ASSESSMENT IN DRIVING SIMULATORS: WHERE WE ARE AND WHERE WE GO

Bart Kappé[1], Leo de Penning[1], Maarten Marsman[2], Erik Roelofs[3]
[1]TNO, Soesterberg
[2]RCEC, Arnhem
[3]CITO, Arnhem
The Netherlands
Email: Bart.kappe@tno.nl

Summary: This paper describes the mindset at the start of a three year project to develop a test on a driving simulator. It reviews the literature, presents background information on driver training simulators and their relation with assessment. It then introduces some of the ideas behind this project, the adaptive cognitive model that will be used, as well as the interoperable assessment module we will develop.

INTRODUCTION

Using driving simulators for assessment is nothing new. They have been used in research and rehabilitation for decades. However, in relation to driver training and –testing, their use is relatively unknown territory. That is not without reason: developing and validating such a test will require hundreds of students. Such research can't be performed in a laboratory setting, and has to be performed in a 'field lab' at a driving school. Also, it is unclear how performance measures are related to driving, how driving in a simulator can be assessed best by experienced assessors, and how information provided by these assessors can be combined to come to an automated estimate of a person's driving ability.

In October 2008 TNO has initiated a three year project to develop and evaluate a test on a driving simulator. Participants in this project are the Dutch licensing authority (CBR), the Research Center for Examination and Certification (RCEC), ANWB driving schools and ADS Technique Inc. The project is supporting two PhD's and aims to develop a generic testing module for driving simulators using a hybrid cognitive model (using symbolic as well as neural network techniques) to assess the relation between simulator measures and driving skills assessed by human raters. In a simulator, it seams reasonable to assume that performance is measured objectively and reliably. Most driving simulators feature deterministic, scripted traffic situations. This allows identical, well defined traffic situations to be presented to examinees as items. The test module will present items selected on the basis of item characteristics described in an item response theory model (IRT; Lord & Novick, 2008). Each item will be chosen to provide the most information cantered around a students' current driving ability estimate (computer adaptive testing, see van der Linden & Glas, 2000). The test module will be developed using the driving simulators of ANWB driving schools, as well as their students and assessors (raters).

BACKGROUND AND LITERATURE

Driving simulators seem a valid tool for assessment purposes, as they measure performance in a setting that is highly similar to the actual driving task. This similarity is much higher than the more traditional computer based assessments using pictures, videos and/or animations.

However, the literature on simulator based assessment in relation to driver training is relatively sparse. Kappé, van Emmerik and van Winsum (2004) reported on the validation of a simulator based 'Intest'. Aspirant driving school students have to perform simple tasks on the driving simulators of ANWB driving schools, e.g. driving off, changing gears, lane keeping, braking and stopping. Based on a regression function on the simulator data the system automatically generates an advice concerning a package of driving lessons best suited for the student (a package with the minimal number of hours for the student to be ready for the driving test). This function was developed using the assessments of an intest examiner, who watched the student perform in the simulator and gave an advice of the best suited package. It was found that the prediction made by a linear regression function on the simulator data (training group $N = 160$) was able to match the instructor's judgment with a correlation of 0.99 (control group $N = 800$, regression function and instructor judge the same student, same simulator ride). A similar comparison was made with the judgment of the student's practical driving instructor at the end of the curriculum ('what package should this student have had?'). Here, a correlation of 0.84 was found. These high correlations should be regarded with caution: the simulator instructors could see the prediction generated by the simulator, which may have biased their judgment. Furthermore, out of the six possible packages of lessons, only three were actually advised, and this assures a relatively high base-line correlation. Finally, the judgment of a rater (how much an expert he/she will be) is prone to have measurement error. Using only one rater introduces an additional form of bias due to imprecise measurement. The rater could potentially be a very strict rater, or be wrong in certain situations. Rater bias, in the form of systematic bias (for instance a HALO effect), or non-systematic (wrong judgment) cannot be detected if only one rater is used to assess a student.

De Winter, De Groot, Mulder, Wieringa, Dankelman and Mulder (2008) compared performance in the driving simulator with performance during practical driver training and at the test ($N = 804$). They found a correlation of 0.18 between fewer steering errors in the simulator and a higher chance of passing the driving test the first time. Furthermore, a shorter duration of practical driving training corresponded with faster task execution, fewer violations and steering errors in the simulator (predictive correlation 0.45).

De Winter's research (de Winter, 2009) shows that performance in a driving simulator can have predictive validity on the practical driving test, and on the number of lessons required to be ready for the test. Their relatively low correlations may have been due to the fact that they focused on performance at the control level (e.g. steering, lane keeping, line-crossing), and did not measure performance on the higher levels of the driving task, like traffic participation and hazard perception. Also, the 'grain' of the collected data, which was at the level of a complete 20 minute simulator lesson might have contributed, as it does not allow focusing on errors made in specific situations.

Allen, Park, Cook and Fiorentino (2009) have found that more risky behaviour during driving simulator training is related to higher accident rates in the real world. The highest fidelity simulator (a full cab and wide projection screen) showed the lowest accident rates.

ASSESSMENT IN DRIVING SIMULATORS

There are many factors that may contribute to the validity, reliability and robustness of a simulator based test. Nobody seems to know what characteristics a driving simulator should have in order to be used as a valid system for assessments. As a starting point we therefore suggest the following rule of thumb for simulator based assessment: 'tasks that can be trained well can also be assessed well'. Therefore, we will first have a look at some of the components of a driving simulator and their relation with driver training.

Hardware configuration. We believe that there is a relation between driving simulator configuration and the task envelope it allows to be trained. The motor patterns associated with procedural aspects of the driving task (e.g. changing gears, scanning patterns, starting and stopping) can only be developed correctly in a system with a mock-up that has all the normally available controls, switches and mirrors in their correct spatial position. A simulator with a wide (180 degree horizontally, or more) field of view will allow the correct scanning patterns to be developed, and will improve spatial orientation and self-motion perception. It also provides a good overview at intersections, when changing lanes, merging, etc. The findings of Allen, Park, Cook and Fiorentino (2007), Allen, Park, Cook and Fiorentino (2009), Tarr, Whitmere and Gupta (2007), Kappé (2000), Kappé, van Winsum and van Wolffelaar (2002) confirm this notion.

Figure 1. The complexity and realism of simulated traffic (left) is still far from real traffic (right).

Traffic. Learning how to negotiate traffic situations is an important aspect of driver training. The traffic models used in a driving simulator can not present the full envelope of situations that the driver may be confronted with during practical driving, see Figure 1. The number and diversity of traffic participants is generally limited, just as the type of infrastructure that can be negotiated. Driving simulators differ in the way traffic situations are generated, and not all driving simulators allow scripted situations to be presented to the driver. Having such scripted, deterministic traffic situations is a prerequisite in an assessment of driving ability.

Performance measurement. Assessment of driving performance has traditionally focused on the specific actions a person performs in a car. Recent developments in driving assessment are that more attention is put into the consequences of these actions, instead of the actions themselves. One possible consequence is safety. Current driver behaviour models focus on the underlying cognitive processes (Groeger, 2000) which is not directly observable in a simulator-based test, or on a task (level) hierarchy (Michon, 1985) which does not tell much about the level of driving ability, but more on task complexity. A different approach, that focuses on the consequences of a drivers' actions, is currently being developed (Roelofs, Vissers, van Onna & Nägele, this conference). The idea is to relate performance measures measured in the simulator to the criteria provided by the approach put forth by Roelofs, et al., assessed by multiple raters. However, the set of possible performance measures is quite large (e.g. Östlund, Nilsson, Carsten, Merat, et al., 2004), and the relevant set of measures may vary from situation to situation, just as the standards for adequate performance. Even when everyone seems to have a notion of what safe driving is, automatic assessment of such high order *competences* in a simulator is still not possible.

Training of traffic participation The limitations in simulated traffic are not a handicap when the simulator is used at the initial stages of the training curriculum. Then, drivers need to learn the basic, procedural aspects of negotiating traffic. That is: when to scan, when to slow down, when to give priority and when to take it, etc. For learning such basic traffic procedures a relatively simple traffic environment is fine, and this is exactly what driving instructors do in their practical driving lessons.

Hazard perception and high order skills. The relative simplicity of the traffic situations presented in a driving simulator poses a challenge on training hazard perception skills. Hazard perception is related to the ability to read and recognize potentially hazardous situations in traffic. In simulators, learning how to recognize hazardous situations can only be trained with relatively simple hazards, like when overtaking a bus at the bus stop and another vehicle can just be observed to start crossing the lane in front of the bus. A simulator generally lacks the fidelity to present more intricate hazards.

The driving simulators in the field lab

The simulators at ANWB driving school, see Figure 2, feature a VW Golf mock-up with force feedback steering wheel and all the normal controls, indicators and dials. It has a 3 channel 180 degree field of view (1024 x 768 pixels per channel) using three rectangular projection surfaces. The database allows city, rural and highway traffic to be negotiated. Since 2002, this driving school has 30 driving simulators in operation, training about 10 – 15% of the curriculum (with a focus on the basic skills at the initial stages). The school trains about 5000 students per year. Practical driving instructors of ANWB driving school rated the simulator students 'above average' when compared to normal students at a similar stage in the curriculum (Kappé 2002). For the simulator group, instructors frequently mentioned to have more time available for teaching higher-order competences, and spend less time on the basic aspects of traffic participation. The simulators at ANWB driving schools have proven their training value, and our preliminary assumption is that these systems can be used for assessing vehicle operation and control and the basics of traffic participation and hazard perception.

Figure 2. Two of the driving simulators at ANWB driving schools that will be used in this project

PROJECT OVERVIEW

In the project a simulator based test is developed on a driving simulator. In this test, candidates have to complete test *items* in the simulator: a series of well defined, parameterized, deterministic traffic scenarios. Each item will be selected to lie close to the current estimate of an examinee's driving ability, modelled under an IRT model. Each item represents a relatively small 'traffic assignment', for example: 'merge onto the highway when there's a truck in the lane left of you' or ' take a left turn at the intersection when there's a priority vehicle arriving from the right'. The items are labelled with context information: the assignment (e.g. merge, turn left), the scenario (e.g. a truck in the next lane), the location (e.g. entry ramp, intersection), item specific data (e.g. difficulty, number of times presented to a candidate) and other relevant data (performance measures, time of day). These labels are stored in standardized formats.
We aim to have a set of about 50 test items. Each individual test item will be driven by a large number of candidates. Their performance data is stored in a database that is associated with each individual item, along with additional data on the candidate (e.g. age, sex, number of simulator lessons, number of practical driving lessons, and score on the theory test).

Initially, driving instructors and examiners will rate the performance of candidates on the individual items on a number of criteria (competences). An IRT model is being developed to that maximizes the information provided by these raters and to derive item characteristics and ability of the student. This information is also stored in the database alongside the performance measures of the student driving each chosen item in the simulator. This will not only allow us to set standards on 'low-level' performance measures (e.g. Time Head Way, Time to Line Crossing, Post Encroachment Time), but also allows us to relate these measures to scores on higher order *competences* like 'safe' or 'social' driving behaviour.

Adaptive Cognitive Model

An IRT model is used to model rater judgments and provides a so called 'true' judgement of a person's performance on a test item. When multiple items are taken, and performed by a large

number of students, item characteristics such as item difficulty and discrimination, and an estimate for a persons' driving ability can be derived from the model. An adaptive cognitive model is then used to determine the relation between the contextual information that is contained in the item labels, the performance measures registered in the simulator and the information concerning the judgement, the item characteristics and the students' ability provided by the IRT model, see figure 3.

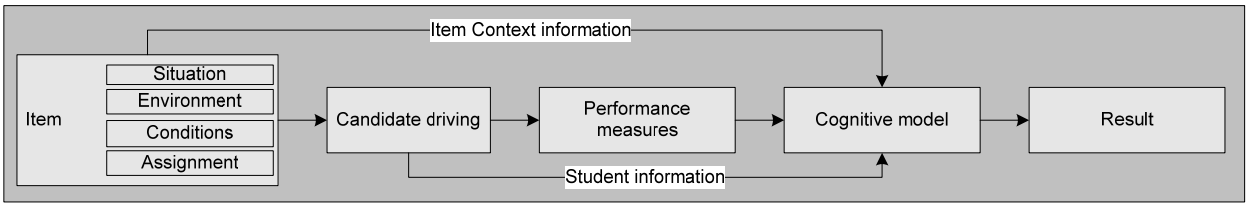

Figure 3. The cognitive model will use performance data, item context information and student information in an automated performance assessment.

The model will use neural-symbolic learning and reasoning. This means that the model can;

- use symbolic information, like existing background knowledge on assessment rules to prime the model,
- learn the relations between the subjective opinions from human assessors on the performance of the candidate and the scored performance measures using neural network learning techniques,
- generate new or improve existing assessment rules based on the learned relations.

We will evaluate the test using up to 30 driving simulators at ANWB driving schools. This will allow us to collect performance data of hundreds of students for each individual item. The test results will be validated using performance data gathered during the practical driver training and with data of the practical driving test.

An interoperable assessment module

An important sub goal of this project is to develop an assessment module that is as generic as possible. As said, we will label and describe each test item on a relatively high level. This context information on the traffic situation is not sufficient for detailed, fully standardized implementation of this meta-information in a script, a logical network, a visual database and a set of performance measures, but it is a start. (Within in the EU project TRAIN-ALL attempts for such standardization are made.)

Whenever possible, we will follow the SCORM and QTI standards for describing item content and context. SCORM is an e-learning standard that allows learning content to be described and used in a standard way. The Question and Test Interoperability (QTI) standard is a related standard aimed at testing and describes item specific data. SCORM and QTI are mature standards, and there's an abundance of SCORM compliant content and tools.

In this project we will use an open source Learning and Content Management System called MOODLE that allows learning content to be stored in standardized databases, along with learner results and instructor opinions. Since it is a system aimed at e-learning, MOODLE is not directly usable in a simulator environment. In the world of simulation the dominant standard is HLA.

HLA is a software framework that allows simulations to be linked and managed. HLA has a focus on interoperability and re-use of simulators and simulation models. Within the HLA community, there is an abundance of related standards and supporting software tools. There is however no standard for didactical aspects within HLA or its related standards.

Recently, the SCORM-Sim study group of SISO (see www.sisostds.org) has worked on an initiative to relate SCORM and HLA. In this light, TNO has developed SimSCORM (de Penning, Boot, Kappé, 2008). SimSCORM is a software platform that allows SCORM compliant content and tools to interact with HLA based simulators that is used in several TNO projects. We will use the SimSCORM platform as the basis for our test module. The SimSCORM platform allows the driving simulator and the test items to interact in real-time. Traffic situations, performance measures and standards for each item will be described in accordance with the SCORM data model, and stored in a SCORM package (i.e. a lesson). When a test item is started in MOODLE it will send this data to the simulator, where the traffic situation is negotiated by the driver. The cognitive model will receive item context information, student information, the performance data of the student and, in its training phase, subjective assessments of the instructor or examiner. The outcome of the assessment, along with the performance data, will be stored in MOODLE.

CONCLUDING REMARKS

This paper describes the ideas, plans and mind set at the start of the project ‘assessment in driving simulators’. It is work in progress, and it has only just started. We aim to have a high degree of transparency in the way we define and parameterize our scenarios and performance measures. This may contribute to an increased standardization of scenarios and performance measures within the driving simulator community. Such a standardization would not only help assessment and testing, but would facilitate research and training as well.

REFERENCES

Allen, R.W., Park, G.D., Cook, M.L. & Fiorentino, D. (2007). The effect of driving simulator fidelity on training effectiveness. Proceedings of the Driving Simulation Conference North America, Iowa City, IA.

Allen, R.W., Park, G.D., Cook, M.L. & Fiorentino, D. (2009). Training and assessment of novice drivers. Proceedings of the Driving Simulator Conference, Monaco, France.

De Penning, L., Boot, E, Kappé, B. (2008). Integrating Training Simulations and e-learning systems: the SimSCORM platform. Proceedings of the IITSEC conference 2008, Orlando Florida.

De Winter, J. C. F., de Groot, S., Mulder, M., Wieringa, P. A., Dankelman, J. and Mulder, J. A.(2008) 'Relationships between driving simulator performance and driving test results',Ergonomics, Oct 28:1- 24.

De Winter, J.C.F. (2009). Advancing simulation-based driver training. Thesis, Delft University.

Groeger, J.A. (2000). *Understanding Driving.* Philadelphia: Taylor & Francis, Inc.

Kappé, B. (2000). Low-cost driving simulation at TNO Human Factors. Proceedings of the Driving Simulator Conference, Paris, France, September 2000.

Kappé, B. (2002). Managementsamenvatting praktijkproef ANWB rijsimulator. Unpublished Memo (in Dutch), TNO, Soesterberg.

Kappé, B. van Winsum, W and van Wolffelaar (2002). A cost-effective driving simulator. Proceedings of the Driving Simulator Conference, Paris, France, September 2002

Kappé, B. van Emmerik, M, van Winsum, W. and Rozendom A. (2003) Virtual instruction in driving simulators Proceedings of the Driving Simulator Conference North America, Dearborn, Michigan, USA, October 2003.

Kappé, B. van Emmerik, M and Winsum, W. van (2004). Validatie Intest. Memo TNO-TM 2004-M013 (in Dutch).

Park, G.D., Allen, R.W., Rosenthal, T.J., Fiorentino, D. (2005). *Proceedings of the Human Factors and Ergonomics Society,* pp. 2201 – 2205.

Lord, F.M., & Novick, M.R. (2008). *Statistical Theories of Mental Test Scores.* USA: Information Age Publishing.

Michon, J.A. (1985). A Critical Review of Driver Behavior Models: What do we know, what should we do? In L. Evans & R.C. Schwing (Eds.), *Human Behavior and Traffic Safety* (pp. 485-520). New York: Plenum.

Östlund, J., Nilsson, L., Carsten, O., Merat, N., Jamson, H., Jamson, S., et al. (2004) *Deliverable 2 - HMI and Safety-Related Driver Performance* (No. GRD1/2000/25361 S12.319626): Human Machine Interface And the Safety of Traffic in Europe (HASTE) Project.

Roelofs, E., Vissers J., van Onna M, & Nägele, R.(2009). Validity of an on-road driver performance assessment within an initial driver training context. *This conference*

Tarr, R., Whitmire II, J.D. and Gupta, K (2007). The virtual check ride as a diagnostic and remediation system. Proceedings of the Driving Simulation Conference North America, Iowa City, IA.

Van der Linden, W.J., & Glas, C.A.W. (2000). *Computerized Adaptive Testing.* Dordrecht:Kluwer Academic Publishers.