%	0.23%	0.00%
Wide	858	479	D40M	Open	Indeterminate	21.28%	0.09%	0.00%
Wide	892	476	D40M	Open	Ramp	47.39%	5.88%	0.00%
Wide	871	435	D40M	Open	Ramp	32.01%	0.59%	0.00%
Wide	850	454	D40M	Open	Indeterminate	12.39%	0.03%	0.00%
Wide	897	495	D40M	Open	WZ	48.86%	8.67%	0.00%

Table 22: Response probabilities for a participant with a wide point distribution for the No Work alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	902	523	No Work	Open	Ramp	47.60%	10.57%	0.00%
Wide	902	464	No Work	Open	Indeterminate	30.48%	10.39%	0.00%
Wide	936	475	No Work	Open	WZ	2.61%	31.89%	0.00%
Wide	915	524	No Work	Open	Ramp	31.00%	22.33%	0.00%
Wide	844	422	No Work	Open	Indeterminate	6.92%	0.01%	0.00%
Wide	876	510	No Work	Open	Ramp	44.41%	0.92%	0.00%
Wide	915	512	No Work	Open	Ramp	28.57%	23.43%	0.00%
Wide	913	516	No Work	Open	Ramp	32.39%	21.17%	0.00%
Wide	879	489	No Work	Open	Ramp	49.97%	1.66%	0.00%
Wide	913	557	No Work	Open	Indeterminate	33.49%	13.63%	0.00%

Table 23: Response probabilities for a participant with a wide point distribution for the PCB alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Wide	879	503	PCB	Open	Ramp	49.30%	1.47%	0.00%
Wide	905	568	PCB	Open	Indeterminate	36.01%	6.19%	0.00%
Wide	905	532	PCB	Open	Ramp	44.40%	11.82%	0.00%
Wide	854	519	PCB	Open	Indeterminate	11.09%	0.02%	0.00%
Wide	895	460	PCB	Open	Indeterminate	37.42%	6.59%	0.00%
Wide	877	456	PCB	Open	Ramp	42.50%	1.36%	0.00%
Wide	877	495	PCB	Open	Ramp	47.86%	1.25%	0.00%
Wide	854	540	PCB	Open	Indeterminate	7.92%	0.01%	0.00%
Wide	856	359	PCB	Open	Indeterminate	5.35%	0.02%	0.00%
Wide	868	459	PCB	Open	Ramp	35.13%	0.46%	0.00%

Table 24: Response probabilities for a participant with a medium point distribution for the D10A alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	130	79	D10A	Closed	Closed	0.00%	0.00%	99.98%
Medium	154	137	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	200	103	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	221	104	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	236	102	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	180	113	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	159	99	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	134	89	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	152	92	D10A	Closed	Closed	0.00%	0.00%	100.00%
Medium	211	119	D10A	Closed	Closed	0.00%	0.00%	100.00%

Table 25: Response probabilities for a participant with a medium point distribution for the D40A alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	973	441	D40A	Closed	WZ	0.01%	4.56%	0.00%
Medium	251	135	D40A	Closed	Closed	0.00%	0.00%	100.00%
Medium	242	102	D40A	Closed	Closed	0.00%	0.00%	100.00%
Medium	157	84	D40A	Closed	Closed	0.00%	0.00%	99.99%
Medium	178	93	D40A	Closed	Closed	0.00%	0.00%	100.00%
Medium	262	77	D40A	Closed	Closed	0.00%	0.00%	99.97%
Medium	139	95	D40A	Closed	Closed	0.00%	0.00%	100.00%
Medium	232	113	D40A	Closed	Closed	0.00%	0.00%	100.00%
Medium	129	99	D40A	Closed	Closed	0.00%	0.00%	100.00%
Medium	191	120	D40A	Closed	Closed	0.00%	0.00%	100.00%

Table 26: Response probabilities for a participant with a medium point distribution for the D40M alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	923	505	D40M	Closed	WZ	21.51%	48.86%	0.00%
Medium	180	103	D40M	Closed	Closed	0.00%	0.00%	100.00%
Medium	218	121	D40M	Closed	Closed	0.00%	0.00%	100.00%
Medium	174	96	D40M	Closed	Closed	0.00%	0.00%	100.00%
Medium	216	111	D40M	Closed	Closed	0.00%	0.00%	100.00%
Medium	170	44	D40M	Closed	Closed	0.00%	0.00%	97.47%
Medium	187	113	D40M	Closed	Closed	0.00%	0.00%	100.00%
Medium	137	68	D40M	Closed	Closed	0.00%	0.00%	99.87%
Medium	227	105	D40M	Closed	Closed	0.00%	0.00%	100.00%
Medium	183	107	D40M	Closed	Closed	0.00%	0.00%	100.00%

Table 27: Response probabilities for a participant with a medium point distribution for the PCB alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	232	85	PCB	Closed	Closed	0.00%	0.00%	99.99%
Medium	59	87	PCB	Closed	Closed	0.00%	0.00%	99.84%
Medium	159	94	PCB	Closed	Closed	0.00%	0.00%	100.00%
Medium	216	126	PCB	Closed	Closed	0.00%	0.00%	100.00%
Medium	143	82	PCB	Closed	Closed	0.00%	0.00%	99.99%
Medium	155	49	PCB	Closed	Closed	0.00%	0.00%	98.52%
Medium	260	120	PCB	Closed	Closed	0.00%	0.00%	100.00%
Medium	208	121	PCB	Closed	Closed	0.00%	0.00%	100.00%
Medium	124	103	PCB	Closed	Closed	0.00%	0.00%	100.00%
Medium	171	92	PCB	Closed	Closed	0.00%	0.00%	100.00%

Table 28: Response probabilities for a participant with a medium point distribution for the D10A alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	910	517	D10A	Open	Ramp	45.90%	36.15%	0.00%
Medium	898	433	D10A	Open	Indeterminate	24.10%	0.04%	0.00%
Medium	950	510	D10A	Open	Indeterminate	3.41%	81.02%	0.00%
Medium	934	478	D10A	Open	WZ	5.20%	24.27%	0.00%
Medium	885	493	D10A	Open	Ramp	57.40%	4.16%	0.00%
Medium	922	512	D10A	Open	WZ	26.16%	52.76%	0.00%
Medium	872	515	D10A	Open	Ramp	46.75%	2.11%	0.00%
Medium	873	466	D10A	Open	Ramp	46.84%	0.18%	0.00%
Medium	891	487	D10A	Open	Ramp	54.78%	5.12%	0.00%
Medium	962	534	D10A	Open	Indeterminate	2.46%	89.28%	0.00%

Table 29: Response probabilities for a participant with a medium point distribution for the D40A alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	891	526	D40A	Open	Ramp	63.67%	12.07%	0.00%
Medium	892	517	D40A	Open	Ramp	62.08%	13.09%	0.00%
Medium	882	505	D40A	Open	Ramp	57.78%	4.84%	0.00%
Medium	898	485	D40A	Open	WZ	47.94%	7.63%	0.00%
Medium	862	492	D40A	Open	Indeterminate	33.05%	0.30%	0.00%
Medium	914	516	D40A	Open	Ramp	39.63%	42.16%	0.00%
Medium	902	517	D40A	Open	Ramp	55.80%	24.54%	0.00%
Medium	914	482	D40A	Open	WZ	24.70%	15.49%	0.00%
Medium	905	500	D40A	Open	WZ	44.40%	21.61%	0.00%
Medium	917	507	D40A	Open	WZ	30.43%	42.22%	0.00%

Table 30: Response probabilities for a participant with a medium point distribution for the D40M alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	893	488	D40M	Open	Ramp	53.69%	6.29%	0.00%
Medium	881	497	D40M	Open	Ramp	56.43%	3.41%	0.00%
Medium	916	506	D40M	Open	WZ	31.35%	40.05%	0.00%
Medium	914	466	D40M	Open	Indeterminate	20.37%	5.17%	0.00%
Medium	909	583	D40M	Open	Indeterminate	18.16%	5.13%	0.00%
Medium	987	478	D40M	Open	WZ	0.00%	56.90%	0.00%
Medium	874	489	D40M	Open	Ramp	49.97%	1.13%	0.00%
Medium	909	520	D40M	Open	Ramp	48.83%	34.94%	0.00%
Medium	889	570	D40M	Open	Indeterminate	34.18%	2.40%	0.00%
Medium	970	528	D40M	Open	Indeterminate	0.88%	93.33%	0.00%

Table 31: Response probabilities for a participant with a medium point distribution for the No Work alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	934	426	No Work	Open	Indeterminate	1.53%	0.15%	0.00%
Medium	916	485	No Work	Open	WZ	22.93%	19.70%	0.00%
Medium	918	499	No Work	Open	WZ	25.32%	36.44%	0.00%
Medium	908	435	No Work	Open	Indeterminate	16.64%	0.11%	0.00%
Medium	881	410	No Work	Open	Indeterminate	10.93%	0.00%	0.00%
Medium	934	511	No Work	Open	Indeterminate	12.55%	67.81%	0.00%
Medium	937	498	No Work	Open	WZ	6.77%	56.43%	0.00%
Medium	892	531	No Work	Open	Ramp	64.01%	12.31%	0.00%
Medium	879	507	No Work	Open	Ramp	55.55%	3.82%	0.00%
Medium	898	482	No Work	Open	Indeterminate	47.01%	6.42%	0.00%

Table 32: Response probabilities for a participant with a medium point distribution for the PCB alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Medium	883	487	PCB	Open	Ramp	56.16%	2.56%	0.00%
Medium	888	507	PCB	Open	Ramp	60.11%	8.49%	0.00%
Medium	904	491	PCB	Open	WZ	42.19%	14.79%	0.00%
Medium	940	471	PCB	Open	WZ	2.41%	18.80%	0.00%
Medium	873	520	PCB	Open	Ramp	47.86%	2.43%	0.00%
Medium	898	470	PCB	Open	Indeterminate	43.19%	2.75%	0.00%
Medium	928	533	PCB	Open	Ramp	28.28%	61.27%	0.00%
Medium	889	488	PCB	Open	Ramp	55.97%	4.59%	0.00%
Medium	916	462	PCB	Open	Indeterminate	17.03%	4.05%	0.00%
Medium	912	516	PCB	Open	Ramp	42.52%	39.04%	0.00%

Table 33: Response probabilities for a participant with a tight point distribution for the D10A alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	977	456	D10A	Closed	WZ	0.00%	0.00%	0.00%
Tight	1159	432	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1180	441	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1124	459	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1195	421	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1226	420	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1175	442	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1145	430	D10A	Closed	WZ	0.00%	99.93%	0.00%
Tight	1194	439	D10A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1129	458	D10A	Closed	WZ	0.00%	100.00%	0.00%

Table 34: Response probabilities for a participant with a tight point distribution for the D40A alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	1185	382	D40A	Closed	WZ	0.00%	0.00%	0.00%
Tight	1149	440	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1121	447	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1127	448	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1170	416	D40A	Closed	WZ	0.00%	94.54%	0.00%
Tight	1170	440	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1174	433	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1181	437	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1166	444	D40A	Closed	WZ	0.00%	100.00%	0.00%
Tight	1135	445	D40A	Closed	WZ	0.00%	100.00%	0.00%

Table 35: Response probabilities for a participant with a tight point distribution for the D40M alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	1118	385	D40M	Closed	WZ	0.00%	0.00%	0.00%
Tight	1121	437	D40M	Closed	WZ	0.00%	99.75%	0.00%
Tight	1103	439	D40M	Closed	WZ	0.00%	93.77%	0.00%
Tight	1133	423	D40M	Closed	WZ	0.00%	44.11%	0.00%
Tight	199	96	D40M	Closed	Closed	0.00%	0.00%	100.00%
Tight	1149	449	D40M	Closed	WZ	0.00%	100.00%	0.00%
Tight	1151	428	D40M	Closed	WZ	0.00%	99.94%	0.00%
Tight	1142	456	D40M	Closed	WZ	0.00%	100.00%	0.00%
Tight	1125	456	D40M	Closed	WZ	0.00%	100.00%	0.00%
Tight	1098	459	D40M	Closed	WZ	0.00%	100.00%	0.00%

Table 36: Response probabilities for a participant with a tight point distribution for the PCB alternative in the closed condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	223	104	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	154	110	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	197	90	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	223	105	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	235	115	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	226	107	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	214	101	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	220	105	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	239	97	PCB	Closed	Closed	0.00%	0.00%	100.00%
Tight	233	109	PCB	Closed	Closed	0.00%	0.00%	100.00%

Table 37: Response probabilities for a participant with a tight point distribution for the D10A alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	882	504	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	879	500	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	882	511	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	506	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	507	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	878	510	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	884	507	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	509	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	883	506	D10A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	512	D10A	Open	Ramp	100.00%	0.00%	0.00%

Table 38: Response probabilities for a participant with a tight point distribution for the D40A alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	886	505	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	506	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	882	501	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	882	508	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	883	505	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	510	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	880	510	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	886	506	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	882	507	D40A	Open	Ramp	100.00%	0.00%	0.00%
Tight	881	509	D40A	Open	Ramp	100.00%	0.00%	0.00%

Table 39: Response probabilities for a participant with a tight point distribution for the D40M alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	968	458	D40M	Open	WZ	0.00%	0.00%	0.00%
Tight	873	508	D40M	Open	Ramp	91.26%	0.00%	0.00%
Tight	877	506	D40M	Open	Ramp	99.99%	0.00%	0.00%
Tight	872	506	D40M	Open	Ramp	81.43%	0.00%	0.00%
Tight	875	508	D40M	Open	Ramp	99.28%	0.00%	0.00%
Tight	876	505	D40M	Open	Ramp	99.91%	0.00%	0.00%
Tight	874	509	D40M	Open	Ramp	96.87%	0.00%	0.00%
Tight	872	508	D40M	Open	Ramp	79.15%	0.00%	0.00%
Tight	876	508	D40M	Open	Ramp	99.86%	0.00%	0.00%
Tight	876	509	D40M	Open	Ramp	99.84%	0.00%	0.00%

Table 40: Response probabilities for a participant with a tight point distribution for the No Work alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	906	511	No Work	Open	Ramp	26.48%	73.35%	0.00%
Tight	915	514	No Work	Open	Ramp	0.00%	100.00%	0.00%
Tight	902	514	No Work	Open	Ramp	97.78%	2.22%	0.00%
Tight	916	518	No Work	Open	Ramp	0.00%	100.00%	0.00%
Tight	906	514	No Work	Open	Ramp	43.77%	56.22%	0.00%
Tight	908	513	No Work	Open	Ramp	7.91%	92.08%	0.00%
Tight	924	509	No Work	Open	WZ	0.00%	100.00%	0.00%
Tight	909	516	No Work	Open	Ramp	11.35%	88.65%	0.00%
Tight	912	508	No Work	Open	Ramp	0.00%	99.13%	0.00%
Tight	919	502	No Work	Open	WZ	0.00%	90.40%	0.00%

Table 41: Response probabilities for a participant with a tight point distribution for the PCB alternative in the open condition (Experiment 1, 3s distance)

subject	x	y	Alternative	Condition	Selected Value	Pr(ramp)	Pr(WZ)	Pr(Closed)
Tight	886	510	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	883	511	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	883	507	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	884	504	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	884	502	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	882	502	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	879	505	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	882	508	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	885	509	PCB	Open	Ramp	100.00%	0.00%	0.00%
Tight	885	508	PCB	Open	Ramp	100.00%	0.00%	0.00%

REFERENCES

Accot, J., & Zhai, S. (2002). More than dotting the i's---foundations for crossing-based interfaces. *Proceedings of the SIGCHI Conference on Human …*, (4), 73. doi:10.1145/503387.503390

Andersen, E. B. (1972). The Numerical Solution of a Set of Conditional Estimation Equations. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(1), 42–54. Retrieved from http://www.jstor.org/stable/2985049

Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261–278. doi:10.1109/47.867942

Bryden, J., Fortuniewicz, J., & Andrew, L. (2000). Intrusion Accidents on Highway Construction Projects. *Transportation Research Record*. doi:10.3141/1715-05

Campbell, J. L. (Batelle), Lichty, M. G. (Batelle), Brown, J. L. (Batelle), Richard, C. M. (Batelle), Graving, J. S. (Batelle), Graham, J. (Midwest R. I., … Harwood, D. (Midwest R. I. (2012). *Human Factors Guidelines for Road Systems*. Washington, D.C. Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_600Second.pdf

Carroll, J. M. (1990). Infinite detail and emulation in an ontologically minimized HCI. In *CHI* (pp. 321–327). doi:10.1145/97243.97303

Chen, H., Zhou, H., Zhao, J., & Hsu, P. (2011). Safety performance evaluation of left-side off-ramps at freeway diverge areas. *Accident; Analysis and Prevention*, *43*, 605–612. doi:10.1016/j.aap.2010.08.019

Chen, N.-S., Yu, P.-T., Kinshuk, & Chang, T.-W. (2012). The effects of presentation method and information density on visual search ability and working memory load. *Computers & Education*. doi:10.1016/j.compedu.2011.09.022

Chin, H.-C., & Quek, S.-T. (1997). Measurement of traffic conflicts. *Safety Science*, *26*(3), 169–185. doi:10.1016/S0925-7535(97)00041-6

Chrysler, S. T., Wright, J., & Williams, A. (2004). *3D Visualization as a Tool to Evaluate Sign Comprehension*. Retrieved from http://d2dtl5nnlpfr0r.cloudfront.net/swutc.tamu.edu/publications/technicalreports/167721-1.pdf

Cole, B. L., & Hughes, P. K. (1990). Drivers don't search: they just notice. *In: D. Brogan (Ed.), Visual Search. Proceedings of the First International Conference on Visual Search (pp. 407-417). London: Taylor & Francis* .

Coren, S., & Girgus, J. S. (1980). Principles of perceptual organization and spatial distortion: the gestalt illusions. *Journal of Experimental Psychology. Human Perception and Performance*, *6*, 404–412. doi:10.1037/0096-1523.6.3.404

Crundall, D., Andrews, B., van Loon, E., Trawley, S., Collins, L., Chapman, P., & Underwood, G. (2012). Some hazards are more attractive than others: Drivers of varying experience respond differently to different types of hazard. *Accident Analysis & Prevention*. doi:10.1016/j.aap.2011.09.049

Donker, A., & Reitsma, P. (2007). Aiming and clicking in young children's use of the computer mouse. *Computers in Human Behavior*, *23*(6), 2863–2874. doi:10.1016/j.chb.2006.06.002

Dudek, C., Finley, M., & Ullman, G. (2001). Sequential Warning-Light System for Work-Zone Lane Closures. *Transportation Research Record*. doi:10.3141/1745-05

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, N.J.: Psychology Press. Retrieved from http://prx.library.gatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=44641&site=ehost-live

Florsheim, J. M., & Kulp, J. H. (1978). Traffic control element. United States.

Graham, J. L., Paulsen, R. J., & Glennon, J. C. (1978). Accident Analyses of Highway Construction Zones. *Transportation Research Record*, *693*, 25–32.

Heeger, D. (1998). Signal Detection Theory. Retrieved October 19, 2015, from http://www-psych.stanford.edu/~lera/psych115s/notes/signal/

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 1–19. doi:10.3389/fnins.2014.00150

Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of Perception in the Low-Prevalence Effect: Evidence From Active and Passive Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, (APRIL), 1–14. doi:10.1037/xhp0000053

Humphreys, G. W. (2015). Feature confirmation in object perception: Feature Integration Theory 26 years on from the Treisman Bartlett lecture. *The Quarterly Journal of Experimental Psychology*, (February), 1–49. doi:10.1080/17470218.2014.988736

Hunter, M., Rodgers, M., Corso, G., Xu, Y., & Greenwood, A. (2014). *Improved Methods for Delineating Diverges in Work Zones.* Atlanta, GA.

ISO. (2010). ISO 9241 Part 210: Human-centred design for interactive systems. In *ISO 9241 Ergonomics of humansystem interaction*. Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Jared, D., Dixon, K., & Daniel, J. (2000). Analysis of Fatal Crashes in Georgia Work Zones. *Transportation Research Record*. doi:10.3141/1715-03

Johnson, J. (2010). *Designing with the mind in mind: simple guide to understanding user interface design rules*. *Children*. Retrieved from http://books.google.com/books?hl=en&lr=&id=woVADVwtL1gC&oi=fnd&pg=PP2&dq=Designing+with+the+mind+in+mind&ots=vo7eeBASon&sig=V2ov3lai0ZQlRaR9zM93FbGu_Rw

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. doi:10.1177/01466210122032064

Khattak, A. J., Khattak, A. J., & Council, F. M. (2002). Effects of work zone presence on injury and non-injury crashes. *Accident; Analysis and Prevention*, *34*, 19–29. doi:10.1016/S0001-4575(00)00099-3

Khorashadi, A. (1998). *Effect of Ramp Type and Geometry on Accidents*. Washington, D.C.

Krupinski, E. a., Graham, A. R., & Weinstein, R. S. (2013). Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human Pathology*, *44*(3), 357–364. doi:10.1016/j.humpath.2012.05.024

Kubinger, K. D. (2008). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement*, *69*(2), 232–244. doi:10.1177/0013164408322021

Lambert, J., & Bard, C. (2005). Acquisition of visuomanual skills and improvement of information processing capacities in 6- to 10-year-old children performing a 2D pointing task. *Neuroscience Letters*, *377*(1), 1–6. doi:10.1016/j.neulet.2004.11.058

Lunenfeld, H., & Alexander, G. J. (1990). *A Users' Guide to Positive Guidance*. Washington, D.C.

M. Davis & Co. (2015). *National Telephone Survey of Reported and Unreported Motor Vehicle Crashes*. Washington, D.C. doi:DOT HS 812 183

MacKenzie, I. S., Kauppinen, T., & Silfverberg, M. (2001). Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 9–16). doi:10.1145/365024.365028

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. *Detection theory: A user's guide (2nd ed.).*

Mair, P. (2007). Extended Rasch Modeling : The eRm Package for the Application of IRT Models in R. *Journal Of Statistical Software*, *20*(9), 1–20.

Mair, Pa., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*(1), 26–43.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*, 529–544. doi:10.1007/BF02302590

McCartt, A. T., Northrup, V. S., & Retting, R. A. (2004). Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in Northern Virginia. *Journal of Safety Research*, *35*, 107–114. doi:10.1016/j.jsr.2003.09.019

Mourant, R. R., & Rockwell, T. H. (1972). Strategies of Visual Search by Novice and Experienced Drivers. *Human Factors*, *14*, 325–335. doi:10.1177/001872087201400405

Murata, A., & Bullinger, H.-J. (1991). An experimental evaluation of mouse, joystick, joycard, lightpen, trackball and touchscreen for pointing-basic study on human interface design. In H.-J. Bullinger (Ed.), *Human Aspects in Computing. Design and Use of Interactive Systems and Work with Terminals. Proceedings of the Fourth International Conference on Human-Computer Interaction*. Dept. of Commun. & Comput. Eng., Fukuoka Inst. of Technol., Japan: Elsevier IFIP IFIP. Retrieved from http://prx.library.gatech.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=inh&AN=4155851&site=eds-live&scope=site

NHTSA. (2013). Traffic Safety Facts 2011. *DOT HS 811 754*, 232.

Nielson, J. (1993). *Usability Engineering*. San Francisco: Morgan Kaufmann.

O'Shaughnessy, M. P., & Kayson, W. A. (1982). Effect of presentation time and Gestalt principles of proximity, similarity, and closure on perceptual accuracy. *Perceptual and Motor Skills*, *55*, 359–362. doi:10.2466/pms.1982.55.2.359

Pain, R. F., McGee, H. W., & Knapp, B. G. (1981). *NCHRP 236: Evaluation of Traffic Controls for Highway Work Zones*. Washington, D.C.

R Core Team. (2014). *R: A language and environment for statistical computing.* Vienna Austria: R Foundation for Statistical Computing.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*. doi:10.1109/TSMC.1983.6313160

Ré, J., & Carlson, P. (2012). *NCHRP Synthesis 431: Practices to Manage Traffic Sign Retroreflectivity*. Washington, D.C. Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_syn_431.pdf

Reason, J. (1990). *Human error*. *Human error.* Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1991-97559-000&site=eds-live

Roadway Safety Consortium. (2010). *Guidelines on Ensuring Positive Guidance in Work Zones*. Retrieved from http://www.workzonesafety.org/files/documents/training/courses_programs/rsa_program/RSP_Guidance_Documents_Download/RSP_Positive_Guidance_Download.pdf

Ross, H. E., Sicking, R., & Zimmer, R. A. (1993). *NCHRP Report 350: Recommended Procedures for the Safety Performance Evaluation of Highway Features*. Washington, D.C. Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_350-a.pdf

Rouphail, N. M., Yang, Z. S., & Fazio, J. (1988). Comparative Study of Short- and Long-Term Urban Freeway Work Zones. *Transportation Research Record*, (1163).

Schwark, J. D., MacDonald, J., Sandry, J., & Dolgov, I. (2013). Prevalence-based decisions undermine visual search. *Visual Cognition*, *21*(5), 541–568. doi:10.1080/13506285.2013.811135

Senders, J. W., & Morray, N. P. (1991). *Human Error*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sharp, H., Rogers, Y., & Preece, J. (2011). *Interaction Design: Beyond Human-Computer Interaction*. *Book* (Vol. 11).

Shinar, D. (2008). Looks are (almost) everything: where drivers look to get information. *Human Factors*, *50*, 380–384. doi:10.1518/001872008X250647

Siegel, P., Han, E., Cohen, D., & Anderson, J. (2013). A dissociation between detection and identification of phobic stimuli: unconscious perception? *Cognition & Emotion*, *27*(7), 1153–67. doi:10.1080/02699931.2013.774264

Soukoreff, R. W., & MacKenzie, I. S. (2004). Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International Journal of Human Computer Studies*, *61*(6), 751–789. doi:10.1016/j.ijhcs.2004.09.001

Stokes, R. W., Rys, M. J., & Russell, E. R. (1996). Motorist Understanding of Selected Warning Signs. *ITE Journal*, 36–41.

Theeuwes, J., & Godthelp, H. (1995). Self-explaining roads. *Safety Science*, *19*(2), 217–225. Retrieved from http://www.sciencedirect.com/science/article/pii/092575359400022U

Transportation Research Board. (2010). *Highway Safety Manual*. AASHTO.

Treisman, a. (1988). Features and objects: the fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *40*(2), 201–237. doi:10.1080/02724988843000104

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. doi:10.1016/0010-0285(80)90005-5

Trick, L. M., Brandigampola, S., & Enns, J. T. (2012). How fleeting emotions affect hazard perception and steering while driving: The impact of image arousal and valence. *Accident Analysis & Prevention*. doi:10.1016/j.aap.2011.07.006

Ullman, B. R., Trout, N. D., & Ullman, G. L. (2012). Driver Comprehension of Messages on Truck-Mounted Changeable Message Signs During Mobile Maintenance Operations. *Transportation Research Record: Journal of the Transportation Research Board*. doi:10.3141/2258-06

Ullman, G., Finley, M., & Bryden, J. (2008). Traffic safety evaluation of nighttime and daytime work zones. *Report*. Retrieved from http://trid.trb.org/view.aspx?id=875318

United Nations. (2011). United Nations Decade of Action for Road Safety 2011-2020. Retrieved January 12, 2013, from http://www.un.org/en/roadsafety/

Van den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. doi:10.1080/0044929031000

Wang, Z., Cao, B., Weiping, D., Lu, J., & Zhang, Z. (2011). Safety Evaluation of Truck-Related Crashes at Freeway Diverge Areas. In *90th Annual Meeting of the Transportation Research Board*. Washington, D.C.: National Academies of Science.

Warton, D. I., & Hui, F. K. C. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, *92*(1), 3–10. doi:10.1890/10-0340.1

Wolfe, J. M., & Van Wert, M. J. (2010). Varying Target Prevalence Reveals Two Dissociable Decision Criteria in Visual Search. *Current Biology*, *20*(2), 121–124. doi:10.1016/j.cub.2009.11.066

Zhai, S., Kong, J., & Ren, X. (2004). Speed-accuracy tradeoff in Fitts' law tasks - On the equivalency of actual and nominal pointing precision. *International Journal of Human Computer Studies*, *61*(6), 823–856. doi:10.1016/j.ijhcs.2004.09.007

Zhu, S., Ma, Y., Feng, J., & Sears, A. (2009). Speech-Based Navigation: Improving Grid-Based Solutions. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. Prates, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2009 SE - 6* (Vol. 5726, pp. 50–62). Springer Berlin Heidelberg. doi:10.1007/978-3-642-03655-2_6

VITA

AARON TODD GREENWOOD

AARON TODD GREENWOOD was born in Chattanooga, Tennessee, and raised in Mount Airy, North Carolina. He received his diploma from the McCallie School in Chattanooga, Tennessee in 2007, a B.S. in Civil Engineering from the Georgia Institute of Technology in Atlanta, Georgia in 2010, and a M.S. in Civil Engineering from the Georgia Institute of Technology in Atlanta, Georgia in 2012 before starting a doctorate in Civil Engineering at Georgia Tech. When he is not working on his research, Mr. Greenwood enjoys marathon training and flying small airplanes.