TEST-RETEST RELIABILITY OF SIMULATED DRIVING PERFORMANCE: A PILOT STUDY

Christopher Irwin, David Shum, Michael Leveritt & Ben Desbrow
Griffith Health Institute, Griffith University
Gold Coast, Queensland, Australia
Email: c.irwin@griffith.edu.au

Summary: Twenty-seven volunteers completed three simulated driving tests to determine test-retest reliability of performance on a low-cost, fixed-base computerized driving simulator. One retest was completed a few hours after the initial drive, and the final retest was completed 7 days following the initial test drive. Driving performance was compared using measures of vehicle control, speed, and reaction time to critical events. A measure of participants' ability to inhibit a pre-potent response was also assessed using an inhibition task during each drive, with the number of incorrect inhibition responses recorded. Practice effects were evident for measures of vehicle control (deviation of lane position and number of line crossings) and participants' ability to withhold responses to inhibition tasks. Good test-retest reliability was observed for measures of vehicle control, speed, reaction time, and variability measures. Poor test-retest reliability was observed for the number of stopping failures observed during driving. The findings from this study suggest that the driving scenario used provides reliable assessment tasks that could be used to track the effects of pharmacological treatments on driving performance. However, an additional familiarization drive should be included as part of future study protocols employing this driving scenario to reduce learning effects during trials. Care should also be taken when interpreting results from tasks with low test-retest reliability.

INTRODUCTION

Driving simulators offer a safe and cost effective method of collecting objective and repeatable measures of driving performance (Allen, Rosenthal, & Cook, 2011). They also provide a means to investigate situations that would otherwise be dangerous (e.g. alcohol impaired driving, sleep deprived driving) (Caird & Horrey, 2011). Test-retest reliability of assessment instruments is important in behavioral based research and is a well-established principle in most areas of psychology. A substantial body of literature exists on the test-retest reliability of standardized neuropsychological assessments (e.g. Wisconsin Card Sorting Test). However, surprisingly few studies have examined the test-retest reliability of driving simulator measures (Akinwuntan, Tank, Vaughn, Wilburn, & Easton, 2009; Bedard, Parkkari, Weaver, Riendeau, & Dahlquist, 2010; Marcotte et al., 2003; Törnros, 1998). In those that have, the repeated administration of driving tests appears to follow after a significant time lapse (2-3 months). Laboratory based studies of driving behavior and performance often involve multiple assessments of individuals. For example, when investigating pharmacological effects (e.g. alcohol) on driving performance, researchers often employ protocols that involve testing before and after exposure to a treatment. In these cases, the duration between initial testing and retesting is likely to be a matter of minutes or hours rather than months. Assessment tasks that have low test-retest reliability have implications for their use in applied or clinical settings. They are limited in the tests sensitivity to detect changes in performance when administered repeatedly (Lowe & Rabbitt, 1998). A limitation of repeated testing is that improvement with practice may occur (Beglinger et al., 2005). This is normally most pronounced when intervals between testing are short and could potentially mask other effects that may be present, leading to confounding results (Collie, Maruff, Darby, & McStephen, 2003). The purpose of this pilot study was to examine the test-retest reliability of driving simulator performance measures over relatively short re-test intervals (hours and days). Test-retest reliability data from this study may provide greater confidence in the interpretation of driving performance changes observed in future studies where retesting is completed after short delay intervals and treatment effects are anticipated.

METHODS

Twenty-seven volunteers (13 male, 14 female) aged between 19 and 34 years (mean 24±4.4 years) participated in this study. Participants had no known neurological conditions or injuries that would influence their driving ability. All participants held a valid driver's license, had at least 2 years driving experience (range 2-18 years), and drove at least 5000 km each year. For each testing session, participants were asked to refrain from alcohol, non-prescription medications, and recreational drugs in the 24 hours prior to each test. In addition, they were asked to avoid consuming any caffeinated food and beverages and to drink at least 1 liter of water in the 2 hours prior to testing to assist with maintaining adequate hydration status. Dehydration has been associated with impairment in cognitive functions and mood, which may influence driving performance (Grandjean & Grandjean, 2007; Lieberman, 2010). Prior to completing test drives, all participants completed a 10 min familiarization drive on the simulator to become accustomed to the controls and driving in the virtual environment. Two of the test drives were conducted on the same day, one completed between the hours of 0800 and 1100 (Test 1), and one completed between the hours of 1300 and 1600 (Test 2). The third test drive was conducted 7 days after the initial test drive between the hours of 1000 and 1400 (Test 3).

The driving simulation task was operated on a desktop computer with peripheral devices for steering wheel, gas and brake pedals, and gear shifter (Figure 1). Visual images were displayed on three 22-inch LCD monitors (3840 x 1024 resolution), set to provide a 100° front field of view. A rear scene was also displayed on the central monitor to provide images associated with the rear view mirror. Images from the simulation software were refreshed at a rate of 60 Hz, with data sampled at a rate of 20 Hz. Auditory and haptic feedback were provided using a stereo sound system and force feedback steering. Kinematic and behavioral data of the controlled vehicle was recorded by the simulator's software program and converted to a spreadsheet data set allowing analysis of mathematical determinants from the vehicle. The simulation display provided a view of the road and vehicle dashboard instruments (Figure 2). The simulated vehicle was set to automatic transmission so participants were not required to adjust the gear lever. Participants controlled the vehicle by moving the steering wheel and manipulating accelerator and brake pedals. Participants were instructed to stay in the center of the left-hand lane and adhere to all normal road rules and speed signs. A GPS in the scenarios provided audio and visual (arrow) directions for the itinerary. Crashes into other vehicles would result in the presentation and sound of a shattered windshield. The program then reset the car in the centre of the left lane at the point of the crash and allowed the participant to resume driving.

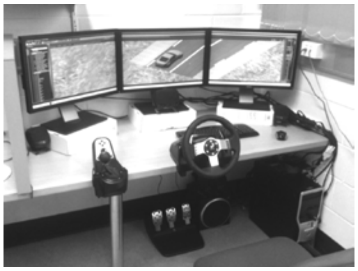

Figure 1. Driving simulator set-up

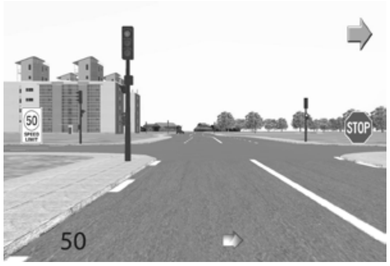

Figure 2. Visual simulation display

In the test drive tasks, participants completed a 10 km course, which took approximately 15 min. The driving scenario was set in daylight conditions and comprised six main sections (Table 1). Other vehicles and pedestrians were present in the scenario but did not actively interact with the participant's vehicle.

Table 1. Driving simulator scenario for experimental test drives

Description	Length (Km)	Configuration	Critical Events
1. Familiarization	3.00	2 lane single carriageway. 50-100 km/hr. 2 intersections with traffic signals, 1 intersection with stop sign. Few buildings. Light traffic present.	2 RI + 2 RT events
2. Highway	0.55	2 lane single carriageway. 80 km/hr. Few buildings. Light traffic present.	1 RI event
3. City	0.70	4 lane dual carriageway. 50 km/hr. 5 intersections with traffic signals. Many buildings. Moderate traffic present.	1 RT event
4. Rural	2.20	2 lane single carriageway. 50 km/hr. 4 intersections with traffic signals, 1 intersection with stop sign. Few buildings. Light traffic present.	1 RI + 2 RT events
5. Highway	2.60	2 lane single carriageway. 80 & 100 km/hr sections. 1 intersection with traffic signal. Few buildings. Light traffic present travelling in opposite direction.	1 RI + 1 RT + 1 Headway* event
6. City	0.95	4 lane dual carriageway. 50 km/hr. 4 intersections with traffic signals. Many buildings. Moderate traffic present.	2 RI + 1 RT events

Example of critical events from one scenario. Parallel versions differed in arrangement of critical events. RI – response inhibition, RT – reaction time. * occurred in a section separate from RI and RT events. Light traffic = 2-3 vehicles, Moderate traffic = 6-8 vehicles.

Critical Events

Critical events were included at random intervals within the scenario to test participant's reaction time and response inhibition behavior. To reduce the predictability of the critical events, three parallel scenarios were used, with the events occurring in different sections of the driving task for each version. In addition, the three parallel versions of the driving test scenario were randomly assigned to the three testing times for each participant. During the simulated test drives, participants were required to respond to stimuli on five occasions to test reaction time. For each reaction time event, the stimulus was the presentation of a stop signal image on the right side of the centre screen. Participants were instructed to brake as quickly as possible when the stimulus appeared. Once they had come to a complete stop, the stimulus disappeared and participants could resume driving. On five separate occasions participants were presented with a response inhibition task. For each event, a stop signal image was presented on the right side of the centre screen. A short auditory tone was played after a 400 millisecond delay and participants were instructed to withhold their brake response to the stop signal stimulus if they heard the auditory sound. This test provided a measure of participant's ability to inhibit a pre-potent response and errors (depressing the brake pedal when the visual and auditory stimuli were both present) were recorded as the number of incorrect inhibition responses. On one occasion during test drives, participants encountered a vehicle on the road ahead of them travelling at a speed set 10 km/hr below the designated speed limit. This event was set to occur at a pre-defined location on a single carriageway road with solid centre line markings to avoid having the participant overtake the vehicle. Participants were required to follow for a total distance of 1.5km. This event was used to examine car following behaviour, with time to collision (TTC) between the front of the interactive vehicle and back of the lead vehicle measured.

During the simulated test drives, participants encountered 15 intersections. One had a stop sign and required the driver to stop completely before resuming driving. The other 14 intersections were equipped with traffic lights. At five of the intersections, the traffic light was red and required the driver to stop. At three intersections the traffic light was green and did not require the driver to stop. At the remaining six intersections, the light turned from yellow to red as the vehicle approached with enough time for the driver to stop. Order of the traffic lights was randomly allocated throughout each test drive. Failing to stop at intersections was recorded (total stops required = 12) and the total number of stopping failures was calculated for each test drive. Several other measures of driving performance were also obtained during the driving tests including average speed, standard deviation of lane lateral position (SD lane position), standard deviation of steering angle (SD steering angle), the number of center and side line crossings, and the number of off-road and other vehicle impacts.

Data Analysis

All statistical procedures were performed using SPSS for Windows, Version 19.0 (SPSS Inc., Chicago, IL). Differences between trials for each of the main dependent variables in the driving task were examined using one-way repeated measures analysis of variance (ANOVA). Pair-wise comparisons (Bonferroni) were performed where significant main effects were present. Effect size was reported as partial eta squared (η_p^2). Intra-class correlation coefficients (ICC) were calculated using the two-way mixed average measures (absolute agreement) model. Coefficients of variation for each of the driving performance measures representative of continuous data were calculated by standard methods using the mean and standard deviations of each variable across the three trials. Statistical significance was accepted at $p<0.05$. All data are reported as mean±standard deviation unless otherwise specified.

RESULTS

All participants completed the three test drives with no complications or simulator sickness reported. Off road and other vehicle impacts were extremely rare ($n=2$), thus precluding any statistical analyses. There was a significant reduction in lane position deviation observed in test 3 compared to test 1 ($p<0.05$). Participants had more center and side line crossings in test 1 compared to the two subsequent tests ($p<0.05$), and a reduction in the number of incorrect inhibition responses (braking when a stop signal stimulus and inhibitory auditory tone was present) was observed in test 3 compared to test 1 and test 2 ($p<0.05$). No difference was seen in performance on this task between test 1 and test 2 ($p>0.05$). No significant differences were observed between tests for any of the other driving performance measures assessed ($p>0.05$).

Table 3. Analysis of practice effects in repeated driving performance tests

Performance Measure	Test 1	Test 2	Test 3	ANOVA		η_p^2
	mean (SD)	mean (SD)	mean (SD)	F (2, 25)	Sig	
Average speed (Km/hr)	42.72 (1.82)	43.13 (2.24)	43.35 (1.51)	2.580	ns	0.17
SD Lane position (m)	0.35 (0.06)	0.34 (0.05)	0.33 (0.05) *	4.041	p<0.05	0.24
SD Steering angle (deg)	0.77 (0.07)	0.77 (0.06)	0.77 (0.06)	0.102	ns	0.01
TTC (s)	2.78 (1.28)	2.92 (0.96)	2.73 (1.07)	0.531	ns	0.04
Center and side line crossings (n)	5.93 (4.57)	4.44 (3.67) *	3.78 (3.52) *	9.998	p<0.05	0.44
Reaction time (s)	0.96 (0.11)	0.94 (0.11)	0.95 (0.14)	1.286	ns	0.09
Failures to stop (n)	0.07 (0.27)	0.22 (0.51)	0.11 (0.32)	0.792	ns	0.06
Incorrect inhibition responses (n)	1.67 (1.21)	1.96 (1.51)	0.67 (1.11) **	13.590	p<0.05	0.52

* denotes significant difference compared to Test 1 (p<0.05), ** denotes significant difference compared to Test 1 and Test 2 (p<0.05).

Significant moderate to high ICC's were found for most assessment measures, indicating good to excellent reliability (Table 4). However, ICC values for the number of failures to stop outcome measure show low levels of test-retest reliability. The degree of variability in individuals' performance across driving tests was determined using coefficient of variation (CV). A low degree of intra-individual variability was observed for all performance measures except TTC performance.

Table 4. ICC and CVs for three test drives

Performance Measure	ICC	95% CI	CV (%)
Average speed	0.69*	0.43 - 0.85	2.4
SD Lane position	0.92*	0.84 - 0.96	5.7
SD Steering angle	0.91*	0.83 - 0.96	2.9
TTC	0.77*	0.56 - 0.89	17.7
Center and side line crossings	0.88*	0.75 - 0.95	-
Reaction time	0.74*	0.51 - 0.88	6.9
Failures to stop	-0.47	-1.80 - 0.28	-
Incorrect inhibition responses	0.47*	0.06 - 0.73	-

* denotes significance at the p<0.05 level.

DISCUSSION

Overall, most of the driving performance measures in this study demonstrated moderate to high test-retest reliability. Participants were able to maintain consistent speed, vehicle control (lane position, steering angle), and response time to critical events across repeated tests. Similar to previous work by Tornros (1998) and Marcotte et al. (2003), high test-retest reliability was observed for speed and lane position. The results also support the work of Akinwuntan et al. (2009) who observed high test-retest reliability for reaction time responses during driving. Whilst a high ICC coefficient was observed for TTC to lead vehicles indicating high test-retest reliability, the high CV value for this variable suggests a large intra-individual variation in car following behavior. Recent work by Brackstone et al. (2009) suggests that drivers are inconsistent in their choice of headway, with individual variations above 19% in adopted headway observed between trials in their study. Collectively, these results suggest that driving headway is likely to be susceptible to intra-individual differences. A low ICC coefficient was observed for the number of stopping failures in this study. Given these findings, this may have implications for the use of this measure as a performance variable in future studies. However, given that there were very few stopping failure instances across all three drives it is possible that decision errors or misjudgments by participants (they thought they could clear the intersection before the red light but failed) explain the observed differences and low reliability.

In the present study there did appear to be some influence of practice on a number of performance measures. Lane position deviation was lower in drive three compared to drive one. In addition, the total number of line crossings was higher in the initial test drive compared to the subsequent drives. Whilst participants completed a single familiarization drive prior to the test drives, these results suggest that inclusion of an additional familiarization drive may help to reduce any learning effect. Participants also had a greater ability to correctly withhold their brake response to inhibition stimuli in the final test drive compared to the first two drives. However, a true practice effect would assume directional change as trials progressed. Participants made fewer incorrect inhibition mistakes in the final drive compared to the first and second drives, yet an increase was observed from drive one to drive two. A possible explanation for these results may relate to the type of task used. Response inhibition tasks involve a reaction component in addition to a measure of accuracy. As such, participants may adopt different strategies that ultimately results in a speed-accuracy trade off (Rabbitt & Vyas, 1970). Reaction time for brake pedal press during the response inhibition task was not measured in this study, but may explain the differences observed between trials. It is possible that participants adopted strategies on the final test drive where speed of response was forfeited to allow fewer errors to be made. Further investigation of test-retest reliability for response inhibition tasks during simulated driving are needed to clarify these results.

One of the limitations of this study is that compliance to pre-experimental conditions was verbally acknowledged. These would be better verified with objective measures (e.g. breath analysis for alcohol, plasma analysis for caffeine). In addition, it is important to acknowledge that this study involved a desktop computer based simulator and it is likely that larger, very-high fidelity simulators with greater fields of view are more realistic of real world driving and may be more sensitive to the measures assessed in this study. The test-retest reliability results presented in this study are based on the equivalent absence of differences between test drives. Conclusions may have been strengthened if the effects of a treatment (e.g. alcohol consumption) had been shown to be consistent across repeated testing, thus demonstrating equivalent sensitivity of effects rather than the equivalent absence of differences. Finally, this study involved a naturalistic drive and participants were given minimal instructions on how to drive during the scenarios, providing no task priorities, incentives or performance feedback. More traditional testing protocols typically involve a highly constrained situation in which test participants have little freedom to choose responses. Driving behavior is different, particularly when it is relatively unconstrained and the absence of differences between some metrics measured in this study may be due to relatively high levels of variability. The use of a driving simulator protocol with more constrained instructions may be better for the purpose of measuring test-retest reliability, reducing driving variability.

In summary, the findings from this study suggest that the driving scenario used provides assessment tasks that may be reliable for tracking the effects of pharmacological treatments on driving abilities, when test-retest assessments are made following relatively short delay periods. The driving scenarios developed in this study will be used to examine the combined effects of dehydration and alcohol consumption on driving performance in a future study.

REFERENCES

Akinwuntan, A. E., Tank, R., Vaughn, L., Wilburn, A., & Easton, S. (2009). Normative values for driving simulation parameters: A pilot study. *Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Big Sky, Montana, 161-168.

Allen, R., Rosenthal, T., & Cook, M. (2011). A Short History of Driving Simulation. In D. L. Fisher, M. Rizzo, J. K. Caird & J. D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. Boca Raton, FL: CRC Press.

Bedard, M. B., Parkkari, M., Weaver, B., Riendeau, J., & Dahlquist, M. (2010). Assessment of driving performance using a simulator protocol: validity and reproducibility. *American Journal of Occupational Therapy, 64*(2), 336-340.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., Fastenau, P. S., & Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology, 20*(4), 517-529.

Brackstone, M., Waterson, B., & McDonald, M. (2009). Determinants of following headway in congested traffic. *Transportation Research Part F: Traffic Psychology and Behaviour, 12*(2), 131-142.

Caird, J. K., & Horrey, W. (2011). Twelve Practical and Useful Questions About Driving Simulation. In D. L. Fisher, M. Rizzo, J. K. Caird & J. D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. Boca Raton, FL: CRC Press.

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society, 9*(3), 419-428.

Grandjean, A. C., & Grandjean, N. R. (2007). Dehydration and cognitive performance. *Journal of the American College of Nutrition, 26*(5 Suppl), 549-554.

Lieberman, H. R. (2010). Hydration and human cognition. *Nutrition Today, 45*(6 Suppl), 33-36.

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. Cambridge Neuropsychological Test Automated Battery. International Study of Post-Operative Cognitive Dysfunction. *Neuropsychologia, 36*(9), 915-923.

Marcotte, T. D., Roberts, E., Rosenthal, T. J., Heaton, R. K., Bentley, H., & Grant, I. (2003). Test-Retest Reliability of Standard Deviation of Lane Position as Assessed on a PC-Based Driving Simulator. *Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Park City, Utah, 199-200.

Rabbitt, P. M. A., & Vyas, S. M. (1970). An elementary preliminary taxonomy for some errors in laboratory choice RT tasks. *Acta Psychologica, 33*(0), 56-76.

Törnros, J. (1998). Driving behaviour in a real and a simulated road tunnel—a validation study. *Accident Analysis and Prevention, 30*(4), 497-503.