Assessing Driving Simulator Validity: A Comparison of Multi-Modal Smartphone Interactions across Simulated and Field Environments

Thomas McWilliams¹, Nathan Ward², Bruce Mehler¹, and Bryan Reimer¹

Abstract

The use of a driving simulator as a tool to evaluate secondary task performance elicits the question of simulator validity. After upgrading an existing driving simulator from a medium-fidelity to a high-fidelity configuration with a new software environment, a study was run to benchmark this simulator against previously published highway-driving data. A primary goal was to assess relative and absolute validity in a simulated highway environment. Data from 72 participants who performed manual and voice-based contact dialing tasks with a center-stack-mounted smartphone in either the driving simulator or driving on the highway in one of two vehicles is considered. This analysis compared secondary task demand between the simulator and on-road vehicles by primarily considering driver off-road glance behavior. Mean total eyes-off-road time, mean single-glance duration, and the number of long off-road glances showed similar patterns relative to the manual versus voice-based tasks in the simulator and the two on-road vehicles. A driving performance metric, percentage change of standard deviation of velocity, showed differing results between the simulator and on-road vehicles. It is concluded that these data make a strong argument for relative validity, and in some cases absolute validity, for this simulator for studying glance behavior associated with in-vehicle devices under a highway-driving scenario.

Driving simulators have been used as an alternative to on-road driving for experimental purposes since as early as the 1930s (1). Driving simulators offer better environmental control and raise fewer safety concerns than real-world driving. Due to technological leaps and the development of personal computers, driving simulators can be relatively easily implemented and are used for research around the world. Driving simulators can be generally divided into three groups: low fidelity, medium fidelity, and high fidelity (2). The fidelity of a driving simulator refers to how realistic an experience it can provide. Low-fidelity simulators are the most basic, often consisting of a computer monitor with a portable steering wheel and pedals placed at a desk. Medium-fidelity simulators typically employ a larger screen, or several screens, to encompass a larger portion of the participant’s field of view. These types of simulators often also have a partial vehicle cab with a driver’s seat, instrument cluster, and center stack with which the driver can interact. High-fidelity simulators generally use multiple, large screens to support a broader range of a driver’s field of view, often use full vehicle cabs, and can provide physical motion feedback through actuators mounted to the base of the simulator. There is evidence of the benefits of using high-fidelity simulators over lower fidelity, for example in driver training (3), but the level of fidelity needed often depends on the research question at hand. Although high-fidelity simulators may be more attractive in relation to their greater face-validity, they can require a sizable amount of space and resources to build and operate, whereas low-fidelity simulators can be relatively inexpensive and can be set up in small spaces. The present report considers data collected from a high-fidelity simulator and compares it with a subset of previously collected data (see publications 4, 5) from two on-road vehicles under actual highway-driving conditions.

At face value, the simulator in question is nearly identical to a commercially available vehicle; however, it is

1MIT Age Lab, Cambridge, MA
2Tufts University, Medford, MA

Corresponding Author:
Address correspondence to Thomas McWilliams: tmcwill@mit.edu
It is crucial to go beyond physical appearances to compare human performance in the simulator to that on the road. The physical validity of a driving simulator refers to the similarity of the simulator to an actual vehicle. This is often seen through a correspondence between the layout of components in the vehicle and the simulator, and similarities in the operation and dynamics of both environments (6, 7). Not surprisingly, high-fidelity simulators are often seen to have higher physical validity than low- and medium-fidelity simulators (6). The simulator in question arguably has physical validity in the sense that the driver sits in an actual Volkswagen New Beetle cab and interacts with a full sized steering wheel and the original vehicle accelerator and brake pedals. The simulated external environments presented to the drivers of presently available simulators are not able to provide the same level of physical validity. The world may be presented in a realistic manner, and interactions may be similar in the real world, but the environment is highly controlled by researchers and the environment does not present any substantive risk to the driver.

Putting physical validity aside, one must also evaluate whether or not a simulator produces behavioral results that are representative of those observed in the real world. Arguably, the more important type of validity is behavioral validity (7). There are two main aspects of behavioral validity: absolute and relative. The extent to which numerical values obtained in the simulator and in real-world driving are consistent is known as absolute validity. For example, if eye-glance behaviors (e.g., amount of time looking off-road during a specific task) are quantitatively the same in a simulator and on road, then one may consider the simulator to have absolute validity. Relative validity refers to the degree to which the direction and magnitude of behavior observed in the simulator and real-world driving correspond. For example, if two different tasks are performed in both a simulator and on road, and participants spend less time looking off-road during one task versus the other task in both environments, the simulator would have relative validity.

The present study was motivated by a major upgrade of an existing medium-fidelity simulator to a high-fidelity configuration that included the implementation of an entirely new simulation software package. The previous medium-fidelity configuration had been validated against on-road behavior in a number of ways. For example, physiological arousal in response to increasing levels of cognitive demand was evaluated in the simulator and on road using a structured working memory task (8). Though the absolute levels of heart rate and skin conductance were found to be higher under actual highway-driving conditions (thus not demonstrating absolute validity), the relative change in arousal with each level of demand was identical across the two environments (supporting relative validity). In a separate effort (6), a number of metrics (initial response time, mean task duration, and a range of glance measures) during interactions with three different configurations of a destination entry task, mapped in a very similar manner in both the simulator and during interaction with same interfaces on-road, showing good relative validity and, in some cases, absolute validity as well. Findings for driving performance metrics were somewhat mixed, as has been observed in other studies (7, 9). Other aspects of validity, considering the relationship between self-reported behavior and behavior observed driving the simulator have also been previously evaluated (10).

Of primary concern with the new simulator configuration, was the extent to which it could be considered a valid tool for continuing an ongoing program evaluating the demands on drivers associated with interacting with various in-vehicle devices and user interfaces as an analog for on-highway behavior. Of particular interest in this regard are glance behaviors that the National Highway Traffic Safety Administration (NHTSA) has utilized for visual demand assessment in driving simulators (11). Off-road glances are a frequently used metric to measure the amount of demand a device places on the driver (2, 6). There has been significant research on the connection between glance behavior and crash risk (12, 13), as well as the issuance of guidelines by NHTSA on glance behavior during secondary task performance while driving (14). In this context, the present paper focuses on the validation of the new driving simulator configuration for assessing the visual demand of manual and voice-based in-vehicle secondary tasks.

**Methods**

**Participants**

Data from a total of 72 participants was considered. A sample of 24 participants was used in the analysis of the simulator portion of the study, all between the ages of 20 and 24 or 55 and 69. Participants were equally balanced across the age and gender groupings (younger \( M = 21.7, \ SD = 1.1; \) older \( M = 59.1, \ SD = 3.4 \)). All reported having a valid driver’s license for more than 3 years, driving on average one or more days per week, and were in self-reported reasonably good health for their age. The on-road analysis sample consisted of two groups of 24 participants each drawn from the earlier on-road study (4, 5) that were balanced across gender and representative of age characteristics similar to the simulator cohort (younger \( M = 22.9, \ SD = 2.3; \) older \( M = 58.28, \ SD = 5.4 \)) although the age ranges were somewhat broader (20 to 28 and 46 to 69). All met the same qualifications as the simulator group, except all on-road participants were required to drive on average three or more
days per week. One half of the on-road participants drove a 2013 Chevrolet Equinox, the other half drove a 2013 Volvo XC60.

**Apparatus**

**Driving Simulator.** This experiment used a high-fidelity driving simulator designed and fabricated by RTI (Realtime Technologies Inc., Ann Arbor, MI), which ran the SimCreator and SimVista software packages. The simulator consisted of the full cab of a 2001 Volkswagen New Beetle. Three exterior screens provided a 180° simulated field of view. A rear projector screen hung behind the car displayed the rear roadway, which was viewable using the vehicle’s standard rearview mirror. Two LCD displays were placed in the cab’s side mirrors, which also presented the rear roadway. Images of the roadway were displayed at 60 frames per second. The steering wheel, gas, and brake pedals all provided force feedback to the participants. Internal and external speakers produced environmental noise during the experiment. The cab was instrumented with three cameras and one microphone to record the environment in the cab. One camera focused on the driver’s face to capture head and eye movements, one camera recorded hand movements and interactions with a center-stack-mounted smartphone. The last camera recorded the forward screen to validate video and simulator data synchronization.

**On-Road Vehicles.** The two on-road vehicles (a 2013 Chevrolet Equinox and a 2013 Volvo XC60) were each instrumented with a data acquisition system designed for time-synchronized recordings. This system recorded information from the vehicle’s controller area network bus, images from five video cameras monitoring the internal and external environments of the vehicles, and audio from a microphone inside the cab. Two cameras recorded the forward roadway (one narrow- and one wide-angle), one camera recorded the rear roadway, one focused on the center stack to monitor hand movements and interactions with a smartphone, and the last camera was focused on the driver’s face to capture head and eye movements.

**Smartphone.** In each on-road vehicle and the simulator a Samsung Galaxy S4 (SCH-1545) running Android 4.3 (Jelly Bean) was mounted to the right side of the center stack. The Chevrolet Equinox had a mount attached to the dashboard above the center stack that allowed the phone to hang down to the right of the screen. The Volvo XC60 had a mount that clipped to the air vent below the center stack screen placing the phone to the right of the screen. The simulator had a mount that clipped to an air vent above the center stack screen placing the phone to the right of this screen. Voice input tasks were performed using Samsung’s S-Voice interface with the hands-free mode turned on.

**Tasks**

Participants were instructed to place phone calls using a list of phone contacts stored in the smartphone, four using a visual-manual interface and four using a voice-based interface on the smartphone. Complete details are provided in Mehler et al. (4). The same four contact targets were used with each interface. In brief, to manually place a call, participants first had to wake up the phone by pressing the home button below the screen. Next they selected the “Contacts” icon along the bottom of the screen. This opened up a list of contacts that participants scrolled through until they came to the target name. For the first two calls, participants selected a contact’s name, which initiated the call as each of these contacts had a single number listed. For the second two calls, selecting the contact name brought participants to another screen where they selected a specified phone (e.g., home, work) that was the target for that task. The voice-based calling tasks involved pressing the home button twice and waiting for the phone to give a greeting that indicated it was ready to accept a voice command. Participants then said “Call” followed by the name of the person they were instructed to call. For the second two calls, participants had to also say the name of the contact and a target phone (e.g., “Call Frank Scott at work”). After placing calls (both manual and voice), the phone connected to a voicemail message instructing the participant to hang up the call.

**Procedure**

Participants read and signed an informed consent form and completed demographic questionnaires. Next, participants were brought to the vehicle or simulator where they completed the remainder of the experiment. After being introduced to the operation of the vehicle or simulator, participants were trained on the tasks that they would perform during the evaluation portion of the experiment.

The driving simulation scenario consisted of a divided highway with two travel lanes in each direction, a posted speed limit of 65 mph, and low to moderate traffic; this scenario was modeled after I-495 north of Boston, MA used in the on-road study. Participants were instructed to follow the posted speed limit and drive in whichever lane they were comfortable.

On-road participants drove with a research assistant in the back seat of the vehicle. Participants drove approximately 20 min north of Boston on I-93 to I-495. On I-
495, participants were instructed to drive in whichever lane and at a speed that they were comfortable with and that was appropriate for the traffic conditions. The posted speed limit was 65 mph.

For analytic purposes, there was a 2-min baseline-driving period before initiation of the blocks of each task type (manual and voice-base phone contact calling). Within each block, individual calling tasks were separated by 30 s of just driving. The presentation order of the manual and voice blocks was counterbalanced across participants. All tasks began with a prerecorded audio prompt instructing the participant who to call and when to begin the calling task.

**Data Reduction and Analysis**

Given the importance of driver glance behaviors for safety, the majority of this analysis focuses on eye-glance measures. Following procedures detailed in Mehler et al. (4), two trained coders independently manually coded glances in each dataset and a third coder mediated any discrepancies. Open source software, which allowed for frame-by-frame review and annotation of driver eye movements, was used for this coding (15).

Values for the individual trials within each block (manual or voice task) were averaged across the block. The statistical tests used a between-subject design for the three environments (simulator, Chevrolet, Volvo) and a within-subject design for the two modalities (manual and voice) for a 3×2 design. ANOVAs were performed in R (16).

**Results**

**Task Completion Time**

Overall, there was a significant main effect of task modality on overall task completion times when the sample is considered as a whole, \( F(1, 71) = 74.56, p < .001 \). Task times were nominally lower during manual tasks compared with voice tasks in all three environments (Figure 1), generally supporting relative validity for this variable. However, a significant interaction between the three environments and task modality was also found, \( F(2, 69) = 7.65, p < .001 \). Although the difference by modality was statistically significant when testing individually for the two on-road vehicles (Volvo: \( F[1, 23] = 69.44, p < .001 \); Chevrolet: \( F[1, 23] = 53.86, p < .001 \)), the difference in the simulator did not reach statistical significance, \( F(1, 23) = 3.83, p = .062 \); thus absolute validity is not supported. Looking at this another way, it can be observed that there was a significant effect of environment for manual tasks, \( F(2, 69) = 9.6, p < .001 \); participants took more time to complete the manual tasks in the simulator than participants took in the two on-road vehicles. However, there was no effect of environment on task completion time during voice tasks, \( F(2, 69) = .91, p = .41 \).

**Total Eyes-Off-Road Time**

The total amount of time that participants spent looking away from the forward road scene was coded and means across the trials making up each block were calculated for each participant. There was a significant main effect of task modality across the sample as a whole, \( F(1, 71) = 8.23, p < .01 \). On average, participants spent more time looking off-road during the manual tasks compared with voice tasks (Figure 2). There was not a significant interaction between environment and modality; however, it is evident that the modality effect was most prominent in the simulator, in which participants spent 4.8-s longer
looking off-road during manual versus voice-based calling. In the on-road vehicles total eyes-off-road time was 1.73-s longer in the Chevrolet and 1.65-s longer in the Volvo for the manual versus voice modality. When each of the samples was tested individually, only the difference in the simulator was statistically significant, $F(1, 23) = 7.27, p = .012$; the corresponding values for the Chevrolet are, $F(1, 23) = 0.94, p = .342$; and for the Volvo, $F(1, 23) = 1.58, p = .222)$. Thus, although the consistency in the patterning for nominally longer eyes off-road times for the manual interface across the environments is supportive of relative validity, absolute validity for the simulation was not demonstrated.

Mean Single-Glance Duration

Mean durations of single off-road glances showed a significant main effect of task modality, $F(1, 71) = 180.15, p < .001$, in which individual glance durations were, on average, longer during the manual tasks than when using the voice-based interface (Figure 3). No interaction between environment and modality was present and this pattern of longer mean single-glance durations during manual tasks ($M_{	ext{simulator}} = 1.04$ s, $M_{	ext{Chevrolet}} = .93$ s, $M_{	ext{Volvo}} = .98$ s) than voice tasks ($M_{	ext{simulator}} = .76$ s, $M_{	ext{Chevrolet}} = .63$ s, $M_{	ext{Volvo}} = .66$ s) was consistent and statistically significant for all three samples: $F(1, 23) = 77.20, p < .001$; $F(1, 23) = 61.29, p < .001$; and $F(1, 23) = 231.50, p < .001$, respectively. This patterning is consistent with relative validity.

Long Off-Road Glances

There was a significant main effect of task modality on the mean number of long-duration (longer than 2 s) off-road glances, $F(1, 71) = 32.24, p < .001$. On average, the number of long-duration glances was higher during manual tasks than during voice-based tasks (Figure 4). This finding was consistent and statistically significant across the three environments with participants making an average of 1.15 long off-road glances during manual tasks in the simulator, 0.6 in the Chevrolet, and 0.7 in the Volvo versus values for voice-based tasks of 0.37 in the simulator, 0.13 in the Chevrolet, and 0.08 in the Volvo. The corresponding post hoc tests within each environment were respectively: $F(1, 23) = 8.69, p = .007$; $F(1, 23) = 10.86, p < .001$; and $F(1, 23) = 19.38, p < .001$. No interaction effects were present. These results can be considered consistent with both relative and absolute validity.

Standard Deviation of Velocity

The standard deviation of velocity during each task was recorded, and the percentage change between baseline periods and the task periods was calculated. A significant effect of modality was found across the sample, $F(1, 67) = 4.07, p = .047$. When considered by modality, there was a significant difference in the percentage change in standard deviation of velocity from the baseline-driving period for manual tasks, $F(2, 65) = 3.4, p = .039$, but not for voice tasks, $F(2, 65) = .58, p = .55$ (Figure 5). Both the vehicles and the simulator had smaller changes in the standard deviation of velocity from the baseline periods during voice tasks compared with manual tasks; however, the simulator showed overall increases in standard deviation of velocity from the baseline and the two vehicles showed decreases in the standard deviation of velocity from the baseline. Thus, neither absolute nor relative validity is apparent for this variable.
Figure 5. Mean percentage change in the standard deviation of velocity from baseline driving by environment and task modality. Error bars represent the standard error of the mean.

Discussion

The same basic behavior patterns appeared across the two on-road vehicles with regard to the relationship between task modality (manual versus voice-based interface) and each of the variables assessed. Except for the standard deviation of velocity variable, the simulator data largely conformed to these patterns in terms of relative validity, though some differences in absolute validity were present.

Task completion times in all environments were shorter for manual contact calling tasks compared with voice-based contact calling tasks with a mounted Samsung Galaxy S4. This is consistent with previous research (4, 5) and is traceable to the duration of verbalizations by the driver, auditory prompts, and responses on the part of the voice system. The absolute difference in task completion time in the simulator was smaller between manual tasks and voice tasks than was observed in the two vehicles; this corresponded to simulator participants taking more time to complete manual calls relative to what was observed on-road, whereas task duration for voice-based calling was not significantly different across the three samples. Possible reasons for this finding are considered below.

A significant effect of task modality was found for total off-road glance time considering participants across the sample. The nominal mean values for each manual sample were greater than the mean value for each sample’s corresponding voice-based calling task, thus in line with relative validity. The finding that a visual-manual task is more visually demanding than their corresponding voice-based alternative was expected and is consistent with numerous other studies (17–19). However, this difference was largely driven by the simulator, which was the only environment for which this difference was statistically significant. This will also be considered below.

Mean single-glance duration is often used to assess the demand required to interact with a system. All environments had a significant difference in mean single-glance time, with the primary visual-manual tasks eliciting longer glances. These results are consistent with previous literature (4, 5). The consistency in patterning and lack of interaction between task modality and environment supports a finding of both relative and absolute validity in this simulator for this task, metric, and scenario.

Another glance metric that is often used to assess the demand brought on by using a device while driving is the number of off-road glances that last longer than 2 s. This measure is often used as a way to assess the risk involved with using a device as long glances are associated with increased accident risk (12, 13). The significantly greater number of long off-road glances during manual tasks is consistent with previous findings (20–22), and is consistent with the idea that manual tasks place more visual demand on drivers. As the patterning of differences between task modalities is consistent across environments and there is no task modality by environment interaction, these findings are generally supportive of both relative and absolute validity.

Thus, the results for all three visual metrics are consistent with a finding of relative validity in the simulator for these variables. This is in line with findings for the medium-fidelity simulator that this high-fidelity simulator replaced, where the relative patterning of visual metrics characterizing how drivers looked off-road to interact with different visual-manual human machine interface (HMI) implementations aligned well with what was observed running the same tasks with the same interfaces under actual highway-driving conditions (6).

The fact that there were no significant differences by environment for these visual metrics when the modalities are considered independently could be taken as evidence for a degree of absolute validity as well. However, it is also worth noting that total off-road glance time during manual tasks, the mean single-glance duration for both task modalities, and the number of long-duration glances for both task modalities were all nominally greater in the simulator. Although not sufficient evidence to rule out absolute validity for these metrics, it is possible to hypothesize a trend toward participants being willing to visually engage somewhat more in off-road glance behavior during simulation than during the functionally more risky real-world driving environment. It can also be noted that participants engaged longer with the manual interface in the simulator than was observed on-road. This might also reflect on absolute validity in that participants may perhaps be comfortable taking their hands off the steering wheel somewhat longer in the simulator.
than they are on road. Nonetheless, these speculations do not affect the strong evidence for relative validity for the visual metrics.

Changes in a vehicle’s velocity can suggest that the driver is experiencing added demand due the performance of a secondary task. Calculating the percentage change in the standard deviation of velocity from a designated baseline-driving period allows for a comparable measure that adjusts for driver-related individual differences and variation in environmental conditions. Again, there was an effect of task modality on the percentage change in standard deviation of velocity for each driving environment. Higher demand on drivers during manual tasks compared with voice tasks is the likely cause of this (17). The most striking aspect of these results is that both secondary tasks in the simulator resulted in an increase in the standard deviation of velocity, whereas engaging in the same tasks in the two on-road vehicles resulted in a decrease. It can be noted that there was only a statistical difference between the simulator and other vehicles during the manual tasks.

One major difference between simulator and highway driving is the consistency of the surrounding traffic. The simulator was programmed to have a consistent flow of traffic that would allow the driver to maintain a consistent speed. Not surprisingly, this is not the case on most highways. Though the highway-driving data was intentionally collected at times when there was a high likelihood of low to moderate traffic, this does not mean that traffic flow variation could be completely avoided, especially when things like accidents and construction can occur and affect traffic patterns. In addition to the variability of highway-driving conditions, participants in the simulator were instructed to drive at the posted speed limit of 65 mph. On the highway the posted speed limit was 65 mph, but participants were instructed to drive at a speed they were comfortable with that was also appropriate for the road and traffic conditions at the time. The divergence in the standard deviation of velocity between the simulator and highway-driving conditions suggests that the simulator did not achieve absolute validity in this case, and that it may continue to be a challenge to achieve in the future. However, the trend of standard deviations in velocity being closer to baseline driving during voice tasks and having an increased difference during manual tasks suggests that this measure may still demonstrate the relative validity of the simulator.

Summary
Previous work has argued for the importance of validating simulation systems for the use cases in which they are employed (10). The present study and analysis was run to assess the extent to which a simulator that was validated for studying driver glance behavior during interactions with HMI (6) continued to be valid for this application following a major upgrade from a medium-fidelity to a high-fidelity configuration that also included a change to an entirely new software simulation package. Consistent with what was found in the previous validation effort, the new simulator demonstrated relative validity for studying HMI-related glance behavior. Specifically, the finding in the full on-road dataset (5) that the voice-based smartphone interface for phone contact calling was associated with shorter total eyes-off-road time, shorter mean single-glance durations, and fewer long-duration glances than the manual interface was also observed in the new simulator. Therefore, it seems appropriate to use this simulator as a tool to assess these aspects of secondary task demand on drivers in place of real-world driving. At the same time, other results reinforce the importance of recognizing that not all behaviors and metrics will show absolute validity, or even necessarily relative validity, between a given simulation and the real world. This is evident in participants’ apparent willingness to spend more time engaged in the manual phone interface interaction in the simulator than was observed on-road, and the finding that a driving performance metric (standard deviation of velocity) changed in opposite directions under secondary task demand in the simulator versus actual on-road conditions.

Acknowledgments
This work was supported by the New England University Transportation Center and the Toyota Class Action Settlement Safety Research and Education Program.

Author Contributions
The authors confirm contribution to the paper as follows: study conception and design: TM, BM, BR; data collection: TM; analysis and interpretation of results: TM, NW, BM, BR; draft manuscript preparation: TM, NW, BM, BR. All authors reviewed the results and approved the final version of the manuscript.

References


The Standing Committee on Simulation and Measurement of Vehicle and Operator Performance (AND30) peer-reviewed this paper (18-06135).

The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs’ class counsel.