Driver performance under simulated and actual driving conditions: validity and orthogonality.

John A Groeger[1,2] and Gillian Murphy[2]

Department of Psychology,

Nottingham Trent University,

50 Shakespeare Street

NOTTINGHAM, NG1 4FQ

[2]School of Applied Psychology,

University College Cork,

Cork, Ireland.

Corresponding Author:

John A Groeger, PhD, FBPsS

Keywords: Driving, simulation, validation, skilled behaviour, transfer of training.

Running head: *Driving sub-tasks*

Driver performance under simulated and actual driving conditions: validity and orthogonality.

Abstract

This study contrasted the performance of drivers under actual and simulated driving conditions, in order to assess the validity of the simulators and test the hypothesis that driving is composed of largely orthogonal sub-tasks. Thirty experienced drivers completed an on-road driving test and drove two different simulators, each simulator drive comprising seven difficulty-moderated driving scenarios. Between-simulator contrasts revealed largely absolute validity, the anticipated effects of increased difficulty within driving scenarios, but weak relationships between performance of different driving scenarios. On-road driving was reliably assessed by a nationally-recognised expert driving assessor, as reflected by standard statistical measures of reliability and consistency. However, on-road driving revealed relatively little cross-category correlation of on-road driving errors, or between on-road and simulator driving. Thus, despite the compelling evidence of absolute and relative validity within and between simulators, there is little evidence of criterion validity (i.e. relationship to on road driving, as assessed by the expert assessor). Moreover, the study provides strong evidence for orthogonality in the driving task- driving comprises large numbers of relatively separate tasks.

INTRODUCTION

Driving simulators continue to offer immense practical and theoretical promise to the study of driving (see Fisher, Rizzo, Caird, & Lee, 2011). However, as has been argued elsewhere, decades of development have not been matched by the realisation of this promise with regard to the development of theory (e.g. Lenné, Groeger & Triggs, 2011). In part this is because of an age-old problem in psychological research- the relevance to, and validity of, the measurements taken in the laboratory to the unconstrained real world performance of the task we are seeking to understand. A practical problem, i.e. determining the validity of a newly developed driving simulator, and theoretical challenge, i.e. the possible lack of transfer of competence within the driving task, motivated the study reported below.

Driving simulation has a surprisingly long history; the first simulator was developed in 1934 (Allen, et al, 2011). In that same volume (i.e. Fisher, Rizzo, Caird, & Lee, 2011), which is replete with examples of how ingeniously driving simulators have been used, it is conceded that "Probably the most important question though is whether driver behavior in a simulation mimics that which is exhibited while driving in the real world" (Caird & Horrey, 2011, p. 7). Mullen, Charlton, Devlin & Bédard (2011), attempt to answer precisely this question by reviewing thirty-three studies, some three-quarters of which were published in peer-reviewed journals, in which some form of validation was attempted. Many of these studies, together with more recent examples (McWilliams, Ward, Mehler & Reimer, 2018), seek to establish whether specific changes that occur during simulated driving also occur during actual driving (e.g. effect of a secondary task), on a single aspect of driving (e.g. speed, lateral position, reaction time), or at a specific location (e.g. road works, tunnel), or in particular groups of motorist (e.g.

age or patient groups). Sometimes studies have sought to quantify the relationship between simulator performance and more distal measures of driving (e.g. crash history, speeding violations, self-rated ability). Generally, studies use statistical methods which compare average tendencies across simulator or on-road conditions, or correlations between such measures of central tendency in simulated and actual driving. Only six of the studies reviewed demonstrated 'absolute' validity (i.e. absence of statistically reliable differences between two circumstances), three or these related to speed, and one each for physiological responses, hazard ratings, and brake reaction time. Interpreting an absence of difference is obviously fraught with challenges, since failure to find a difference may reflect lack of statistical power, or that the two circumstances are actually not different. Most studies reported some measure of 'relative' validity (i.e. correlation), generally for some particular aspect of driving (i.e. speed, lateral position, time). The conclusion Mullen and colleagues draw is that "driving behaviour in simulators approximates (relative validity), but does not exactly replicate (absolute validity), on-road behaviour", claiming that "This is sufficient for the majority of research, training and assessment purposes for which simulators are used" (op. cit. p. 15). They later go on to caution that "each simulator set-up should be validated for its ability to measure the driving behaviour of the cohort for which it is to be used" (op. cit., p.16). In essence, the issues we raise below address both 'sufficiency' and the 'necessity', as well as a broader spectrum of how validity can be construed.

At its heart, "validity", is the expectation is that performance in one circumstance will be related to performance in another. More specifically, as mentioned above, the relationship may be 'absolute', such as when the simulated and on-road driving environments yield the same numerical values for a specific behaviour, or 'relative', in that the numerical values obtained are not identical, but related in terms of direction and magnitude of the effects of some manipulation. Absolute and relative validity may be demonstrated concurrently, or predictively, such that measurements at one time point correlate significantly with the same measures taken at a later point. Such 'behavioural' validity, exhibited over an extended period of time and broader circumstances, reflects 'ecological' validity. This is typically claimed where the more specific measures in a constrained circumstance, such as a driving simulator, correlate with more general measures on-road. Task measurements which show consistent relationships across circumstances reflect 'convergent' validity, that is, lateral position while driving on bends of identical curvature would be correlated. Finding a pattern of non-significant correlations across circumstances, where these were expected not to be significant, would reflect 'discriminant' validity. That is, where driving speeds on sets of curves, while related to each other, are unrelated to the speeds at which drivers drove sections of straight road (these being correlated with each other, see Campbell & Fiske, 1959 for discussion of this multi-trait multi-matrix approach to convergent and divergent validity). Two other aspects of validity are typically distinguished: 'content' validity, in which an assessment is made that what is measured in a given circumstance reasonably reflects the full extent of what is involved in that circumstance; and 'face' validity, where a judgement is made as to whether what is measured in one circumstance is superficially similar to that which is measured in another. Together, absolute, relative, convergent, discriminant, ecological, content and face validity encompass what Cronbach & Meehl (1955) describe as 'construct validity'. This list of different types of validity is not quite as comprehensive as it might first seem. Elsewhere, Guion (1980), identifies criterion-related-, construct-, and content-validity are reflecting a 'holy trinity' of validity representing: some external variable of interest one might wish to predict (criterion-related); corroborating evidence from other measured presumed to assess related or unrelated variables (construct-validity) and the extent to which what is measured reflects the domain as a whole (content-validity). As we understand these distinctions, 'criterion-related' is similar to ecological/behavioural validity, but carries with it the notion that the variable predicted must be external or distinct, rather than just mere replication; 'content validity' assumes some attempt to assess a whole domain rather than a distinct behaviour, while 'construct-validity' encompasses degrees of absolute, relative, convergent and discriminant validity.

This study reported below set out to assess the validity of two simulators- one in which the driver sat in a full-sized vehicle with a horizontal visual environment of some 300^0, the other a desktop emulation of the full-sized simulator, where three screens showed a 60^0 field of view of the forward scene. Both simulators ran the same software, which enabled identical multidimensional measurement of a driver's actions when driving through a set of driving sub-tasks which required different extents of interaction with the driving environment. Following the principles laid out by Guion (1980) and others discussed above, in order to assess construct-related validity, this interaction is quantified in a variety of ways, including measures of central tendency and variability, but typically with regard to speed, lateral position and decision making. Each of these seven sub-tasks: using the horn to respond to a brake light onset, following a lead vehicle whose speed varied systematically, responding to a car emerging from a driveway, a pedestrian crossing the road, driving ahead on straight or curved roads, and changing lane in response to a road-side message sign, involved within-task manipulations of difficulty. These seven scenarios may not represent the full panoply of the driving task, but they are a more extensive investigation of content validity than has hitherto been attempted. These different driving sub-tasks were selected on the basis of having similar or different 'common-elements'. Finally, on-road driving competence was assessed, as a means of establishing criterion-related validity, Drivers drove in both simulators, as well as an on-road urban route, in the presence of a state-registered, highly qualified driving assessor, who recorded the seriousness of errors committed as they drove his standard, dual-control, test vehicle.

We predicted, on the basis previous attempts to investigate simulator validity (Mullen et al. 2011), that, relative, if not absolute validity would be established between the two simulators, and that content validity between both simulators would be far greater than that between either simulator and the on-road test, because almost identical requirements are made of the driver during simulation. We would also expect that criterion-related validity (i.e. correspondence between simulated and on-road driving) would depend on the extent to which the driving task, as a whole, is measured in each circumstance.

METHOD

Participants

Thirty participants completed the study (20M/10F), ranging in age from 18-27 years (M = 24.8, SD = 5.44), all had held full driving licenses for at least one year. Men and women were similar in terms of age t(28) = .42 (p>.6), and driving experience in terms average number of kilometres typically driven per week (t(17) = -.71, p >.5; M = 181km, SD = 217km), per year (t(18) = .55, p>.5; M = 10657 km, SD = 9089 km), and years' licensed (t(28) = .86, p>.4; M = 7.3yrs, SD = 5.1). All participants completed three drives, lasting approximately 30 minutes each; one on actual roads, and one in each driving simulator. The order of these drives was fully counterbalanced. In each case, before commencing the test route, or driving in simulators, participants had a brief (c10 mins) familiarisation drive.

On road driving

Participants drove for approximately 30 minutes along a fixed route comprising urban, suburban and rural stretches of roadway which was judged to be of equivalent challenge to that typically encountered as part of the state driving examination. That judgement was made by a highly experienced, nationally recognised, driver trainer and assessor. In addition to his leading status in the profession, the assessor had undertaken extensive training in driver instruction and assessment. He accompanied participants while they drove an otherwise standard, dual control, family saloon- typically used for driver training. As they drove, the assessor, noted the type and seriousness of errors committed en route, following the state driver assessment protocol (Road Safety Authority, 2009), but provided no feedback or comment at any point.

To be entitled to drive alone, drivers in in Ireland, must pass a practical on-road assessment of their driving competence. This was the test used in the current study and the driving assessor had significant experience in preparing candidates for this test, and in assessing the fitness of neurological patients to drive. Driving faults were recorded in categories according to the manoeuvre/traffic circumstance[1], the nature of the error and its seriousness (see Road Safety Authority, 2009). Faults are regarded as: Grade 1 fault:- Minor fault (i.e. a mistake which does not cause immediate danger, e.g. driving in

[1] Categories errors with respect to: Rules/Checks, Positioning, Observation, Reaction to Hazards, Mirrors, Clearance /Overtaking of objects, Signalling, Alighting, Making Progress, Vehicle Controls, Speed, Traffic Controls, Right of Way, Reversing, Turning about in road, Parking

incorrect gear), Grade 2 fault:- More serious fault (combination of minor faults in same aspect of driving or potentially dangerous error), Grade 3 fault:- Dangerous/Potentially Dangerous fault (i.e. immediate danger to self or others), or total disregard of traffic controls. Failure of the test arises when an applicant incurs any of the following: 1 or more grade 3 faults, 4 of the same grade 2 faults for a single aspect, 6 or more grade 2 faults under the same category, 9 or more grade 2 faults overall. Grade 1 faults do not affect the test result and were not recorded during this study.

Driving Simulators

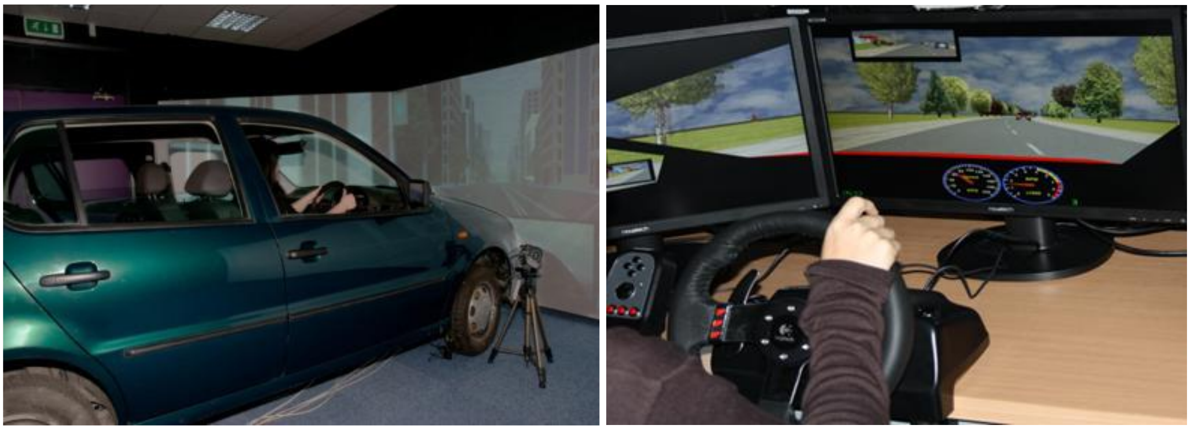


Figure 1. Full-Scale (side on view) and Desk-Top Simulators (driver's view)

Both simulators ran the same software (STISIM 400W) and driving scenarios, and differed only in respect of the physical setting in which driving took place. In the Full-Scale Simulator, drivers sat in a real vehicle, the controls of which were connected to the STISIM console which enabled the force-feedback steering wheel, pedal force registration, etc. The visual environment (160^0 forward view) was augmented with active side view mirrors (where the actual mirrors were replaced by similarly sized VDU taking a live feed from video cameras of the mirror scenes displayed outside the driver's field of view), and a rear screen on which was back-projected the rear view scene such that it would be fully visible when the driver used the rear view mirror. The Desk-Top Simulator used three linked PC screens (45cm, c100^{o} Field of view), using a Logitech27 steering wheel with turn sign indicators, gear lever, as well as accelerator, brake and clutch pedals emulating standard foot controls. Both simulators used identical auditory output presented from speakers immediately behind or ahead of the driver.

Simulators have the capacity to yield huge quantities of data depending on the data rate, and numbers of parameters specified when developing individual scenarios. There is therefore considerable potential for multiplicity and other statistical problems. The seven traffic scenarios analysed here were also driven by other participants in other studies (N<>120) and this allowed us to address these issues. The scenarios described below identify particular indices of driving performance that account for most of the common variance in scenario specific PCA analyses of all 40+ data parameters across all participants across several studies.

Simulated Driving Tasks

Each participant drove a fixed distance (sub)urban route composed of seven closely specified driving tasks, which were randomly ordered across six different versions of the route. Orders were identical in both simulators. Unless required by the specific driving task, all roads were straight and had identical lane widths (3m), with randomly occurring traffic, pedestrians on pavements, trees and shops or housing. Participants were randomly allocated to different versions for their simulator drives, subject to the constraint that all versions of the drives were used an equal number of times across the experiment as a whole. The order of on-road testing and simulator drives was fully counterbalanced across the study.

Free driving: The driver drove along 300m straight stretches of roadway, where the posted speeds were 50, 60 or 100 Kph. Dependent variables were the average speed and its standard deviation over each section.

Braking Task: Participants were instructed to follow a lead vehicle, which travelled at a sinusoidally varying speed below the posted speed (50,60 100Kph), at a safe distance. As it did so, at unpredictable intervals which ranged between 1 and 3 seconds, the lead vehicle brake lights illuminated, and the participant was required to respond to this, by using the horn. There were 12 of these brake light reacting time tasks. Responses following a brake light onset were considered to be 'hits', unless the response time was absent, or longer than 3 seconds, in which case they were treated as 'misses'; those preceding a brake light onset were considered 'false alarms'. Dependent variables were the response latency for 'hits', and the proportion of hit-responses.

Car-following Task: When driving through some 50, 60 and 100 Kph zones, a parked sedan facing in the same direction, pulled into the roadway ahead of the driver's vehicle. The stream of oncoming traffic was such that overtaking was impossible, and the changes in speed of the lead vehicle, and vehicles behind, ensured that the driver followed the vehicle ahead. All other vehicles obeyed the posted speed limit, but their speed varied systematically such that collisions would never occur, nor would the lead vehicle exceed a maximum headway. The variation in vehicle speeds and headway criteria was identical for each speed and across simulators in order to allow meaningful comparisons within and between simulators. The headway (i.e. temporal distance between the lead and following vehicle) and lateral position of the following vehicle were quantified in terms of their mean and standard deviation, and for headway the headway and relative standard deviation (i.e. coefficient of variation, SD headway/ mean headway).

Pedestrian Obstruction Task: On six occasions, during urban sections of the route, an adult pedestrian attempted to cross the road at a speed equivalent to 5 kph, walking into the carriageway in front of the driver from the adjacent sidewalk. This occurred such that the driver, at current speed, would have 2 or 4 seconds to respond before colliding with the pedestrian. Crossings occurred in 50, 60 and 100 Kph speed zones. Drivers were free to drive at whatever speed they chose, and to respond by changing their vehicle's speed using the foot brake, or their vehicle's course by using their hands to steer their vehicle. Removing their foot from the accelerator alone would not slow the vehicle sufficiently, and a steering response would require that the centre line was crossed. In principle, had more than one response occurred, the earlier of the two responses would have been considered primary, but this criterion was not required. Collisions between vehicles and pedestrians never occurred, because had the response occurred very late or not at all, the pedestrian was programmed to move sideways rapidly to avoid being hit. Response latency, variability in response latency, failures to respond and the preponderance of brake/steering responses were treated as dependent variables.

Vehicle Obstruction Task: On six occasions a saloon vehicle, parked in a driveway outside a block of housing, facing the house, reversed into the driver's path. Vehicles emerged at 20 Kph, with movement beginning when the driver was 2s or 4s from a potential collision. As with the Pedestrian task, these events occurred in 50, 60 and 100 Kph speed zones, with all other aspects of the emerging vehicle task equivalent to the pedestrian task already described. Dependent variables were as for the Pedestrian version of this task.

Lane-change Task: In this task the simulator fixed the speed (60 Kph) at which participants drove along a straight, three-lane highway. At variable intervals a roadside message sign activated showing arrows that indicated that drivers should steer the vehicle into a particular lane other than that in which they were travelling (i.e. centre to left lane, centre to right, right to centre, left to centre, right to left, left to right, each on two occasions). Performance was quantified in terms of whether the instruction was followed correctly, the decision time (appearance of instruction to initial steering movement, ms), completion time, and steering mean and variation when undertaking the manoeuvre.

Curve-driving Task: Each version of the route included nine right and left curves, made up of three in each speed zone (i.e. 50, 60 100Kph), which had radii of 7, 4 and 3, metres, necessitating respectively greater steering adjustment. Across curve direction (i.e. to drivers left or right) was systematically varied across the order-controlled drives. Dependent variables, mean and standard deviation of forward speed and steering, were calculated over four quadrants of the curve's phase (e.g. initial phase, to apex, from apex, final phase). For brevity the directional analyses are not presented here. There were no pedestrians, parked vehicles, housing or trees on roadsides.

Throughout all simulations, data reflecting the state of all standard STISIM driving parameters (e.g. longitudinal, lateral velocity, road position, pedal forces, steering wheel movements, proximity to other vehicles etc.), were recorded at 20Hz.

RESULTS

Inevitably, with a task as complex as driving, there are multiple ways of quantifying performance, with several possible indices from individual tasks (e.g. averages or variability of speed, distance, time etc.). This makes it very difficult to compare across different driving scenarios. Because this is exactly what this paper seeks to do, in each case we report on the measures typically used for any given scenario and then combine all the measures used for a given scenario, into a single composite measure. On-road driving is considered first, before detailed presentation of the effects of increasing difficulty in each of the separate simulated driving scenarios. The final section relies on the aforementioned composite measures to compare and contrast driving in the two different simulators and on-road.

On-road driving

Details of driver performance for those of the Road Safety Authority (RSA)-designated eighteen error categories in which one or more Serious or Dangerous fault was committed are presented in Table 1. Cronbach's Alpha (0.74) suggested an acceptable level of coherence across these assessment categories, indicating a high internal consistency of the categories taken as a whole. Averaging over sub-samples of drivers revealed split half correlations exceeding 0.9 (First/second half .94, Odd/Even .99, both p<.001), indicating that the assessor performed consistently across driving assessments. Internal consistencies within categories was, however, quite variable, only road-positioning (Cronbach's Alpha= 0.74), and perhaps mirror use (Cronbach's Alpha= 0.65) categories are internally consistent.

Table 1 Serious and Dangerous Faults during on-road driving

	Serious Faults		Dangerous Faults		Overall	
	Mean	SD	Mean	SD	Mean	SD
Controls	4.70	3.80	0.03	0.18	4.07	3.41
Clearing/ Overtaking	5.07	5.18	0.07	0.25	4.88	4.29
Hazard Reaction	4.00	3.78	0.07	0.25	3.9	3.29
Traffic Controls	0.1	0.31	0.07	0.25	0.21	0.47
Mirrors	1.13	1.28	0.03	0.18	1.07	1.09
Signalling	2.20	2.86	0.03	0.18	2.21	2.47
Observation	0.70	1.24	0.03	0.18	0.90	1.24
Progress	4.00	2.84	0.17	0.38	4.14	2.99
Right of way	0.17	0.53	0.00	0.00	0.22	0.54
Parking	0.20	0.48	0.07	0.25	0.26	0.49
Total	22.27	15.41	0.57	0.90	22.83	15.88

Of the categories shown in Table 1, fault frequencies are correlated for only a small number of categories. Faults in negotiating Rights of Way and in compliance with Traffic Controls were reliably correlated ($r_{(28)}$= 0.44, p<.05), as were faults in Observation and Mirror use (r=0.38, p<.05), use of Vehicle Controls ($r_{(28)}$=0.45 p<.05) and problems overtaking/clearing other vehicles or obstacles ($r_{(28)}$=0.45 p<.05). Those who make inadequate progress also tend to have poor road positioning ($r_{(28)}$= 0.45, p<.05). Obviously, while these correlations are plausible, the fact that just 5 from a possible 36 correlations are statistically reliable raises the possibility of these merely reflecting type 2 errors, rather than genuine relationships, and it is noteworthy that none of the five survive statistical control for the False Discovery Rate (henceforth FDR; see Benjamini & Hochberg, 1995).

Finally, 11 participants were responsible for the 17 Dangerous driving faults recorded. Seven participants committed a single Dangerous fault. Two participants committed 2, and two committed 3, Dangerous faults. Dangerous faults were largely concerned with dangerous control of Speed/Distance (6), poor observation (5), improper use of Right of Way (2) or Road Positioning (2). The numbers of Dangerous faults under each heading was correlated with the number of Serious faults under these headings ($r_{(18)}$ = 0.68, p<.05), as were the numbers of Serious and Dangerous faults committed by individual participants ($r_{(30)}$= .51, p<.01). Committing a Dangerous fault during the actual state driving test results in test failure. Here those who would have failed their driving test on this basis also committed more Serious errors (Mean= 31.18 $\underline{+}$17.67; 17.11$\underline{+}$11.51, $t_{(28)}$ 2.65, p<.01). As mentioned above, failure in Ireland's on-road driving test can arise through commission of Dangerous errors, or substantial numbers of Serious errors overall (>9), or repeated Serious errors within a single category. All drivers who committed Dangerous errors would also have failed under the multiple-serious-fault criterion, but only two had 4+ repeated faults in the same category. Eight other drivers had both multiple within-category faults, and had committed more than 19 serious errors overall. These too would have failed their driving test under current RSA rules.

Thus, in summary, on road driving was reliably assessed, but commission of a particular type of error was relatively independent of the number of faults that driver makes in a different category. Slightly over half of all participants drove sufficiently badly as to have failed their driving test.

Simulated Driving: Construct Validity

Driving in the seven difference scenarios will be considered in turn. A range of different measures of performance will be reported in each case, as appropriate to the behaviour in each scenario. In each case the effect of some systematic change to the driving behaviours in question will be assessed (i.e. relative validity) as well as quantifying the difference in performance in the two simulators (i.e. absolute validity). The development of composite measures of performance for behaviour each scenario will be described. When all scenarios have been considered, the correlations between and within simulators will be reported (i.e. convergent and discriminant validity). We then re-consider on-road driving, our criterion measure, attempting to integrate these data sources, with a view to establishing the degree of construct-, content- and criterion-validity observed.

Free Driving: The average speed at with participants drove differed across speed zones ($F_{(2,58)}$= 183.29, p<.001, η_p^2= 0.74; 50Kph: 48.37Kph ± 6.69 < 60Kph: 53.22±9.99 <100Kph: 76.51 ± 15.57, p<.001, see Figure 2). Speed stability or variation also differed across speed zones ($F_{(2,58)}$= 22.70, p<.001, η_p^2= 0.26; 50Kph: 3.08Kph ± 2.17 < 60Kph: 4.39±2.54 <100Kph: 6.17 ± 3.61, p<.001, see Figure 2). In neither case was there a statistically reliable main effect or interaction involving simulator type (all F<1). That is, for Free Driving, there is very good evidence of relative and absolute validity. [NB The average speeds reported above are below the posted speed limit in each case. Arguably, people might be expected to drive at the posted speed, and as such posted speed might serve as a criterion variable. One-sample t-tests showed that only the higher speed zones were statistically different from the posted speed: 50kph zone: $t_{(29)}$= 1.33, 60kph zone: $t_{(29)}$= 3.72*, 100kph zone: $t_{(29)}$= 8.26*, *both p<.001].

.

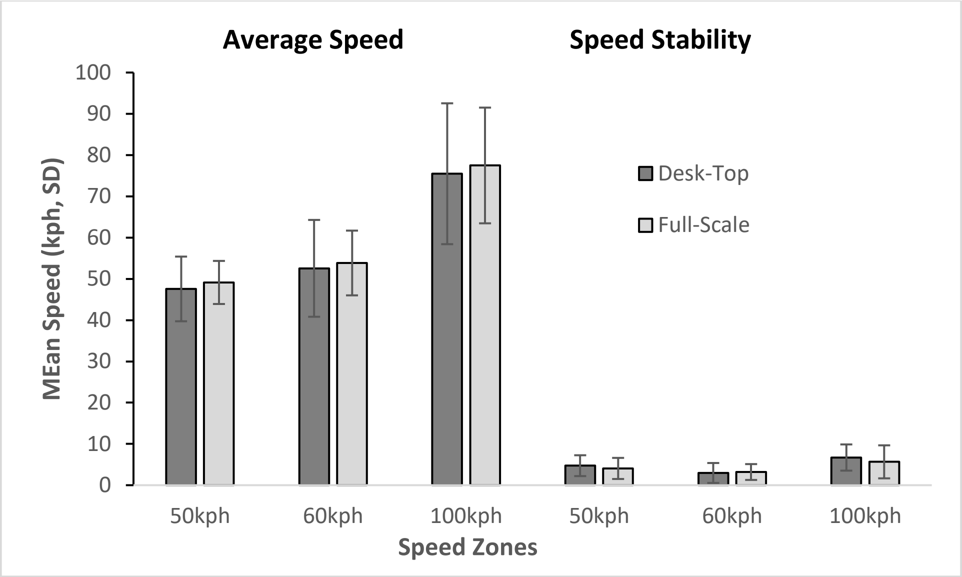


Figure 2 Driving speed, and its stability, as a function of posted speed and simulator type

For use in later analyses, raw average speeds were normalised across simulators, thus yielding a z-score for Speed. This was positively related to the actual speeds driven with correlations from $r_{(30)}$= 0.61 to $r_{(30)}$= 0.82 for the Desk-Top simulator, and $r_{(30)}$= 0.80 to $r_{(30)}$= 0.88 (all p<.001) across simulators.

Braking Task: Over 80% of brake light onsets were responded to in the Desk-Top (M= 92.5% ± 20.5) and Full-Scale (M= 84.1% ± 31.9, $t_{(29)}$= 1.13, p=.27) simulators, and responses which anticipated signals were few and similar in each simulator (Desk-Top: M= 2.17% ± 1.58) Full-Scale: M= 1.73% ± 1.55, t(29)= 1.18, p=.25). The 'hit-rate' (i.e. Hits/(False alarms + Hits)) was also similar ($t_{(29)}$= 0.94, p=.35). That is, there is evidence of absolute validity with respect to brake light detection.

Correct responses were slower in the Full-Scale (M= 958ms ± 284) than in the Desk-Top simulator (M= 842ms ± 319; $t_{(25)}$= 3.90, p<.001).

A derived measure, which combined both response speed and accuracy (i.e. average of Z-scored hit rate and reciprocal of Z-scored hit latency), did not differ between simulators (Desk-Top: M= .10 ± .601; Full-Scale: M= .16 ± .48, $t_{(25)}$= .62, p=.53). Response speed ($r_{(26)}$= .64, p<.001), and combined speed-accuracy ($r_{(26)}$= .58, p<.005) were correlated across simulators, but accuracy measures were not (Hits: $r_{(30)}$= -.15, p<>.4; Hit-rate: $r_{(30)}$= -.05, p<>.4). This suggests that while the simulators were very similar, the data indicate that the Desk-Top simulator may facilitate somewhat faster and more accurate responding to small targets (i.e. in this case the brake lights of the car ahead). Overall, there is evidence of relative, and to some extent absolute validity with respect to brake light detection.

Car-following Task: Steering behaviour was similar when following in both simulators and at each speed, both in terms of its average value and variation during the manoeuvre (all F<1, except steering variation in simulators $F_{(1,29)}$= 1.94, p= .17). That is, there is evidence of absolute simulator validity with respect to lateral position when following, and neither simulator showed an effect of following speed on steering.

The minimum headway adopted when following was similar in both simulators (570ms ± 25ms; F<1), but the average headway was longer in the Full-Scale simulator ($F_{(1,29)}$= 5.95, p<0.02, η_p^2= .17; Desk-Top 6.911s ± 163ms, Full-Scale 7.271s ± 109ms). There was a statistically reliable, but small, effect of speed on headway variability ($F_{(2,58)}$= 3.35, p<.05, η_p^2=.11), but no effect of simulator type. If headway variability is weighted by the mean headway by calculating a coefficient of variation, there was a substantial difference between the two simulators ($F_{(1,29)}$= 12.77, p<.001, η_p^2= .31), with headway being relatively more stable in the Full-Scale (0.72 ±.01) compared with the Desk-Top (mean: 0.76±.01) simulator. That is, with respect to longitudinal positioning when following a lead vehicle, there is some evidence of both absolute and relative validity.

Averaging across speed conditions, correlations between simulators for different elements of following performance were positive and significant only for longitudinal aspects of car following (Average headway: r(30)= .47, Minimum headway: $r_{(30)}$= .65, Headway variation: $r_{(30)}$= .52, all p<.005; Coefficient of headway variation: $r_{(30)}$= .38, p<.05) but not for lateral position (Steering variation: $r_{(30)}$= .12; Steering mean: $r_{(30)}$= -.05, both n.s.).

As before, a composite performance measure was calculated to reflect car following performance by normalising and then averaging scores for steering variation, average headway variation (coefficient of variation), and inverse of the mean and minimum headway, such that a low average score across these four elements indicates better performance. This measure was correlated across simulators ($r_{(30)}$= .41, p<.05).

Taken together these data confirm that the simulators display at least relative validity with regard to car following.

Pedestrian Obstruction Task: Almost all responses to the pedestrian's attempt to cross the road were either braking (78.2%), or steering (16%), drivers failed to respond in some 5.8% of cases. Overall, decision type was almost identical in the two simulators (Desk-Top: Brake-79%, Steer-16%; Full-Scale: Brake-79%, Steer-15%).

Braking likelihood decreased with speed ($F_{(2,56)}$= 8.52, p<.001, η_p^2= .23; Bonferroni corrected post hoc contrasts 50Kph, 60Kph> 100Kph, 92%,86%>73%), and this interacted with simulator type (see Figure 3). Other than a difference in braking likelihood which depended on pedestrian preview ($F_{(1,28)}$= 13.86, p<.001, η_p^2= .33, see Figure 3), there were no significant differences in decision making in the two simulators. This interaction shows braking to be the more likely response in the Desk-Top simulator when the pedestrian had been visible for longer, whereas the opposite response pattern was true of the Full-Scale simulator. This suggests that in the larger, more car-like, Full-Scale simulator, steering is a more viable avoidance strategy when the obstruction is distant, and braking is the preferred response when the obstruction is close. Thus, while the absence of differences suggests that decisions were similar in both simulators, thus reflecting absolute validity, different effects of speed and pedestrian preview undermines this conclusion.

Reaction times to the pedestrian movement are similar in both simulators ($F_{(1,28)}$= 2.56, p=.12, η_p^2=.08; Desk-Top: 890ms ± 373ms, Full-Scale: 778ms ±209ms), are slower at higher approach speeds ($F_{(1,28)}$= 14.91, p<.001, η_p^2=.345, 722ms ±277ms (60Kph)<> 732ms ± 238ms (50Kph)<1049ms ±411ms (100Kph)) and when the pedestrian was further away when his movement began (F(1,28)= 13.31, p<.001, η_p^2=.32; Preview 2s: 708ms, ± 202ms; Preview 4s: 961ms ± 378ms). None of the interactions between these variables approaches significance (all F<1), suggesting that behaviour in the two simulators is similar. Analysing only response time when the driver decided to brake, drivers averaged 728ms ±129 in the Full-Scale simulator and 767ms ±302 in the Desk-Top simulator across all conditions ($t_{(30)}$= 0.92, p= .36). A repeated measures ANOVA revealed that approach speed ($F_{(2,32)}$= 4.074, p<.05, .203) and preview ($F_{(1,16)}$= 8.67, p<.01, η_p^2=.35) both influence brake response times, but these main effects were not subject to any interactions (all F<1).

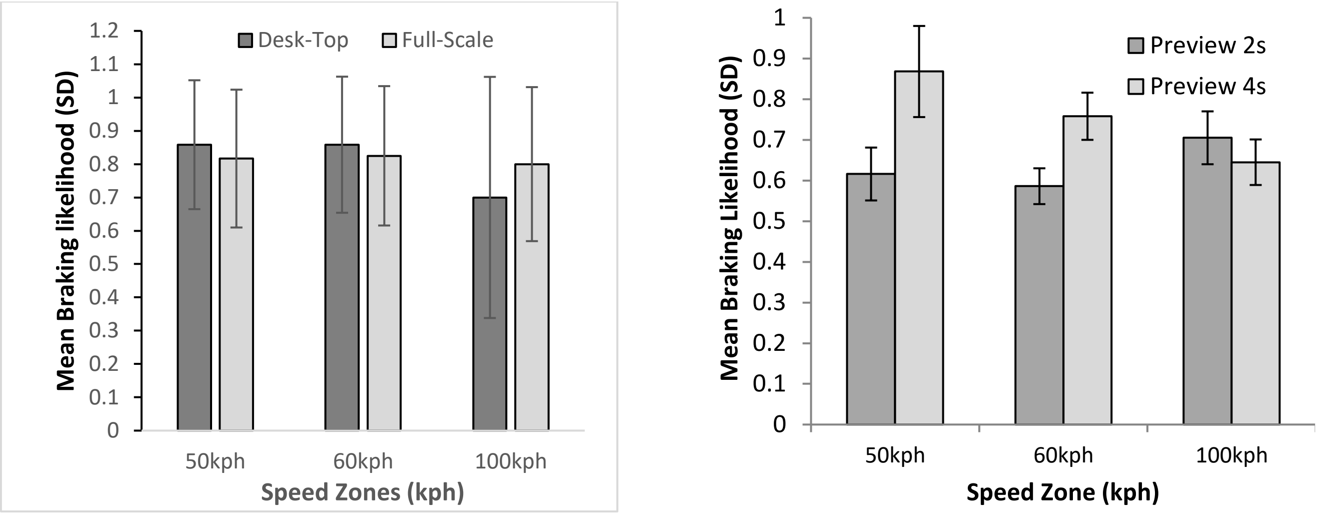


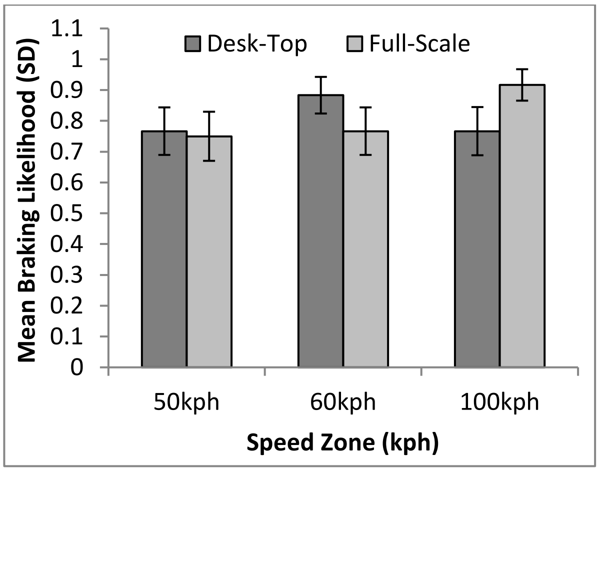
Figure 3 Likelihood of braking with a pedestrian obstruction, effects of (a) simulator type and (b) speed zone

There were statistically significant correlations between the frequency of braking ($r_{(29)}$= 0.48) and steering ($r_{(29)}$= 0.37) decisions in the two simulators, and between the time taken to make these decisions in each case (Brake: $r_{(29)}$= 0.66; Steer $r_{(29)}$= 0.38) . Finally, a composite measure was calculated for each simulator by normalising speed, and in this case the ratio of steering to braking decisions. These two were reliably correlated in the Desk-Top ($r_{(30)}$= .72, p<.01)), but not Full-Scale simulator ($r_{(30)}$= .15), but only decision type was correlated across simulators ($r_{(30)}$= .37, p<.05).

In summary, while the absence of differences in the overall preponderance of braking/steering decisions suggests absolute validity, different effects of speed and pedestrian preview in the two simulators on that decision making, undermines this conclusion. With respect to the time taken to make decisions when pedestrians are encountered, both simulators show similar effects, with both being similarly affected by manipulations of approach speed and the time available for decision making. Thus at least relative validity might be claimed for decision type, and absolute validity with regard to decision time.

Vehicle Obstruction Task: Missing responses were identical in both simulators (4.5%). Response tendencies were identical in the two simulators when the obstacle ahead was a vehicle, with braking (78.7%) or steering (17.8%). It is noteworthy that deceleration, as with the pedestrian task, only occurred when the approach speed was 100Kph. There were no statistically reliable main effects or interactions, other than a difference in braking likelihood which depended on approach speed and simulator ($F_{(2,60)}$= 3.31, p<.05, $\boldsymbol{\eta}_{p}^{2}$= .10, see Figure 4). This small, but statistically reliable effect shows that the obstructing vehicle was responded to differently at higher speeds in the two simulators, with braking less likely in the Desk-Top simulator at 100Kph, individual comparisons did not reach statistical significance in post hoc tests.

Thus, the simulators were similar with respect to the propensity for drivers to brake, rather than steer around an emerging vehicle, but the speed at which the vehicle was approached influenced this somewhat differently in both cases. As with the pedestrian task described above, this suggests mixed evidence with regard to absolute and relative validity.



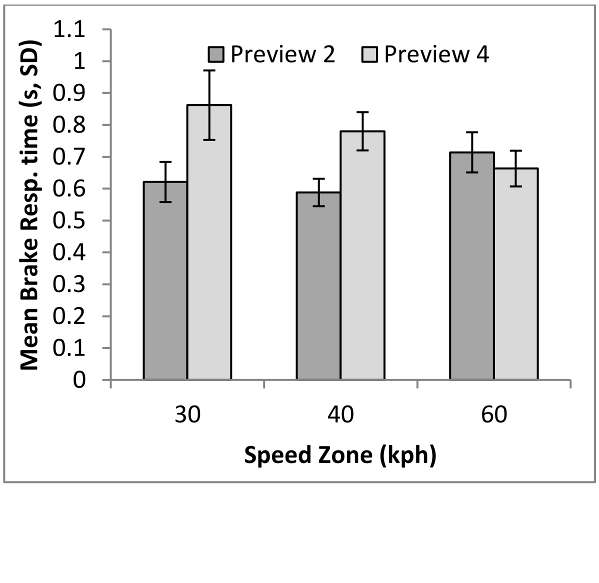


Figure 4 Effects of Preview on (a) brake reaction likelihood, and (b) brake reaction time, in different speed zones

Reaction times to the emerging vehicle were similar in both simulators (Desk-Top: 740ms ±320ms; Full-Scale 671ms ±208ms; $t_{(29)}$= 1.35, p=.19; $r_{(30)}$= .51, p<.005). Braking and steering response times are similar in the Desk-Top simulator (Brake 647 ±232; Steer 878 ±756, $t_{(18)}$= 1.46, p=.16) and in the Full-Scale simulator (Brake 689 ±231; Steer 761 ±661, $t_{(17)}$=.421, p=.68).

A repeated measures ANOVA revealed no significant main efforts or interactions involving simulator type, but responses were faster overall when the time available to respond before hitting the emerging vehicle was short ($F_{(1,29)}$= 7.81, p<.01, η_p^2=.21), but this effect was present only at lower approach speeds ($F_{(2,58)}$= 3.16, p<.05, η_p^2=.10 see Figure 4). An ANOVA based on actual brake response times revealed no significant main effects or interactions. This suggests that with respect to response times, there is good evidence of absolute validity, with differences in approach speed, and time viewing the emerging vehicle affecting response time in both simulators.

As with the Pedestrian task, a composite measure was calculated for each simulator by normalising speed and the ratio of steering to braking decisions. These two were reliably correlated in the Desk-Top ($r_{(30)}$= .71, p<.01) and Full-Scale simulator ($r_{(30)}$= .60, p<.01), but for this scenario decision time was correlated across simulators ($r_{(30)}$= .58, p<.05).

With respect to validity, as with the pedestrian task, there is mixed evidence of relative and absolute validity with respect to decision type, but strong evidence of absolute validity with respect to the time taken to make decisions.

Lane-change Task: Over 90% of all instructions were responded to correctly, 94% and 95% in the Desk-Top and Full-Scale simulators respectively. Instructions which required moving to the drivers right (i.e. road edge towards central reserve) were carried out more accurately than those which required moving towards the left ($F_{(1,29)}$= 25.66, p<.001, η_p^2=.47), and the extent of the movement required also affected accuracy ($F_{(2,58)}$= 9.01, p<.001, η_p^2=.24). These main effects interacted significantly ($F_{(2,58)}$= 9.65, p<.001, η_p^2= .25, see Figure 5), revealing that having to move from the extreme right to the extreme left lane (i.e. what in the UK/Ireland would be road centre to road edge, or outside to inner-most lane), was particularly worse. No other response-type results were statistically significant, indicating that the same decision-making propensities obtained in both simulators

Time taken to detect the instruction and to make the initial steering movement, averaged 841ms (±511) and was almost identical in both simulators, and neither direction or extent of movement significantly affected detection time (all F<1). However, the time taken to complete the instructed action was faster in the Desk-Top (3192 ± 942ms) than in the Full-Scale simulator (3852 ±848ms; $F_{(1,17)}$= 20.71, p<.001, η_p^2=.55), and, in both simulators, was affected by whether the movement required the driver to move to the adjacent lane or to one further away (($F_{(1,34)}$= 111.38, p<.001, η_p^2= .88; 2692ms & 3065ms < 4807ms).

The average lateral distance travelled differed in the two simulators (Full-Scale simulator: 4828 ±2452ms, Desk-Top simulator: 3400 ±942ms; F(1,20)=13.90, p<.001, η_p^2=.41). This was affected by what the goal of the instruction was ($F_{(2,58)}$= 4.38, p<.05, η_p^2=.13). The extent of the movement required influenced the variability within positioning during the manoeuvre ($F_{(2,28)}$=4.31, p<.05, $\boldsymbol{\eta_p^2}$=.24), and there was a marginal difference between simulators ($F_{(1,14)}$=4.36, p=.056), $\boldsymbol{\eta_p^2}$=.24). The main effects interacted significantly $F_{(2,28)}$=4.208, p<.05, $\boldsymbol{\eta_p^2}$=.231; see Figure 5), path efficiency was better in the Full-Scale simulator, and inefficiency increased with the degree of movement required. As with the other manoeuvres, a composite performance measure was calculated to reflect lane changing performance by normalising and then averaging scores for detecting the instruction quickly, the direction of the path chosen (i.e. the average lateral position from start to end position), and as well as the efficiency of the path chosen (i.e. the variability of the previous measure). In each case the inverse was used, such that smaller numbers reflect optimal performance. This measure was correlated across simulators ($r_{(30)}$= .39, p<.05).

It is worth noting in passing that, as with the Brake Reaction task, within an actual avoidance manoeuvre, reaction times in both simulators are very similar. It is also noticeable that response times in the avoidance scenario are faster than to the onset of a brake light on the car ahead- although this may be due to the requirement to respond using the car horn, rather than the more typical braking or steering response.

In summary, the simulators were similar with respect to reaction times and drivers performing better on right-moving lane changes than left-moving lane changes, However there was better path efficiency in the Full-Scale simulator. This suggests mixed evidence with regard to absolute and relative validity.

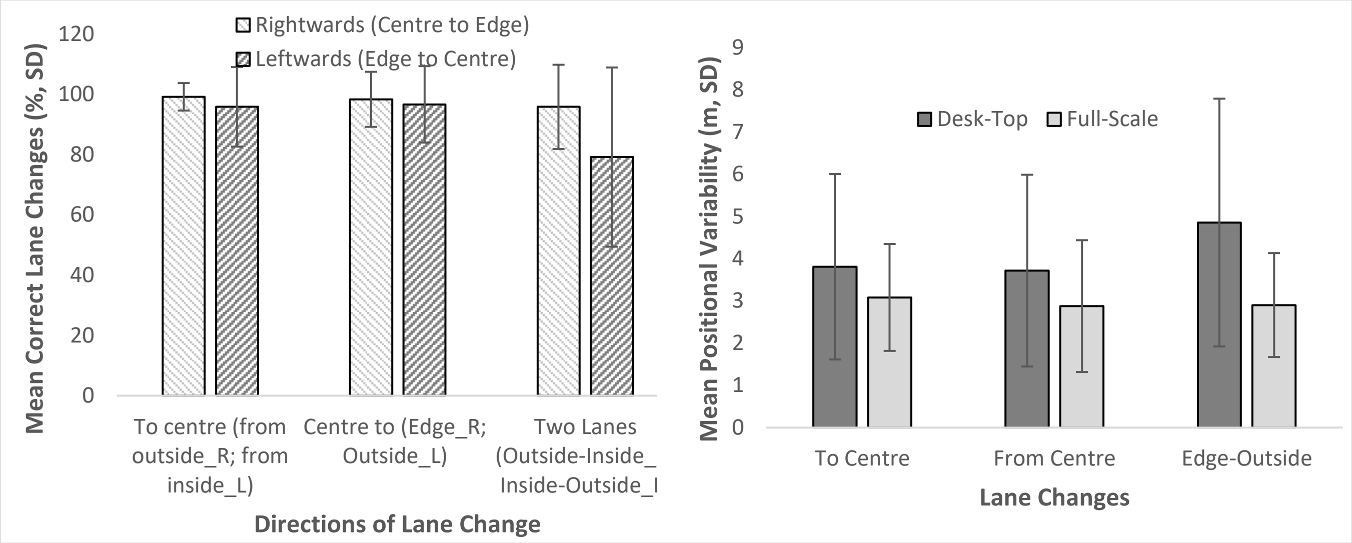


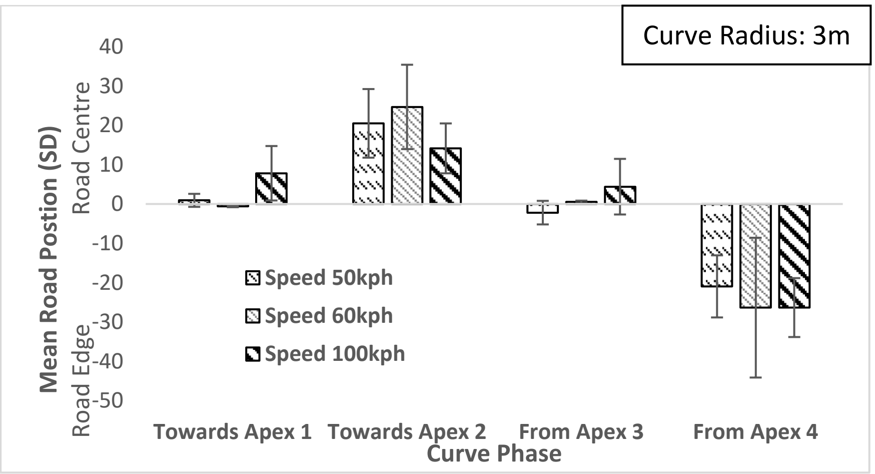
Figure 5 Lane changing as a function of (a) direction and extent of travel, (b) simulator type and extent of travel

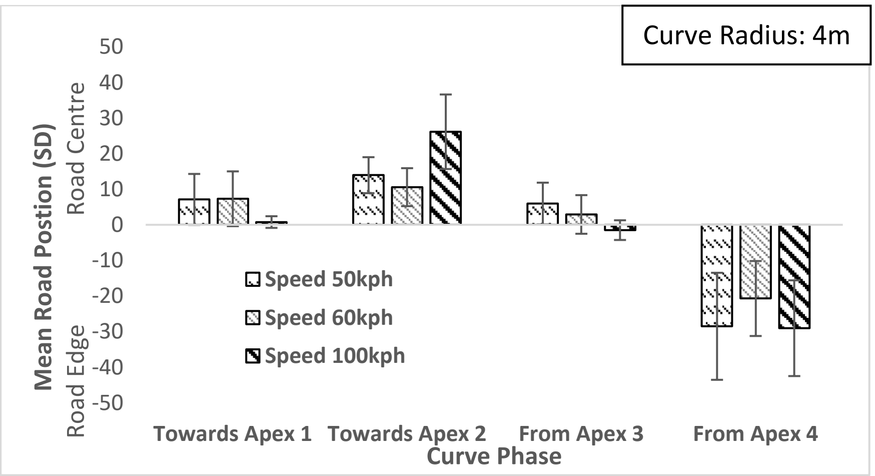
Curve driving Task: Perhaps surprisingly, our curve driving scenarios proved the most complex to analyse. It is to be expected that a driver's position with respect to the road centre and road would change as the curve is traversed, gradually moving away from the road edge, towards the road centre, correcting this, and returning toward the road edge. If the curve radius manipulation affected behaviour, the steering paths would be expected to be different, with more correction needed with 'tighter' curves. Speed would also be expected to change, as would the variation in that speed. Although the simulators differed very little with respect to how curves were driven, once again emphasising the validity of the measurements taken, the effects of our attempts to manipulate curve driving are worthy of closer consideration.

In terms of steering trajectory (i.e. the mean road position over time), only the main effect of curve phase was statistically reliable ($F_{(3,87)}$= 1146.87, p<.001, $\boldsymbol{\eta}_{p}^{2=}$.98), although the difference between simulators approached significance ($F_{(1,29)}$= 3.37, p<>.08, $\boldsymbol{\eta}_{p}^{2=}$.10), with drivers tending to stay closer to the road edge in the Full-Scale simulator. There were two statistically reliable interactions, between curve radius*curve phase and speed zone*curve phase, and these combined into the only other statistically significant result curve radius*curve phase*speed ($F_{(12,\ 348)}$= 8.52, p<.001, eta $\boldsymbol{\eta}_p^{2=}$ .23), see Figure 6. In general, curves were approached with a trajectory which resulted in the vehicle moving towards and beyond the centre of the roadway, a correction took place close to the apex of the curve which resulted in the vehicle heading back towards the driver's nearside verge. Where the curve radius was smaller, i.e. the tightest curve, movement towards the centre is greater, and there is an additional correction, effectively introducing a further 'hinge' into the curve. Speed zone, and hence presumably approach speed, modified this, at higher speeds, road position is closer to the verge, and the trajectory towards the curve apex is a more extreme correction, especially where the curve has a smaller radius. The inability to make this correction, in time or at all, would result in crashes where the vehicle is found on the off-side verge.

Steering movements made were affected by curve radius ($F_{(2,58)}$= 30.97, p<.001, $\boldsymbol{\eta}_p^{2=}$ .52), being less variable on tighter curves (Radius 3m: 8.58±2.29 <> Radius 4m : 8.14±2.59< Radius 7m: 13.13±4.73). Curve phase also affected steering movements ($F_{(3,87)}$= 7.64, p<.001, $\boldsymbol{\eta}_p^{2=}$ .21; Initial: 8.57±5.047, Into Apex: 12.72±3.97, Out of Apex: 9.39±2.85, Final: 9.12±4.26), increasing towards the apex of the curve and reducing again. Steering was more variable in higher speed zones ($F_{(2,58)}$= 40.81, p<.001, $\boldsymbol{\eta}_p^2$=.59; 50Kph: 7.759±2.37, 60Kph: 7.46±2.037, 100Kph: 14.63±5.83). Steering movements did not differ between simulators.

As might be expected, the speed zone in which the curve was placed, and its curvature, affected the average curve speed ($F_{(4,116)}$=27.62, p<.001, $\boldsymbol{\eta}_p^{2=}$.49). This was modified by where on the curve the driver was ($F_{(6,174)}$= 4.06, p<.001, $\boldsymbol{\eta_p}^{2=}$.12). In general, speeds were higher on curve approaches than on the sections before and immediately after the curve apex, especially on tighter curves, and speed zone and curve phase also interacted ($F_{(6,174)}$= 5.15, p<.001, $\boldsymbol{\eta_p}^{2=}$.15), with greater reductions in average speed around the curve apex in higher speed zones. The three-way interaction between road curvature, speed zone and curve quadrant failed to reach statistical significance (p>.3), and neither the main effect nor interactions involving simulator type approached significance (all F <1).





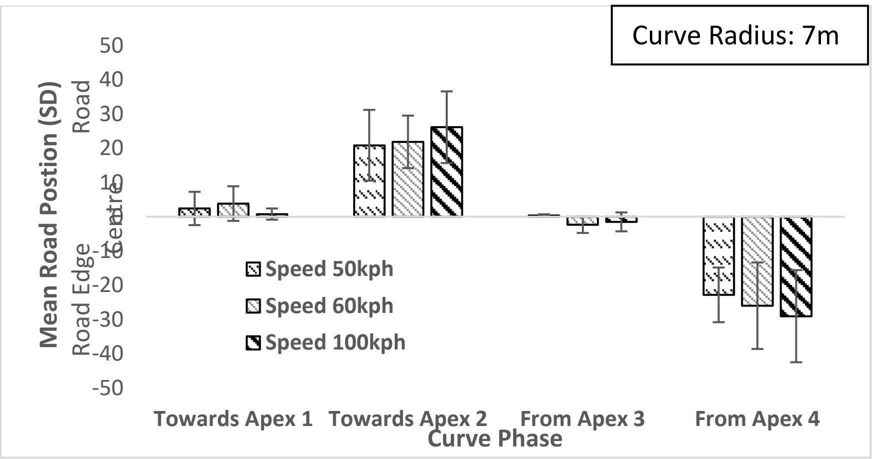


Figure 6abc Curve driving trajectory as a function of curve phase, speed zone and curvature in curves with different radii (3,4, 7m)

Consistent with this, speed variation across the curve depended on curvature ($F_{(2,50)}$= 5.01, p<.01, $\boldsymbol{\eta_p}^{2=}$.17, being higher on the largest curve), speed zone ($F_{(2,50)}$= 2.45, p=.10, $\boldsymbol{\eta_p}^{2=}$.09, with greater variation in the higher speed zone), and curve phase ($F_{(3,75)}$= 10.59, p<.001, $\boldsymbol{\eta_p}^{2=}$.30, with greater variation on curve entry than curve exit). These main effects interacted significantly ($F_{(12,300)}$= 1.88, p<.05, $\boldsymbol{\eta_p}^{2=}$.07), with these general trends being more in evidence where road curvature and speed zones were more permissive. Once again, simulators did not differ overall in terms of speed variation, nor did any interactions involving simulators approach statistical significance (all F<>1, except Simulator*Curve phase ($F_{(3,75)}$= 2.21, p=.09, $\boldsymbol{\eta_p}^{2=}$.08). Finally, as with all other manoeuvres, a single index was calculated to reflect overall driving performance on curves, by normalising performance across simulators for each index, and averaging these across different speed and curvature conditions.

In terms of validity, there is a strong case to be made for the absolute validity of the simulators in terms of curve driving. Both showed clear effects of the difficulty manipulations we intended, but did not differ in the extent of the effects of these manipulations.

Combining On-road and Simulated driving

As mentioned above, composite scores were calculated for each manoeuvre by normalising scores across simulators and then averaging across components. The results of doing so are shown in Table 2, with the diagonal grey area indicating the correlation between simulators and the upper (Full-Scale) and lower (Desk-Top) triangles formed by the diagonal showing correlations within and across manoeuvres separately for the two simulators. There were statistically significant correlations between performance of the same manoeuvres in the different simulators, indicating that those who driver better in one simulator also do so in the other, although for three of the seven manoeuvres these correlations failed to survive FDR correction. None of the scenario-based paired t-tests between Full-Scale and Desktop simulators was statistically significant. Together these indicate both relative and absolute validity at manoeuvre level, consistent with the conclusion reached above, and also substantial convergent validity.

However, as is also clear from the upper and lower triangles in Table 2, there is very little evidence in either simulator of strong correlations within manoeuvres. The only exceptions which survive FDR correction, was an interrelationship between car following, curve driving and lane changing, and driving along a straight road and car following. These three significant correlations are, however, the exception, and just 15% of all correlations in each case. That is, in both simulators, as with the error types recorded during the on-road driving test, orthogonality is the norm. There is certainly no evidence of the 'positive manifold' that might be expected if all of the behaviours measures were part of some single underlying entity.

Table 2 Correlation between and within driving scenarios in Full-Scale and Desk-Top Simulators

		FULL-SCALE SIMULATOR						
		Brake Reaction	Pedestrian Obstruction	Vehicle Obstruction	Curve Driving	Lane Change	Free Driving	Car following
DESK-TOP SIMULATOR	Brake Reaction	.378*	0.275	-0.026	-0.049	-0.11	-0.055	0.112
	Pedestrian Obstruction	0.158	.447*	0.057	0.353	0.046	0.193	0.333
	Vehicle Obstruction	0.03	0.155	.479**	0.072	0.175	0.203	0.059
	Curve Driving	-0.024	0.324	0.11	.941**	0.196	.888**	.465**
	Lane Change	-0.199	-0.29	0.22	-0.07	.467**	0.074	-0.094
	Free Driving	0.006	0.209	-0.117	.790**	-0.236	.612**	.663**
	Car following	-0.149	.368*	-0.052	.769**	-0.144	.471**	.388*

Recall that performance on our criterion measure, on-road driving, was such that eleven of the thirty drivers would have failed their driving test had they committed the Dangerous errors they did on this occasion. Those who would have passed or failed their driving test did not perform significantly differently on any of the simulated scenarios in the Desk-Top (range $t_{(28)}$= .14 to 1.25, all p> .23) or Full-Scale (range $t_{(28)}$= .064 to 1. 85, all p> .12), or on a measure based on performance aggregated across scenarios in either simulator (Desk-Top: $t_{(28)}$= 1.3, p> .20; Full-Scale: $t_{(28)}$= 0.31, p> .75). This overall performance measure was uncorrelated with the numbers of errors committed during the driving test in neither the Desk-Top ($r_{(30)}$= 0.05) or Full-Scale simulator ($r_{(30)}$= 0.11), but strongly correlated between simulators ($r_{(30)}$= 0.63, p<.001).

[This pattern of results is similar irrespective of whether parametric or non-parametric correlations are used.]

DISCUSSION

In essence two questions motivated the research reported above: Do our simulators validly reflect real world driving? To what extent do different aspects of driving relate to each other? As the results imply, the answers to these questions are intimately linked.

Driving simulator behaviour, when the tasks performed are identical and when measured in the same way, appears to be more or less independent of the physical setting in which it takes place. There are no substantial differences in absolute or relative terms between responding to the same driving challenges from a Full-Scale car with surround scenes or whilst seated at a Desk-Top simulator with a similar view of the forward scene. There is some indication that speed, especially at the higher end, is not well simulated in either simulator, but it is somewhat better in larger simulator. These positive indications of validity are consistent with those reported for other simulators (e.g. Godley, Triggs & Fildes, 2002; Mayhew, Simpson, Wood et al, 2011).

However, in our unique approach to validation, we have shown that there is very little relationship between individual manoeuvres, even when these rely upon and measure in the same way: reaction time, lateral position and control, and longitudinal velocity. That is, truly identical elements of driving are related to each other, but similar elements deployed as part of some other manoeuvre, are weakly related, if related at all.

There is thus little evidence for the suggestion that when different tasks rely on similar operations, the performance of these tasks will itself be related- as might be expected from a transfer-appropriate-processing account of training (e.g. Lee, 1988). These results are consistent with recent reviews which are pessimistic about the extent of transfer of training, especially between relatively dissimilar tasks (e.g. Barnett & Ceci, 2002; Sala & Gobet, 2017). Rather than driving reflecting a single underlying competence, as might be implied by a positive manifold, driving may instead be comprised of numerous orthogonal tasks.

Except where intending validators have sought to simulate an actual stretch of roadway in their simulator (see Blana, 1996) relationships between on-road and driving simulator behaviour are at best suggestive, rather than compelling (e.g. Mayhew, Simpson, Wood et al, 2011; de Winter, de Groot, Mulder et al, 2009). Consistent with this, in the study reported above, direct relationships between behaviour in either simulator and the specific types of error committed during an on-road driving test were weak and sparse, and would not necessarily have been predicted a priori because the way in which the behaviours differ. This should, we suggest, give some pause for thought about what a researcher's intentions might be when attempt[ting to validate a simulator.

A distinctive characteristic of the current study is the use of a real-world, on-road assessment test, scored by a qualified and highly experienced assessor, as the outcome measure for the on-road portion of the study. Inevitably, this measure is perhaps less precise and more subjective than the measures used in the driving simulators, but importantly, it reflects the current best practice for driving assessment in most countries and thus best corresponds to our desired assessment of criterion validity (Guion, 1980). Our findings suggest that performance in the driving simulators was not a strong predictor of performance in the on-road assessment. Thus, despite compelling evidence of absolute and relative validity, both across and within simulators, there is no real evidence of criterion validity. It might be argued, because of its inherent subjectivity and measurement properties, that this criterion is in appropriate. Future research might, through vehicle instrumentation when on-road driving is assessed, reduce the challenge this poses for future simulator users and validators. However, as the study above very clearly shows, the situations measured and types of measures used, must closely approximate each other if there is to be any real chance for validity to be observed.

The robust but highly specific relationship between driving simulators, and unspecific, perhaps higher order, relationship between on road driving and simulators, has implications beyond establishing validity. The results imply that there is no single driving competence, nor a set of generic skills which, once acquired, relate to each other across different driving situations. As speculated previously, on the basis of reviewing the paucity of transfer of training in other literatures (Groeger & Banks, 2007), this means that when we learn to drive our ability to perform in a given traffic situation will depend on our history of driving in highly similar circumstances. Without such highly specific practice, and the attendant feedback and instruction from supervising drivers, our capacity to perform remains limited. There is recent evidence which shows that errors committed by drivers undergoing training decrease with practice (see Durbin, Mirman, Curry et al., 2014). Other evidence shows that the rate at which accompanying instructors reduce their feedback and instruction (Groeger & Clegg, 2008) suggests that competence is developed on the basis of the accumulation of experience of specific manoeuvres, rather than more general practice- which has a far weaker effect. Whether ‘manoeuvres’, or less specifically ‘traffic scenarios’, are the basic behavioural units which underlie safe, skilful, driving, cannot be resolved on the basis of the findings reported above. What the results strongly support is that however drivers typically 'parse' their driving experience, these units of experience, and the operations they require, are likely to be orthogonal to each other. As such, they are likely to be learned and maintained through the repeated encountering of highly similar scenarios. The implications of this are that when learning to drive, drivers should gain experience under varying, but largely consistent circumstance, until a degree of competency is achieved- before the challenge of driving is systematically increased.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of Ireland's Road Safety Authority under "Fitness to Drive following Neurological Disorder: Integrating Off-Road, Simulator and On-Road Driver Performance", as well as the technical assistance at various stages of the overall project by Dr Amanda Stephens, Dr Steve Trawley and Dr Chris Burns. Particular thanks are due to Barbara Howe, who collected the data, but regrettably left the project before reporting.

REFERENCES

Allen, R. W., Rosenthall, T.J. and Cook, M.L., 2011. A short history of driving simulation. In D. Fisher, M. Rizzo, J.K. Caird, & J.D. Lee, *Handbook of driving simulation for engineering, medicine, and psychology*, Boca Raton, Fl., CRC Press.

Barnett, S. M. & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128, 612–637.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* 57, 1, 289–300.

Blana, E. (1996). Driving Simulator Validation Studies: A Literature Review. Working Paper. Institute of Transport Studies, University of Leeds, Leeds, UK. *http://eprints.whiterose.ac.uk/2111/*Caird, J.K. & Horrey, W.J. (2011). Twelve Practical and Useful Questions about Driving Simulation, In D. Fisher, M. Rizzo, J.K. Caird, & J.D. Lee, *Handbook of driving simulation for engineering, medicine, and psychology*, Boca Raton, Fl., CRC Press.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56, 81-105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Deary, I. J., Strand, S., Smith, P., Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35, 13–21.

De Winter, J. C. F., De Groot, S., Mulder, M., Wieringa, P. A., Dankelman, J. & Mulder, J. A. (2009). Relationships between driving simulator performance and driving test results. *Ergonomics,* 52, 2, 137-153.

Durbin, D. R., Mirman, J. H., Curry, A. C., Wang, W., Fisher-Thiel, M. C. Schultheis, M., & Winston, F. K. (2014). Driving errors of learner teens: Frequency, nature and their association with practice. *Accident Analysis & Prevention*, 72, 433-439. Fisher, D.L., Rizzo, M., Caird, J.K., & Lee, J.D. Handbook of driving simulation for engineering, medicine, and psychology, Boca Raton, Fl.: CRC Press.

Godley S. T., Triggs T. J., & Fildes B. N. (2002). Driving simulator validation for speed research. *Accident Analysis & Prevention*, 34(5):589-600.

Groeger, J. A. (2000). *Understanding driving: Applying Cognitive Psychology to a complex everyday task.* Hove, UK: Psychology Press.

Groeger, J. A. & Banks, A. P. (2007). Anticipating the content and circumstances of skill transfer: Unrealistic expectations of driver training and graduated licensing? *Ergonomics*, 50, 1250–1263.

Guion, R.M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385-398.Lee, T.D. (1988). Transfer-Appropriate Processing: A Framework for

Conceptualizing Practice Effects in Motor Learning. In Onno G. Meijer & Klaus Roth (Eds.) *Advances in Psychology: Complex Movement Behaviour 'The' Motor-action controversy*. Elsevier B.V; Amsterdam, 201-215.Lee, Y., Lu, M., & Ko, H. (2007). Effect of skill training on working memory capacity. *Learning and Instruction*, 17, 336–344.

Lemieux, C.,& Stinchcombe, A., Gagnon, S. & Bédard, M. (2014). Comparison of simulated driving performance across platforms: From 'low-cost desktop' to 'mid-level' driving simulators. *Advances in Transportation Studies*. 34. 33-42.

Lenné, M. G., Groeger, J. A., & Triggs, T. J. (2011). Contemporary use of simulation in traffic psychology research: Bringing home the Bacon? *Transportation Research, part F: Traffic Psychology and Behaviour*, 14, 6, 431-434.

Mayhew D. R., Simpson H. M., Wood K. M., Lonero L., Clinton K. M., & Johnson A. G. (2011). On-road and simulated driving: concurrent and discriminant validation. *Journal of Safety Research,* 42,4, 267-75.

McWilliams, T., Ward, N., Mehler, B. & Reimer, B. (2018). Assessing Driving Simulator Validity: A Comparison of Multi-Modal Smartphone Interactions across Simulated and Field Environments. *Transportation Research Record: Journal of the Transportation Research Board*, 2672, 37, 164-171.

Mullen, N., Charlton, J., Devlin, A. & Bédard, M. (2011). Simulator Validity: Behaviors Observed on the Simulator and on the Road, In D. Fisher, M. Rizzo, J.K. Caird, & J.D. Lee, *Handbook of driving simulation for engineering, medicine, and psychology*, Boca Raton, Fl., CRC Press.

Perkins, D. N., & Solomon, G. (1988). Teaching for transfer. *Educational Leadership*, 46,1, 22-32.

Road Safety Authority (2009) *Driving Fault: Marking guidelines*. RSA: Ballina, Co. Mayo.(http://www.rsa.ie/Documents/ADI/Information%20Pack/DrivingFault_Marking_Guides.pdf)

Sala, G., & Gobet, F. (2017). Does Far Transfer Exist? Negative Evidence From Chess, Music, and Working Memory Training. *Current Directions in Psychological Science*, 26(6), 515–520.

Salas, E. & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, 52, 471-499.

Schellenberg, E. G., (2006). Long-term positive associations between music lessons and IQ. *Journal of Educational Psychology*, 98, 457–468.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions (I). *Psychological Review*, 8, 247–261.