Driver performance assessment in driving simulators

Bart Kappé [1], Leo de Penning [1], Maarten Marsman [2]
[1] TNO
Kampweg 5, Soesterberg,
[2] CITO
Nieuwe Oeverstraat 50, Arnhem
bart.kappe@tno.nl, Leo.depenning@tno.nl, maarten.marsman@cito.nl

Abstract *– Assessment of driver performance in practical driver training and – testing faces two challenges. First, there is no control of the traffic situations the driver will be presented with, and second, factors other than the performance of the student may play a role in the assessment. Driving simulators allow scripted, deterministic, traffic scenarios to be presented to the driver, and may use automated performance assessment to ensure objective and reliable assessment. In a three year project, we are developing a standardized, interoperable simulator based driver performance assessment. In a field lab of 30 simulators, we will present deterministic traffic scenarios to large groups of students. Using a cognitive model, we will combine scenario background information and performance measures with the assessments made by human observers. This paper presents the project and its goals, and discusses the different approaches we will use to collect assessment data.*

Introduction

Performance assessment in practical driving

In both driver training and the formal driving test, driving performance is generally assessed during practical driving. Driving instructors and examiners assess performance while the driver is negotiating a variety of traffic situations. As each and every situation is different, performance is always assessed in relation to the traffic situation at hand. The observed performance does not solely depend on the skill levels of the driver, but on the nature of the encountered situations as well, see Figure 1. As one never knows what situations will be encountered, practical driving assessment is inherently fuzzy.

The variability and unpredictability of traffic situations poses some challenges in the assessment of practical driving skills. First, it may hamper the validity and reliability of the assessment. When only relatively simple situations are met, both

skilled and unskilled drivers will tend to pass. When relatively difficult situations happen to occur during the assessment, both skilled and unskilled drivers may fail. When driving congested highways or city centers, it is difficult to generalize the relatively narrow set of assessed driving skills. Thus, the outcome of the assessment depends to some extent to the traffic situations that are met, which is a factor that is not under full control of the instructor or examiner.

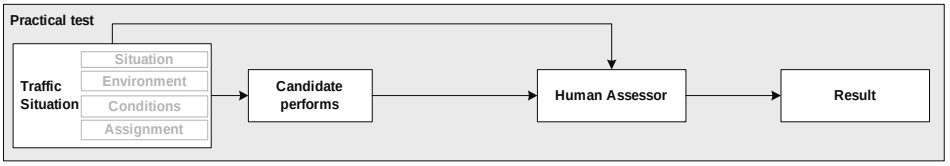

Figure 1. In practical testing, driver performance is assessed in relation to the traffic situation

Second, the variability of traffic situations makes it very difficult to define accurate assessment standards. Assessment manuals currently mention vague standards like brake 'in time' or adjust speed 'appropriately' in respect to 'the traffic scenario at hand', without being able to specify when a braking maneuver should be initiated, or what speed should be maintained. Such vague assessment standards allow room for individual differences in the assessment of driver performance. They also obscure a clear understanding of the variables that define a traffic situation, and their relation with performance measures and standards is vague. In other words, we do not know how 'brake in time' and 'adjust to appropriate speed' vary with the characteristics of the situation.

A third issue in practical driving assessment relates to the human nature of the assessment itself. Assessors can be systematically influenced in their judgment by factors other than the performance of the student. Sex, age and other factors may play a role in the assessment, and it is difficult to get a grip on these factors. Also, similar performance may be judged differently due to severity of judgment.

The variability of traffic, and possible systematic biases may hamper adequate assessment in both driver training and -testing. It will be difficult to meet these issues in a practical driving assessment. We feel they can only be met if one is able to control the traffic situations, and is able to assess performance automatically.

Performance assessment in driving simulators

In a driving simulator, the simulated environment can be deterministic to a large extent. If scripted correctly, a traffic scenario will present a similar traffic situation to the driver, each time it is driven. In our definition, a scenario is a brief 'clip' of a specific traffic situation, such as 'turn left on a signaled intersection with traffic from the left', 'merge onto the highway with a row of trucks on the lane next to you'. In a driving simulator, we may know in advance what traffic situation the driver will be presented during the assessment, and we may allow these situations to be presented in any order.

The traffic situation is not the only aspect that is under control in the simulator. In fact, in the simulator, there is data available on many other aspects that describe a scenario (the 5 Ws: *who* is driving *where*, *what* are they doing *when*, and *why* we should present this scenario).

In the simulator, driving performance can be expressed in many different performance measures (e.g. Pauwelussen, Wilschut & Hoedemaeker (2009), FESTA[1]). And, just like practical driving, we can have an instructor or examiner assess the performance of the driver.

The *difficulty* of a scenario is also a relevant factor. Difficulty levels can be determined subjectively, by having assessors rate the difficulty of a scenario. Difficulty can also be determined statistically, if we are able to present such scenarios to large groups of drivers. Then, scenario difficulty can be based on the actual performance of the students.

By combining scenario descriptors, performance data and human assessments, we may be able to solve some of the above mentioned issues of practical driving assessment in a driving simulator. It could allow us to shed some light on the relevant performance measures and their relation with scenario descriptors. If we include driver and assessor background data (age, sex, experience etc.) we may be able to get grip on the subjective aspects that may play a role in practical driving assessment. We believe that this type of research may ultimately lead to the development of a valid and reliable simulator based assessment.

In 2009, TNO has initiated a three year project to develop a driver performance assessment in driving simulators, in cooperation with CITO (an institute for educational measurement), ANWB driver training (a driving school using simulators) and Rozendom Technologies (a driving simulator manufacturer). The simulator based assessment will be developed and evaluated using the driving simulators of ANWB driver training as our field lab (30 systems, 5000 students/y), see Kappé, de Penning, Marsman & Roelofs (2009) for an introduction.

In the first phase, we have made an inventory of scenario descriptors (for the 5 Ws), of standards to describe content and item data, of performance measures in driving simulators, driving and assessor background data and of cognitive models for assessment in simulators.

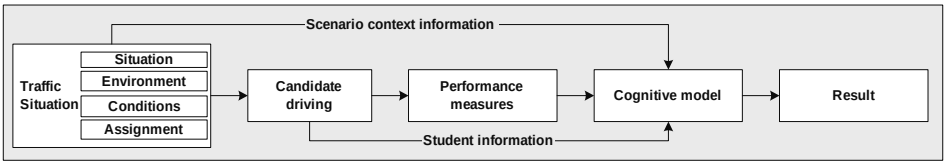

Figure 2. In a driving simulator, the traffic situation that will be presented is known. A cognitive model of an assessor may not only be fed with performance data, but with scenario context and student information as well

[1] See http://www.its.leeds.ac.uk/festa/

We developed a prototype of a Neural Symbolic Cognitive Model that may be used to automatically assess driving performance. The model is able to learn the relations between driver performance, scenario descriptors and the observations of a human assessor, see Figure 2. The model can be fed with both formal and behavioral rules, but is also able to elicit new rules from its data (de Penning, Kappé & Bosch van den (2009), Penning, Kappé & Boot (2009); Kappé, de Penning, Marsman & Kuiper, 2010).

Interoperability through standardization

We realized that a simulator based assessment tend to be developed for simulators of a single manufacturer. As the development of a test is very laborious, we wanted to avoid having to start a new line of research for simulators of a different manufacturer. Thus we try to standardize our scenario data as much as possible. We would like to be able to present identical situations on different simulators, that is, that our simulator based test is *interoperable*. As there is currently no standard scripting language commonly accepted between simulator manufacturers, this can only be done at a meta-level, describing the essentials of a traffic scenario. Therefore, we decided to describe content, results and item specific data in their corresponding standards from the e-learning & e-testing domain (SCORM[2], QTI[3], IMS LIP[4]).

By describing test content on a meta-level, in an e-learning environment that is separated from a specific brand of driving simulator, we hope to take a large step in standardization and interoperability.

TNO has developed the SimSCORM platform (de Penning, Boot & Kappé, 2008). SimSCORM allows SCORM compliant content to be played from (open source) Learning and Content Management systems like MOODLE[5], on any HLA[6] compliant (driving) simulator. (The High Level Architecture (HLA) is the dominant standard for interfacing and connecting simulators). With SimSCORM we can use all the facilities that are offered by modern LCMSs, like databases for storing content, results and student data, and use built in provisions like sequencing and navigation of test content, forums, wiki's etc. As it is web-based, we can access individual simulators from the web, add or manipulate test content, and download performance data and instructor observations. Thus, we can remotely access and control the simulators in our field lab at the driving school.

The SimSCORM platform also serves the cognitive model. The cognitive model has access to the meta-data that we use to describe the traffic scenario, to the performance data of each individual student in that scenario, and to the observations made by human assessors that watch the student negotiate that traffic situation in the simulator. Using SimSCORM's data-logging facilities, we are able to use both live assessment as well as post-hoc performance assessments based on replays of recorded performances in the simulator.

---

[2] http://www.adlnet.gov
[3] http://www.imsglobal.org/question/
[4] http://www.imsproject.org/profiles/lipinfo01.html
[5] http://moodle.org/
[6] http://www.sisostds.org/

Performance assessment methods

This year a prototype of the assessment module, with a database of about 20 testing scenarios will be installed at the driving school. Using this database we aim to collect assessment data in three different ways.

Observer

We will ask instructors to assess a student's driving performance during and after scenario run-time. With these data, we may be able to discriminate 'acceptable' and 'unacceptable' driving performance. We will ask instructors to assess performance at several pre-defined low- and high order aspects of the driving task (guided and unguided by the assessment module). We know that instructors are likely to be influenced by cognitive biases and factors like gender and age of the driver. Direct observation of the driver negotiating traffic situations in the simulator will allow some room for these subjective aspects giving better insight in the influence these factors have in the assessments of human observers.

We realize that during simulator operation, we cannot expect instructors to assess performance at multiple aspects for all students and all scenarios. Therefore the data will logged during simulator operation and can be played back afterwards for assessment when the instructor has more time. This will also allow other instructors to assess the same logged scenario, which improves the validity of the assessment and thus the validity of the cognitive model that learns from these assessments.

Data only

A 'data-only' method does not require human observers. It relies solely on scenario descriptors, performance data, and other readily available data. If we accept that more experienced students will perform better than novice students, we may be able to use their driving experience (e.g. number of driving lessons or –hours) as a rough performance measure for their driving skills.

Using a statistical analysis of the data registered in a simulator curriculum, De Winter (2009) has shown that such an approach is able to discriminate different types of drivers in the simulator and that there is a correlation of these groups with the success at the practical driving test.

Unbiased assessments

We realize that assessors can be systematically influenced in their judgment by factors other than the performance of the student. Also, different assessors can judge similar performance differently due to severity of judgment.

The first aspect, systematic influence by factors other than the performance of the student, is problematic if the factor is a characteristic of the student and there is live assessment. This is because the assessor can see the student, and his or her characteristics, while rating the performance. For instance, when an assessor judges men different than women, because they think that men drive better than

women. The assessor then judges similar performance by a male and a female student differently. If a female student is then judged to perform poorer than a male student, it is not possible to disentangle actual performance from a bias in assessment, and it will consequently be addressed to the student. In our system, the simulator records the performance of a student in the simulator. This recorded performance can be displayed elsewhere on a later moment. This makes it possible to display performance in the simulated environment, without displaying the driver, to an assessor at a different location (preferably in a driving simulator). This replaying of recorded behaviour enables the scoring of the behavior of a student, without bias based on student characteristics.

The second aspect pertains to differences in severity of judgment. This is because different assessors have different internal benchmarks to which they compare performance. To handle this there are two possibilities: First, include assessor effects in the IRT model (see for instance Patz, Junker, Johnson & Mariano, 2002**),** or, second, provide an external benchmark to compare performance to. An external benchmark can be derived by first collecting a small sample of performances of students (say 20). These performances need to be diverse in quality of performance. A group of driver training and examination experts are then asked to individually rank the set of performances on quality of performance. Note that this means that for each task, performance is ranked on a number of sub-domains deemed relevant for competent performance. A statistically optimal ranking of performance can then be provided to a group of experts (possibly the same). The group of experts can then indicate which performance from the ordering can be considered to be on the boundary between sufficient and insufficient. The selected performance can then be used as an external benchmark in scoring performance from a large group of students.

Each of these three assessment methods has its own merits and pitfalls. A data driven approach will be able to use all the performance data that is recorded for training the cognitive model, but will not provide assessment standards. Asking instructors or examiners to rate performance while observing drivers performing the test in the simulator, is relatively simple to realize, be it that they are likely to have cognitive biases in their assessment. Subjective aspects can only be avoided by having instructors perform the unbiased assessment method. This will yield high quality data, but at a cost, as the method is labor intensive. We aim to use all three assessment methods. A comparison of the results may be able to reveal how well a human observer is able to assess true driving performance, and, if present, quantify the nature of their cognitive biases.

Concluding remarks

We believe a simulator based performance assessment may result in more objective assessment of driving performance. By focusing on individual traffic scenarios, deterministic and described in detail, we will be able to take ‘situational’ aspects of driver performance assessment into account. If we are able to get a grip on subjective and individual biases of human assessors, we will be able to train the cognitive model with high quality assessment data. This will open a way for automated performance assessment in driving simulators. We will learn which

performance measures are the most relevant ones, and how these should be standardized. The data generated in our field lab are not only useful for the present research, but they may also be used for the development and refinement of driver- and traffic models.

Bibliography

De Winter, J. (2009) Advancing simulation-based driver training. Doctoral dissertation, Technical University Delft.

Kappé, B., de Penning, L., Marsman, M., Roelofs E. (2009) *Assessment in Driving Simulators: Where we Are and Where we Go*. Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. P 183 – 190.

Kappé, B., de Penning, L, Marsman, M. Kuiper, H. (2010) Human Performance Assessment In Driving Simulators Phase 1: Theoretical Backgrounds. Report TNO Defence and Safety, in Press.

Pauwelussen, J. Wilschut, E.S., Hoedemaeker, M. (2009) HMI validation: objective measures & tools. Report TNO-DV 2009 C062, TNO, Soesterberg, The Netherlands.

Patz, R.J., Junker, B.W., Johnson, M.S., & Mariano, L.T. (2002). The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data. Journal of Educational and Behavioral Statistics, 27(4), p. 341-384.

Penning, de H.L.H., Boot, E. & Kappé, B. (2008) *Integrating Training Simulations and e-Learning Systems: The SimSCORM platform*. In Proceedings of the Conference on Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, USA.

Penning, de H.L.H., Kappé, B., Bosch, van den K. (2009a) A Neural-Symbolic System for Automated Assessment in Training Simulators: Position Paper. In Workshop on Neural-Symbolic Learning and Reasoning of the International Joint Conference on Artificial Intelligence (IJCAI), Pasadena, USA.

Penning, de H.L.H., Kappé, B. & Boot, E.W. (2009b) Automated Performance Assessment and Adaptive Training in Training Simulators with SimSCORM. In Proceedings of the Conference on Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, USA.

