Serendipity in Simulation: Building Environmentally Valid Driving Distraction Evaluations of Google Glass™ and an Android™ Smartphone

B. D. Sawyer[1, 2], A. A. Calvo[3], V.S. Finomore[4], P. A. Hancock[1]

[1]Psychology Department, University of Central Florida, Orlando, USA
[2]Oak Ridge Institute for Science and Education, Wright-Patterson Air Force Base, USA
[3] Ball Aerospace, Wright Patterson Air Force Base, USA
[4] Air Force Research Laboratory, Wright Patterson Air Force Base, USA

In evaluating Google Glass, our team created a novel dependent variable collection strategy in which time-synchronized devices reported time-stamps for each user action, permitting precise temporal targeting of events in a simulated driving task. This device status reporting (DSR) trigger system was used to expose vehicle operators in the midst of reading an incoming text message to a consistent, controlled worst-case scenario roadway event: a braking lead vehicle. We argue that such strict control over the temporal delivery of the treatment effect provides results which are more environmentally valid and interpretable than strategies which rely on high base rates. After establishing the theoretical need we provide a technical description of a working DSR system, discuss associated technical and methodological concerns, and look to future utility and implications.

Keywords: wearable device, environmental validity, traffic safety, methodology

1. Introduction

It begins with a fateful message alert. The driver retrieves and readies the device in one practiced motion, never looking away from the roadway. After a final forward scan, their eyes glance down and moments pass as the message is read. The story here turns to surprise and consequence. A car ahead slams on its brakes, or a dog runs into the road, or a man does, or a child. Such stories moved the scientific phrase 'driver distraction' into law, and then into day-to-day English, where it became Websters' 'Word of the Year (Agnes, 2004). These real-world cautionary tales are powerful precisely because they describe the most unfortunate of serendipities: a worst case scenario in which a driver's attention shifts 500.

In the psychological study of driving, the above scenario can be framed in the large amount of underload drivers endure relative to rare overload situations (as in Hancock & Warm, 1989). Such contrast in workload has been referred to as "hours of boredom and moments of terror" (Hancock, 1997). Drivers often compensate by engaging in satisficing (Simon, 1969), performing not to the best of their ability, but to the minimum requirements of the present situation. This strategy further frees driver resources for peripheral, which break the monotony at a cost to driving performance. Distracted, drivers reveal their impairment by moving more slowly (Törnros & Bolling, 2006), more erratically (He, McCarley & Kramer, 2013), and by compensating for their detriment by maintaining greater following distances to vehicles ahead (Sawyer, Finomore, Calvo & Hancock, 2014; He et al., 2014). Such changes in driving behavior do not immediately subside when any secondary task ends. This hysteresis (as in Morgan & Hancock, 2011) suggests that the 'only a moment' of taking a message while driving may, in terms of impact to handling future roadway events, have a long tail. When load increases suddenly, as it does in roadway incursions, the reduction in available resources limits the ability of the driver to react appropriately.

For distracted drivers, 'moments of terror' (Hancock, 1997) are more likely to result in dynamic instability and subsequent failure (Hancock & Warm, 1989). Some divide exists in the research community between those that understand this risk chiefly as a consequence diversion of visual and manual resources, and those that propose it is largely due to cognitive factors related to language, strategy, and working memory demands. Supporting either the structural or cognitive interference theoretical perspective is not the purpose of the present work. Rather, we submit that performing meaningful tests of either theory requires an increase in the amount of experimental control in driving research at large. In the present work we describe the theoretical need, describe a system addressing that need, discuss relevant associated technical and methodological concerns, and look to future utility.

2. **The Need**

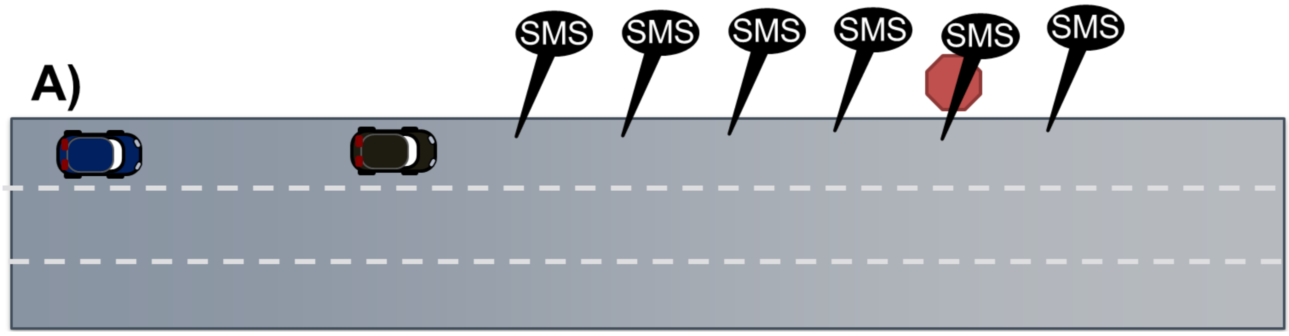

Figure 1: A ‘pace car paradigm’ experiment, in which the participant’s car (blue, left) follows a computer controlled lead car (black, right) while the participant engages in an additional task or tasks (here, SMS text messages). When the lead car brakes suddenly, response time data is collected. The challenge of aligning the treatment effect (messaging) with the event producing the DV (brake) has led to a common strategy of flooding the environment. Above, repeated text messaging leads to such an overlap, but it is neither controlled nor guaranteed. Further, such tactics can lead to suboptimal experimental procedure that deviates considerably from real-world situations. Still, what better solution exists when participant free will renders consistent temporal placement of the message relative to the brake event impossible?

When measuring discrete roadway events (as in Sawyer & Hancock, 2013; Sawyer et al., 2014; Drews, Yazdani, Godfrey, Cooper, & Strayer, 2009; and many of the works described in Caird, Johnston, Willness, Asbridge, & Steel, 2014), participant free will creates difficulties with precise temporal placement of a treatment relative to a measured behavior. How can a researcher be certain engagement with a secondary task will coincide with the measurement of a dependant variable when the participant can do whatever they wish? To compensate, researchers have, in general, tended to rely on a high base rate of the manipulation over the duration of the experiment. If a participant never stops performing the secondary task, a complex issue of targeting the manipulation is masked by the presumed homogeneity of that task space. This approach to managing participant free will, in the present work, will be referred to as *flooding*.

The *pace car* paradigm (see Figure 1), in which a lead vehicle brakes at unexpected intervals, is often paired with flooding strategies. In continuous conversational vocal cell phone use (as in Strayer & Drews, 2004), this strategy is arguably quite a naturalistic one, and a solution to the identified issue of temporal placement. In a discrete task, (as in Sawyer & Hancock, 2013; Sawyer et al., 2014; for a review of such studies see Caird et al., 2014) flooding is, in terms of environmental validity, less optimal. For example, Drews and colleagues (2009), presented 42 brake events at “freeway speeds”. Sawyer and Hancock (2013) sent text messages continuously, a new one delivered as soon as the previous message had been responded to, while a single brake event was measured. However, it is difficult to argue that either situation proves to be a common one on the road.

Beyond environmental validity concerns, flooding is an inherently imprecise experimental approach. It relies upon hysteresis effects (Morgan & Hancock, 2011) to ‘smooth out’ temporal mismatches caused by differences in participant behavior, offering very little in the way of precise experimental control. Such strategies lack the fidelity to examine tasks as constellations of sub-tasks. Messaging, for example, involves a series of sub-tasks corresponding to interface elements. These differ by device and are responsible for the outcome differences seen between, say, drivers communicating using Google Glass as compared to an Android smartphone (as in Sawyer et al., 2014). Most studies which rely on flooding produce broad binary answers: drivers who engage in given secondary tasks either 1) exhibit driving detriment or 2) do not, when compared to baseline driving. Binary data evaluating a device as a whole is used because, without the ability to temporally target manipulations in a device task space, meaningful evaluations of individual user interface decision affordances become impractically complex. The best studies employing flooding tactics code matches between messaging activity and response maneuvers, and *then* analyse (as in Drews et al., 2009). Post hoc

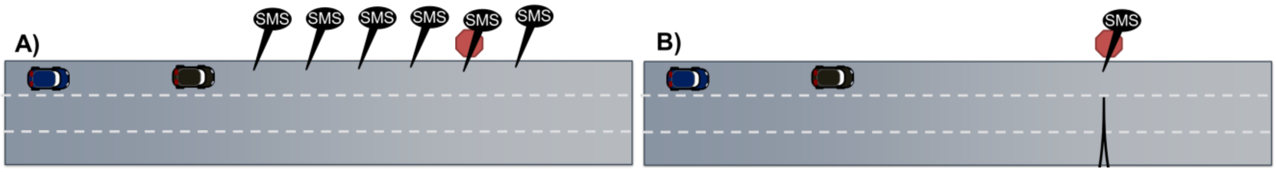

Figure 2: As opposed to past studies (A, and see Figure 1), the Device Status Reporting (DSR) system (B) allowed the precise targeting of a brake event contingent upon user interaction with a device. Such a system can, for example, interrupt a specific subtask within operation of a device. To the participant, the event is unexpected. To the researcher, the event is precisely placed for every participant in the study.

results of these sort still arguably lack strong experimental control advantages over naturalistic observational studies (such as Fitch et al., 2013) although it should be noted that simulator studies undeniably retain *cost* advantages.

A better approach, a more detailed lens, was needed. Our team's response to these needs was to 'build' reliable, repeatable serendipity to replace the flooding strategy of creating serendipity through high base rates. Our resultant device status reporting (DSR) trigger tool can elicit simulator events, such as braking, in response to participant interaction with in-vehicle technology. The advantage of such a trigger tool is greater experimental control, and the ability to build more naturalistic tasks.

3. The System

The initial goal of the system was to allow, in simulation, the consistent production of a driving distraction worst-case scenario: a dangerous roadway event in the course of engaging with a message (Sawyer et al., 2014). Building such serendipitous situations in the laboratory has historically been a major undertaking (for an example, see Hancock & DeRidder, 2003). It had already been decided that participants would receive short messages including difficult mathematical problems (for rationale of this decision see Sawyer et al., 2014). It was desirable that participants have the free will to answer text messages at their leisure. As a result a series of pace car style scenarios was built in which the participant could receive a text message and drive at length behind a lead car. It was decided to use the participant's interaction with the in-vehicle device to initiate the pace car's braking behavior. The act of reading the message was targeted for the event because, of the various user interface elements, the two devices of interest were determined to differ most significantly in terms of display (see Sawyer et al., 2014).

The in-vehicle devices of interest were a Google Glass (hereafter 'Glass') unit running software version XE11 and a Galaxy Nexus smartphone running Android 4.3, and in both devices there was a desire to use the stock messaging application. The simulator in question was a fixed platform PatrolSim, with three screens displaying the virtual environment over 270 degrees of visual angle, as well as a dash, automatic transmission, wheel and seat from a Crown Victoria police cruiser. The Patrolsim's five computer network, which included three visual channels, an interpreter for the dash, and an operator's console, ran custom-built software for data collection and integration (see Sawyer & Hancock, 2012 for a more detailed description).

The feature of the system on which the present work will focus is device status reporting (DSR), the ability for in-vehicle devices to report the timing and nature of each action taken by the human. Our DSR system was built to interact with a central dispatch application, programmed in C++, capable of running on the operator's console and reading a data multicast providing millisecond-precision updates of the x,y position of the participant's vehicle. This central dispatch was also capable of injecting commands to the simulator over the network. Trigger logic was loaded in comma separated value (CSV) script files containing a) x,y position, b) radius for detection, and c) actions to be sent to the phone and/or simulator. Each such set of values constituted a single trigger in a system capable of temporally targeting events in the simulation (See Figure 1).

In Sawyer & colleagues' 2014 study, three goals were identified for the trigger system: 1) to interrupt participants with the brake event as they read an incoming message (a question) and considered the answer and 2) to measure recovery from the time the participant locked the phone. In pilots, participant self-report and observation gave us an average time of 3600ms from phone unlock to first key press in a reply, and 3400ms from Glass unlock to beginning the dictation of a reply. Therefore, we were interested in determining when a participant unlocked the in-vehicle device and 1800 ms later triggering a brake event in the pace car (Figure 3B). In order to properly segment time for our DVs, we also needed to record when they began typing a reply, and when they sent a reply.

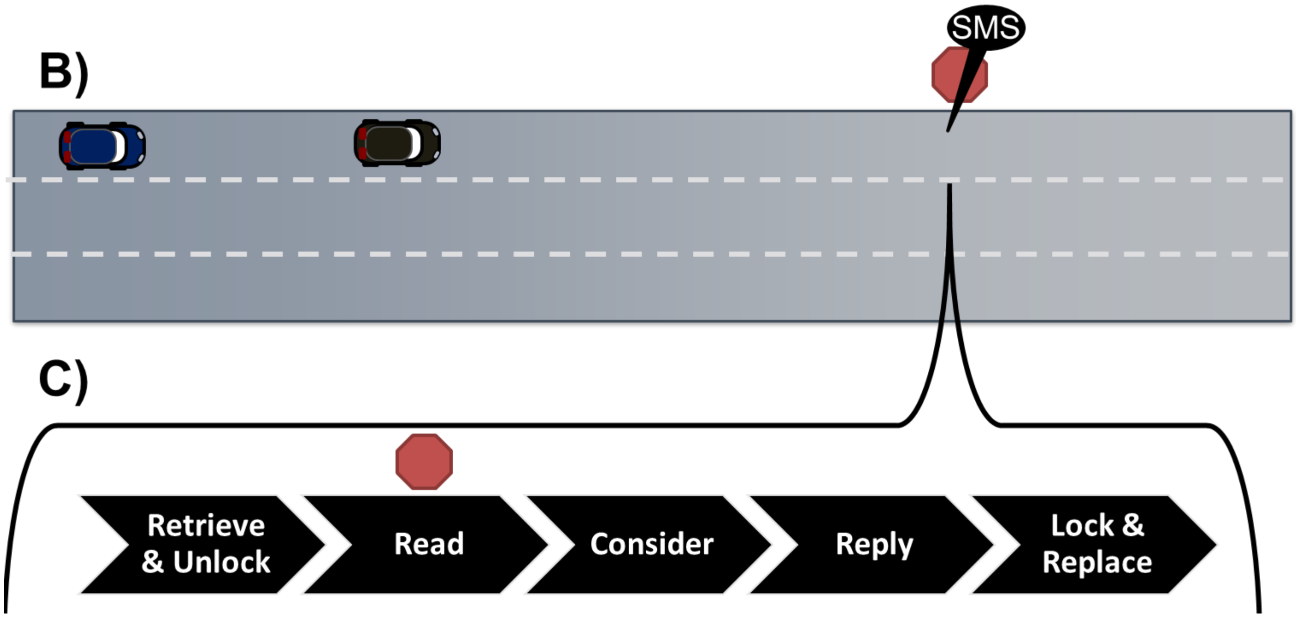

Figure 3: In the present project (B) a user unlocked the phone and 1800ms later the lead car braked sharply. The intent of this timing was to inject the brake event into the time in which the participant was (C) reading and cognitively processing the answer. The system could likewise be used to trigger actions in the simulation relative to any user interaction event. In the described project we only used interaction with onscreen interface items for such interactions, but a native app might interact with sensors as well.

To identify and temporally target these events, we reverse engineered the stock SMS apps for both Glass and Android. Both output log messages for debugging purposes as users responded to text messages, and although these debugging files were invisible to a user and typically inaccessible from an external app, we were able to read the messages from a custom app on rooted devices. Our app ran in the background, determined when a participant took an action, and wirelessly forwarded this information to a central dispatch application on the operator's console.

The resulting system (see Figure 3 B and C) was able to both record on-device and broadcast wirelessly the timing of the manipulation of all interface elements. Although our technical solution was specific to our simulator, many simulators and test-track environments allow for similar network report of vehicle position and elicitation of roadway events. It seems likely the DSR triggering scheme our team used might generalize well to many other experimental setups. Our system relied on the native SMS apps because Glass did not have a native development kit at the time of the study. Since then, Google has released the Glass Development Kit (GDK). It makes possible a much more straightforward approach: a custom app that mimics the interface of the native SMS app while logging all relevant information. On Android, this approach would allow an app to run on any modern device (unlike our app, which only runs on Nexus devices). On Glass, this approach would work on an un-rooted device, and it would be able to log additional data our approach could not.

4. Discussion and Future Directions

Of the technical challenges overcome in implementing our DSR trigger system, the issue of time is the one we feel most likely to be encountered in similar systems. Human response time in braking can be under a second, and differences between groups can run to the hundreds of milliseconds (Drews et al, 2009, Sawyer et al., 2014). Sub-second time inconsistencies, unlikely to be noticed in the course of experimentation, can nonetheless confound such sensitive temporal measures and subsequent analysis. In our experiment, random (relative to experimental condition and participant task) temporal inconsistencies were of a magnitude that would have obscured the differences we strove to detect. What is worse, we also encountered (and mitigated) temporal inconsistencies that systematically varied by device.

One cause of these inconsistencies was latency, the delay in communication between systems or parts of a system. Latency between hardwired PCs in our PatrolSim network was generally below 1ms and seldom higher than 10ms, but latency to either wireless Android device initially averaged 500ms, with a standard deviation, in one early test, of ~350ms. We mitigated this issue by forcing disabling power-saving features in software. This forced the Android device's processor and WiFi card to remain on and at full operational capacity at all times, even when the phone's screen was off.

Another source of temporal inconsistencies was clock drift. Hardware clocks tend to drift apart over time; even devices of the exact same model will drift at different rates (Murdoch, 2006). Google Glass and the operator's console, in our example, could easily drift 100ms apart over an hour. We used Network Time Protocol (NTP) synchronization at the beginning of each trial to mitigate the effects of clock drift. Technical details are here less important than a general principle: commercial operating systems are not real-time, and latency in different parts of a system can vary by orders of magnitude. Claims of 'real time' recording in a system should therefore be treated with extreme scepticism, and 'millisecond accuracy' should prompt the question of "how many milliseconds". A proper goal, in building systems for psychological experimentation, is a level of temporal error and drift which is as far as possible below the threshold of the differences expected.

Our DSR system reliably produced the desired worst-case scenario in our simulation, providing experimental control without interfering with the free will of our participants. DSR-based trigger systems, like that described here, have considerable application beyond the present project. We are, for example, presently using a similar system to explore prevalence effects in driving distraction, and to look at interface-component-level contributions to secondary task load. DSR systems have the potential to move beyond binary device distraction evaluations and explore the impact of all subtasks within a device, potentially identifying user interface design responsible for undue detriment or surprising lack of detriment to the driving task (as in Sawyer et al., 2014). DSR could assist the better understand of hysteresis effects (as in Morgan & Hancock, 2011). The fine experimental targeting of treatment effects which DSR permits allow for many of the broad findings of driving distraction research to be re-examined with greater fidelity.

Complex in-vehicle devices cause detriment to the driving task, and technological advance means there is no end in sight for the production of new candidates. The Apple Watch is out, and the view of this team is that it could fuel a paper exposing the fact that driving while distracted by a screen on your wrist is as bad an idea as driving while distracted by one in your lap (as in phones), or a virtual screen (as in Glass). At issue here is that merely identifying dangerous objects pays little indemnity, and the loss of lives to these thieves of attention (Hancock & Sawyer, 2015) continues regardless. Indeed, in a growing number of contexts the decision to engage in secondary tasks during a complex primary task like driving is both rational and necessary. Prohibition, appropriate for civilian drivers, is not a reasonable reaction to the multitasking challenges of officers of the law, emergency vehicle operators, or military personnel. To address the challenges such professions face, driving distraction research must move beyond litmus tests of in-vehicle interface. The present challenge is the delivery of information to operators involved in complex, potentially dangerous primary tasks with minimal decrement to performance and maximum comprehension of the message. Such future design will necessarily rely on in-depth, interface element by interface element analysis of present systems. We intend DSR to provide a lens of greater fidelity for such vital efforts.

References

Agnes, M. (2004). *Webster's new world college dictionary*. Indianapolis: Wiley Pub

Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention*, *71*, 311–318.

Drews, F. A., Yazdani, H., Godfrey, C. N., Cooper, J. M., & Strayer, D. L. (2009). Text messaging during simulated driving. *Human Factors, 51*(5), 762-770.

Fitch, G. A., Soccolich, S. A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., Perez, M. A., Hanowski, R. J., Hankey, J. M., & Dingus, T. A. (2013). *The impact of hand-held and hands-free cell smartphone use on driving performance and safety-critical event risk.* (Report No. DOT HS 811 757). Washington, DC: National Highway Traffic Safety Administration.

Hancock, P. A. (1997). Hours of boredom, moments of terror, or months of monotony, milliseconds of mayhem. *Proceedings of the Ninth International Symposium on Aviation Psychology, April.* Columbus, OH.

Hancock, P. A., & De Ridder, S. N. (2003). Behavioural accident avoidance science: understanding response in collision incipient conditions. *Ergonomics*, *46*(12), 1111-1135.

Hancock, P. A., & Sawyer, B. D. (2015). Judging thieves of attention. *Human Factors*. In press.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors, 31*(5), 519-537.

He, J., McCarley, J. S., & Kramer, A. F. (2013). Lane keeping under cognitive load performance changes and mechanisms. *Human Factors, 56*(2), 414–426.

He, J., Chaparro, A., Nguyen, B., Burge, R. J., Crandall, J., Chaparro, B., Ni, R., Cao, S. (2014). Texting while driving: Is speech-based text entry less risky than handheld text entry? *Accident Analysis & Prevention, 72*, 287–295.

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload an example of hysteresis in driving. *Human Factors*, *53*(1), 75-86.

Murdoch, S. J. (2006). Hot or not: Revealing hidden services by their clock skew. *Proceedings of the 13th Conference on Computer and Communications Security.* Alexandria, Virginia.

Sawyer, B. D., & Hancock, P. A. (2012). Development of a linked simulation network to evaluate intelligent transportation system vehicle to vehicle solutions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeti*ng, *56*(1), 2316-2320.

Sawyer, B. D., & Hancock, P. A. (2013). Performance degradation due to automation in texting while driving. *Proceedings of 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, 68*, 446-452. Bolten, NY.

Sawyer, B. D., Finomore, V. S., Calvo, A. A., & Hancock, P. A. (2014). Google glass: a driver distraction cause or cure? *Human Factors, 56*(7), 1307-1321.

Simon, H. A. (1969). *The sciences of the artificial*. Boston, MA: MIT Press.

Strayer, D. L., & Drew, F. A. (2004). Profiles in driver distraction: Effects of cell phone conversations on younger and older drivers. *Human Factors, 46*(4), 640–649.

Törnros, J., & Bolling, A. (2006). Mobile phone use – effects of conversation on mental workload and driving speed in rural and urban environments. *Transportation Research Part F: Traffic Psychology and Behaviour, 9*(4), 298-306.