DEVELOPING DRIVING TASK SCENARIOS FOR DEVELOPMENTALLY TAILORED DRIVING ASSESSMENTS: USING AN EVIDENCE-CENTERED DESIGN MODEL

Erik Roelofs[1], Marieke van Onna[1], & Karel Brookhuis[2,3]
[1] Cito, National Institute for Educational Measurement, Arnhem
[2] Delft University of Technology, Delft
[3] University of Groningen, Groningen
The Netherlands
Email: erik.roelofs@cito.nl

Summary: A systematic procedure was described by which task scenarios can be derived as a basis for educationally informative and developmentally tailored driving assessments. To this end, Mislevy´s evidence centered design model for assessment was applied to the driving context. Borrowing from recent theories on driving and driving errors, task environment attributes were derived which may complicate the sub processes of driving and thus may result in varying task difficulty. A universe of assessment tasks was defined by combining basic driving tasks and critical task environment attributes. A collection of 55 critical driving task scenarios was selected from 39 video recorded driving lessons, throughout different stages of driving education. Results of a difficulty rating study pertaining to these scenarios including experienced driving instructors show that the scenarios discriminate well between beginning and advanced learner drivers. Successful scenario solution can be predicted by using an IRT function, where solution probability is a function of driver ability and task difficulty. Implications for assessment design activities are discussed.

INTRODUCTION

During the last decade a shift towards competence-oriented driver training has taken place. Until recently the drivers' task was conceived of as a set of elementary driving tasks pertaining to vehicle control at a low, automated level, and to maneuvering in interaction with other traffic, applying traffic rules. Nowadays, driving is considered as a broad domain of competence, in which the driver is expected to make decisions, taking into account the task environment, combining his interests and those of other traffic participants. Higher order aspects of driving are considered crucial in this process: risk tolerance, reflection on one's own driving behavior and hazard perception. The increased emphasis on integration of lower and higher order aspects of driving is reflected in goals for driver education (the GDE-matrix; Hatakka et al., 2002). Moreover, competence frameworks are being elaborated which specify driver roles and the accompanying knowledge and skills needed for driving. Initial driver training programs increasingly build on these frameworks. As a result, driver training has evolved into a hierarchical learning sequence, particularly stressing the mental processes involved in driving. Driver training is provided by professional driving instructors, who supply tasks in daily traffic. In some European countries a two-phased driver training program, including a pre-license and post-license phase, is put into practice. In addition, initiatives for permanent (ongoing) road safety education have emerged. In sum, acquisition and maintenance of driving competence can be considered an ongoing or even a life-long process.

An indispensible part of driver training and learning is the use of formative assessments, which are used to inform and support the (learner) driver about the levels of driving competence, and the underlying performance aspects that need further attention. Using this information subsequent training may be tailored to the drivers' specific needs (Stiggins, 2002). However, to facilitate higher order level learning, formative assessments should go beyond isolated testing of knowledge about traffic rules and technical driving skills. Fitting in with a competence oriented approach, formative assessments need to: a) address critical elements of the driving task and its circumstances as they appear in daily driving and throughout the driving career; b) fit in with the stage of development of the (learner) driver, c) be informative about mental *processes*.

To tailor formative assessments to different stages of driver competence it is vital to identify driving assessment tasks that discriminate between drivers from different stages. The aim of this study was to develop a systematic procedure by which to arrive at such task scenarios.

Evidence centered design model for assessment

The evidence centered design model for assessments as described by Mislevy, Steinberg, and Almond (2002) seems suited to arrive at a collection of assessment tasks for educational purposes. This model for assessment design helps to sort out the relationships among attributes of a candidate's competence, observations which prove competence and situations which elicit relevant driver performance. The design model is composed of six interrelated sub-models: the Student Model, the Evidence Model, the Task Model, the Assembly Model, the Presentation Model, and the Delivery Model. The first three models mentioned generally constitute the basis for assessment design activities.

The Student Model contains variables representing the aspects of (driving) proficiency that are the targets of inference in the assessment and their inter relationships. The Evidence Model describes how to extract the key items of evidence from driver behavior, and describes the relationship of these observable variables to student-model variables. The Task Model describes the features of a task that needs to be specified when an assessment task is created. In the context of driving a task can relate to e.g. a specified route to be driven, a simulation scenario, or a task presented on a computer screen. In the current study, the design model was elaborated until the level of the task model, which provides the basis for the collection of task scenarios. The following research questions were addressed:

1. Is it possible to identify a collection of driving tasks of varying complexity and difficulty levels using the evidence centered design model?
2. To which extent do driver instructors reliably rate the levels of task difficulty related to these tasks, when they are applied to learner drivers in different training stages?

METHOD

In this study, task scenarios with different task demands were collected while the task difficulty levels were estimated by driving instructors. The categorization of the tasks and the rating variable were based on an elaboration of the evidence centered design model applied to the assessment of driving.

Elaboration of a task model for driving assessments

Applied to driving assessment, the *Student Model* describes the processes that take place during driving and the criteria for competent performance. Competent driving is defined as the ability to carry out driving tasks as they exist within a universe of traffic situations, varying in complexity. The quality of the mental processes must be expressed in terms of performance measures if we wish to make inferences about driving competence. In the current study measures are used that are related to the following outcomes of driving (cf. Roelofs, Van Onna & Vissers, 2010), i.e. 1) safety: the driver's awareness of the situation, resulting in correct timing of actions, adapting speed and using "space cushions"; 2) facilitating traffic flow, which implies not impeding the progress of other road users, and using the road efficiently; 3) consideration with other road users, which means giving others opportunities to fulfill their tasks, or adapting to their mistakes; 4) controlled driving, referring to steering and controlling the car smoothly, without stutters and jerks or departures of smooth lines. The four measures may correlate substantially.

As stated in modern models of driving, the driver is confronted with different levels of task demands which can be close, below or above the ability to solve the concomitant problems (Fuller, 2005). The odds of being in control of the traffic situation are dependent on the balance between task demands and driver abilities. In cases of very high task demands even experienced drivers may encounter difficulties in solving traffic situations. In the *Evidence Model* of this study, each performance measure in the student model is seen as an ability that can have a broad range of values. Item response theory (IRT) models explicitly balance the (driver) ability and task difficulty in predicting the odds of a successful solution of the task by the driver. The evidence model in this study relies on IRT. For each assessment item, an item response function gives the probability that a person with a given ability level (expressed by the parameter θ), will respond correctly on a task. Drivers with low ability have a low probability while drivers with high ability are very likely to respond correctly. Each task may have a difficulty level β. A basic IRT model is the Rasch model (Rasch, 1960). The probability of a successful solution of task j (X_j=1), is a function of the difference between ability θ and task difficulty β_j:

$$P(X_j = 1 | \theta) = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)} \quad (1)$$

IRT offers the possibility to select test items that are tuned to the level of ability of the learner. In the current project we aimed to identify a collection of driving tasks that differ in terms of task demands. However, IRT analyses are useless without sound item construction, based on elaborated ideas of task difficulty. Without those ideas, the range of task difficulties may be too narrow, or the average task difficulty may be off target. The Task Model specifies these ideas on task difficulty.

Critical task environmental attributes that hamper or facilitate the driving process were derived from error-taxonomy studies (Reason, 1990; Stanton & Salmon, 2009). The attributes were categorized with respect to the driving process they hamper. In this way the features of the task environment that may influence the task difficulty were elaborated (see Table 1). The task model was further specified using a distinction in tasks as used in the Dutch driving curriculum resulting in the following basic driving tasks: turning, merging, longitudinal driving, stopping,

crossing, passing, lane changing. In addition, driving task environment were elaborated into detail, referring to e.g. road conditions, weather conditions, traffic intensity, other road users.

Table 1. Critical attributes of the driving task environment that may hinder driving task processes

	Process being complicated		
Critical attributes of the driving task environment	Perception	Decision making	Action execution
Sight obstruction	x		
Hearing obstruction	x		
Discontinuous traffic environment	x	x	
Other participants arrive at scene at the same time		x	
Reduced space to carry out actions		x	x
Inferior road conditions			x
Weather conditions hindering lateral vehicle control			x
Road characteristics hindering lateral vehicle control (hills, curves)			x

Collection of driving task scenarios

Driving lessons from 13 instructors enrolling 39 learner drivers were recorded, using three digital video cameras: one was directed at the driver, one at the road ahead, one at the road behind. All instructors used the training method of Driver Training Stepwise (DTS; Nägele, & Vissers, 2003) which consists of four consecutive learning stages: (1) vehicle control (2) driving in simple traffic situations; (3) driving in complex traffic situations (4) independent driving. To ensure a variety in task complexity, lessons throughout all stages were filmed.

The mixed footage (front shield, drivers' face and rear window) was analyzed for the occurrence of situations in which the learner driver committed an error or a near error, as indicated by verbal comments or pedal interference. The analysis resulted in a collection of 55 scenarios. Each of them represented one specific driving task and involved the complication of one or more sub processes of driving. The greater part of the scenarios related to turning and cruising (both 15 scenarios, 27.3%). A smaller number of scenarios pertained to merging (8, 14.5%), crossing (5; 9.1%), passing (8; 14.5%) and lane changing (4; 7.3%). Out of the 55 scenarios 15 (27.3%) referred to a complicated perception process, 31 (56.4%) to a complicated decision process, and 9 (16.4%) to a complicated action execution.

Subjects

Eight experienced driving instructors (mean age 45.7 years (SD=7.0) and 15 years of experience on average (SD=7.0)) participated. None of them had taken part in the recorded lessons. However, the instructors themselves used the DTS method during their regular training.

The rating procedure

All scenarios were rated in terms of complexity using the following procedure. During two three-hour group sessions driver instructors were asked to view the video scenes which were projected on a screen (60 by 80 inches). The drivers' intention was mentioned by the session moderator;

e.g. “the driver intends to pass a row of parked cars”. After that, the actual execution of the task was displayed on the screen twice. The instructors were asked to respond to the following question: ‘In how many cases out of 10 will the learner driver at the end of DTS stage 2 and 4 respectively solve this traffic task safely, efficiently and independently?’ Safe and efficient were explained as: the learner driver or other road users do *not* need to reduce speed strongly, wait for long times, change their course or evade, touch other road users or objects or to cross road lines. ‘Independently’ was explained as referring to a situation in which no verbal or operational instructor intervention is needed to have the learner driver fulfill the task.

Method of data analysis

Each instructor rating was expressed in terms of a probability, e.g. 0.70 (7 out of 10). The rater agreement was computed by calculating the inter rater reliability coefficient and the Gower similarity index. In case of sufficient agreement the ratings would be pooled. Using the Rasch IRT model, the instructors’ ratings can be conceived as the success probability of the learner driver with ability θ on an item with difficulty β, as stated in a rewritten variant of Equation 1:

$$\ln\left[\frac{P(X_j = 1 \mid \theta)}{1 - P(X_j = 1 \mid \theta)}\right] = \theta - \beta_j \qquad (2)$$

This means that a logit transformation on the average instructor probabilities allows for linear modeling of abilities of drivers and item difficulties. In this study, only two levels of driver ability, corresponding to DTS stages 2 and 4, were considered. In addition, 55 item difficulties were involved. These parameters were estimated by analysis of variance with only main effects.

RESULTS

Since some of the instructors either missed the first or the second session, the dataset was split into three overlapping subsets that were analyzed separately. Version 1 contained the ratings of five instructors who rated all scenarios, version 2 the ratings of seven instructors who rated the first 28 scenarios, version 3 the ratings of six instructors who rated the last 28 scenarios.

The inter-rater reliability of ratings for learner drivers at DTS 2 was good for all data subsets (r>= 0.82). The inter-rater reliability of the ratings for learner drivers at DTS4 was sufficient for all data subsets (r>=0.69). The Gower indexes were all larger than 0.85. These levels of agreement indicate that the estimates of the different driving instructors could be pooled for further analysis with the Rasch model. All individual ratings for a given task scenario were averaged across instructors to obtain pooled estimates of the probability of a successful solution of task *j*, $P(X_j=1|\theta)$. These pooled probabilities averaged about 0.19 for DTS stage 2, and about 0.74 DTS stage 4. They are indicated by the arrows in Figure 1.

A restricted ANOVA with only main effects was run on the logit transformations of the pooled estimates of $P(X_j=1|\theta)$. Error in the logits accounted for only 5.5% of the total variance. The larger part of the variance in the logits was accounted for by the DTS stages (80.7%); the levels of scenario difficulty accounted for 13.7% of the total variance. Note that the scenarios had a

relatively low variation in difficulty levels. Only two scenarios were significantly easier, and three were significantly more difficult than the average item difficulty.

In Figure 1b, the predicted values of $P(X_j=1|\theta)$ given the estimated values of β_j, are plotted for four of the tasks j. Task scenario 34 was relatively easy, resulting in relatively large probabilities of a successful solution. Scenarios 13 and 40 were of average difficulty and task scenario 55 was relatively difficult. The predicted values of $P(X_j=1|\theta)$ at the ability levels corresponding to DTS stages 2 and 4, differed somewhat from the observed pooled estimates. Of the predicted probabilities 19% diverged more than 0.1, but none of them diverged more than 0.2. This seems to indicate a reasonable fit of the Rasch model.

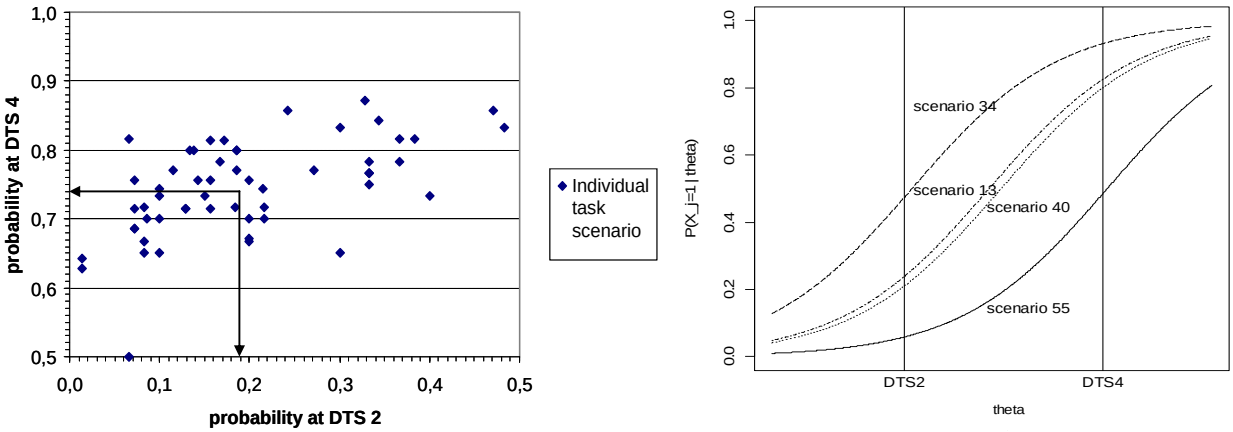

Figure 1. Rated and predicted success probabilities for driving task scenarios in two DTS learning stages; (a) Mean observed success probabilities for individual task scenarios pooled across 8 instructors and (b) Predicted probabilities of a successful solution for four task scenarios as a function of ability

An attempt was made to relate the difficulty levels of the items to the traffic scenario characteristics (traffic tasks and complicating sources). This, however, did not result in a clear difficulty ordering of either the traffic tasks or the complicating sources.

DISCUSSION

This study aimed to develop a procedure by which a collection of driving assessment tasks could be developed. This study illustrates the usefulness of the evidence centered design model. Borrowing from Fuller's theory driving is seen as the dynamic interaction between the determinants of task demands and driver capability, involving various sub-processes. Using insights from error studies the task environment attributes were derived which may complicate these sub-processes. By combining basic driving tasks with critical environment attributes the universe of assessment tasks was defined.

A collection of informative driving assessment task scenarios was selected out of videotaped driving lessons throughout different stages of driving acquisition. Experienced driving instructors viewed the video scenarios and estimated solution probabilities for each of them, taking into consideration beginning and advanced learner drivers. The estimated solution probabilities differed meaningfully between the two envisioned target groups DTS stage 2 and DTS stage 4. However, most of the evaluated driving tasks were considered difficult for

beginning drivers. Regarding the intended formative purposes, these tasks would not fit to the needs of this target group. The high average difficulty level suggests that easier tasks should have been identified.

Encouraging were the high levels of inter-rater reliability among driver instructors, suggesting that the procedure of estimating task difficulty is feasible in practice. The Rasch model could be used as part of the evidence model for assessment. The data seemed to support a reasonable fit of this model. The desirable measurement properties of the Rasch model include the fact that the number-correct test score is a sufficient statistic for the ability level. Also, the Rasch model offers the possibility to model adaptive testing to meet a variety of (learner) drivers.

Within the limited scope of this study we could not sort out the effects of the supposed complicating factors within the task scenario on the estimated task difficulty. Maybe this is because each complicating source can vary in the degree of complication it causes. This limits the possibility to uniquely order the different kinds of sources with respect to the degree of complication they cause.

For reasons of further validation of the design model, the next step in our design study pertains to the construction of actual assessment tasks. In a follow-up study, simulator assessment scenarios will be constructed based on the collected task scenarios, enabling a comparison between actual task performance and expert estimates.

REFERENCES

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis and Prevention, 37*, 461-472.

Hatakka, M., Keskinen, E., Baughan, C., Goldenbeld, Ch., Gregersen, N.P., Groot, H., Siegrist, S. Willmes-Lenz, G. and Winkelbauer, M. (2003) *Basic driver training: New models.* Turku, University of Turku.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice.* Mahwah, NJ: Erlbaum.

Nägele, R.C & Vissers, J.A.M.M. (2003). *'Driver Training Stepwise (DTS). Evaluation of the follow-up in the province of Gelderland.* Veenendaal, Traffic Test.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Roelofs, E.C., Onna, M. van, Vissers, J. (2010). Development of the Driver Performance Assessment: Informing Learner Drivers of their Driving Progress. In L. Dorn (Ed.) *Driver behavior and training, volume IV* (pp. 37-50). Hampshire: Ashgate Publishing Limited.

Reason, J., 1990. *Human Error.* Cambridge University Press, Cambridge.

Stanton, N.A., & Salmon, P.M. (2009).Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science, 47*, 227–237

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan, 83* (10), 758-765.