A COMPETENCE BASED EXAM FOR PROSPECTIVE DRIVING INSTRUCTORS: CONSTRUCTION, VALIDATION AND IMPLICATIONS

Erik Roelofs[1], Maria Bolsinova[1], Angela Verschoor[1], Jan Vissers[2]
[1] Cito, National Institute for Educational Measurement, the Netherlands
[2] Royal Haskoning DHV, Amersfoort, the Netherlands
Email: Erik.Roelofs@cito.nl

Summary: In line with changed views on driver training and driver instructor preparation a competence-based instructor exam was introduced in the Netherlands. The exam consists of two parts: 1) multimedia theory tests, 2) a performance assessment. An implicit idea behind the innovated exam is that it can have a positive backwash effect on the quality of driver instructor preparation programs. This study aims to evaluate the reliability, validity and fairness of the theoretical tests, which appear in different versions across time. Data of 4741 prospective driving instructors, enrolled during the period between January 1010 and October 2012, were used for analysis. The results of IRT-analyses show that the theory tests yielded reliable and fair decisions, although misclassifications occurred across versions. The predictive validity of the theory tests for the final performance assessment was low. Implications for the design and maintenance of exam programs are discussed. Follow-up studies will focus on the question, whether the improved instructor exam contribute to safer drivers in the end.

INTRODUCTION

A growing body of evidence supports the notion that driver training should place greater emphasis on higher-order, cognitive and motivational functions underlying driving behavior. Higher-order skill training addresses factors like driving anger, sensation seeking, and boredom and appears to counteract overconfidence (Isler, Starkey, & Shepard, 2011; Beanland, Goode, Salmon, & Lenné, 2013). In Europe this changed view on driver training was reflected in the Goals for Driver Education issued in 2002 (Hatakka et al., 2002).
Parallel to the changed conception of driver training, the need for innovated driving instructor preparation programs is emphasized. In a review study Bartl, Gregeresen, & Sanders (2005) found that most European programs do not cover higher order skills and rely on teacher-focused approaches, which typically fall short in developing higher order skills.
In line with the changed view, a new law on driving education was introduced in the Netherlands in 2009, including competence-based outcome standards and an innovated exam for prospective driving instructors. An implicit idea behind this was that a valid and reliable exam that only allows proficient prospective instructors to enter the profession can have a positive backwash effect on driver instructor education programs (Crooks, 1988).
These developments resulted in a two stage competence-based exam, introduced in 2009. Since then, over 6000 prospective driving instructors (PDI's) have enrolled in the exam. The question is whether the new exam yields valid and fair decisions about PDI's. This paper focuses on the psychometric quality of the separate theory tests, comprising stage 1 and of the final performance assessment lesson, comprising stage 2. In addition, the fairness question focuses on the comparability of different versions of theory tests used during a period of two years. In summary, four research questions are addressed:

1. To what extent are the individual parts of the exam psychometrically reliable?
2. Do the used cut-off score across different versions of the theoretical tests reflect equivalent required levels of proficiency?
3. To what extent are results on theoretical tests and on the performance assessment for instructional ability related?

DESIGN OF THE COMPETENCE BASED EXAM

The competence based driving instructor exam consists of two parts. The first part relates to theoretical knowledge regarding driving and driving pedagogy. After having passed the first part, PDI's receive a provisional instructor license enabling them to enroll in a half year internship at a (certified) professional driving school. In the second part, after having finished their internships, PDI's are judged on their professional instructional abilities during a masterpiece lesson involving one of their own learner drivers. After they have passed, they get a full instructor license for the next five years. Below the design of the exam is described in more detail.

Instructor competence model

Starting from a literature search on what comprises good teaching in general and more specifically on good driving instruction, a competence model was constructed and validated with the aid of stakeholders. This resulted in the formulation of four domains of competence (see Table 1): 1) Conscious traffic participation as first and second driver, 2) Lesson preparation, 3) Instruction and coaching, 4) Evaluation, reflection and revision.

Table 1: domains of instructor competence

1. Proficient and conscious driving
1.1 Driving responsibly as a first driver: The prospective driving instructor (PDI) is able to drive a vehicle safely, smoothly, socially considerate, and in an eco-friendly way according to Dutch driving standards and traffic rules.
1.2 Verbalizing mental processes of driving: the PDI is able to verbalize the mental task processes that take place when carrying out driving tasks in different traffic situations.
2. Lesson preparation
2.1 Adaptive planning; the PDI is able to construct an educational program for the long term (curriculum) and for the short term (lesson design) adapted to the needs of the individual learner driver (LD).
2.2 Elaborating driving pedagogy: the PDI is able to prepare a learning environment for learner drivers.
2.3 Organizing learning: the PDI is able to organize lessons in such a way that activities run smoothly and without interruptions, ensuring a maximum amount of productive learning time.
3. Instruction and coaching
3.1 Providing instruction: the PDI is able to provide instruction that is geared to the actual developmental level of the learner driver. It enables the LD to progress towards self-regulated performance in increasingly complex tasks.
3.2 Providing coaching: the PDI is able to monitor learner driver development and guide the LD towards self-regulation in solving driving tasks and driving related tasks.
4. Evaluation, reflection and revision
4.1 Assessing learner progress: The PDI is able to assess the progress in driver competence by judging the level of performance himself and by using expertise of professional colleagues.
4.2 Reflection and revision: the PDI is able to reflect on his own actions and to use the results for adapting his approach.

The exam composition

To measure the listed aspects of competence, three theory tests and two Performance assessments were employed.

Proficient and conscious driving: knowledge and performance. To test the PDI's own driving proficiency and his ability to verbalize mental task processes, the PDI had to complete a 60 minute drive, which was judged by a trained assessor. Five performance criteria were used for this purpose: 1) driving safely, 2) aiding traffic flow, 3) driving socially considerately, 4) driving eco-friendly, and 5) controlling the car. At two intermediate stops the PDI was asked to retrospectively verbalize his mental processes that he had went through while solving the traffic situations. The quality of his verbalization was judged on the identification of four psycho motor processes: 1) perception of the key factors in the traffic situation, 2) anticipation of consequences given an intended line of action, 3) decisions to make a maneuver, 4) the way in which the maneuver had been carried out.
To measure knowledge of the theory of driving, an item bank of over 300 items was developed to form the basis of 60-item computer based test versions of the Theory of Driving Test. The items addressed the four psycho motor processes mentioned above. Each item posed a perception question (or anticipation, decision making, action execution question) about a traffic scenario. Key factors in a situation could for instance be traffic signs, an applicable traffic rule, or other road users in a certain position. The situations were always presented from the perspective of the wind shield, i.e. the driver's seat. The Theory of Driving Test items were scored dichotomously (0,1) for incorrect and correct answers respectively. The cut-off score for passing the test was 42 items correct.

Knowledge of driving pedagogy. To measure pedagogical knowledge two items banks with 300 innovative multiple choice items each were developed, for the use in two computer-based tests: Lesson Preparation (60 items) and Instruction and Coaching (60 items). Two types of items were used. First, case based items, which address knowledge of concepts and cause-effect rules, embedded in a rich driving instruction context. An example of such an item is: "An instructor starts a lesson with an explanation on a first lesson topic. He does not explain what the learner driver is able to do at the end of the lesson. What is the most likely consequence for the learner?" To respond, the PDI could choose one out of four options: A. the learner will learn less that desirable, B. the learner will not fully understand your explanation, C. the learner will have less time to practice new driving tasks, D. the learner cannot direct his attention to the essential parts of the lesson (correct). A second item type related to situational judgment items. These items address decision making skills. An example regarding lesson planning is: "In this lesson you are going to instruct the learner driver how to park backwards into a parking bay. Which of the parking bays can you choose best for this learner driver?" To respond, the PDI could choose one option out of four pictures, that represented parking situations with a different complexity. Both the Theory of Lesson Preparation Test and the Theory of Instruction and Coaching were scored dichotomously. The cut-off score for passing these two tests was 38 items.
Performance assessment tasks. The quality of instruction, coaching and evaluation was judged during a lesson with a real learner driver. To this end trained assessors used a 34 item scoring form. The form addressed four aspects of instruction: providing overview, explaining and modeling, guiding practice, providing feedback. In addition, five aspects of coaching were covered: observing and diagnosing driving performance, providing task support, adapting guidance to individual students, interpersonal communication, and providing motivational support. The items on the Performance Assessment Lesson were scored though a three point rubric, representing 'counterproductive performance', 'beginning productive performance' and 'optimal performance'. A scoring guide was available for assessors. Initial rater agreement index

scores using Gower's similarity index (Gower, 1971) showed acceptable levels of agreement, .67 for instruction and .75 for coaching. The cut-off score for passing the performance assessment was 71 points (out of 102 points).

METHOD

Subjects and Data

Test data from 4741 prospective driving instructors who enrolled the program between January 1st 2010 and October 1st 2012 were selected. 79 per cent of them were male and 21 per cent female. The mean age was 34.9 years (SD=10.9). Of them 3079 (74.4%) were born in the Netherlands. The remaining 25.6% originally came from 79 different countries. The majority of them were immigrants from Morocco (n=199), Suriname (n=190), Turkey (n=151), Afghanistan (n=112), Iraq (n=89), and Iran (46). A total of 4644 PDI's completed at least one of the theory tests. Of them 2977 passed all their theory tests, from which 1941 PDI's took part in the Performance Assessment Lesson. From the remaining PDI's about half (n=508) did not participate in the Performance Assessment Lesson within more than a year after their last successful theory test. The remaining part (n=528) did not finish their internship. 368 PDI's got dispensation to participate in the performance assessment, although they failed on a theory test. In total, 2315 PDI's participated at least once in the Performance Assessment Lesson.

Analyses

Psychometric analyses were applied on the data of the three theory tests. Each test had been administered in many different versions, drawn from an item bank. For each of the theory tests 15 versions with a substantive number of participants were selected for analysis. This resulted in sample sizes of n=3013, n=2524 and n=2771 for the tests Driving, Lesson Preparation and Instruction and Coaching respectively. The number of items involved in these three tests were k=211, k=201 and k=148 respectively. Using a one parameter logistic Item Response Theory (IRT) model (Verhelst, & Glas, 1995) the true ability was estimated. As these tests are used for certification purposes, it is important to know the measurement accuracy at the pass-fail boundary ability level. The use of an IRT model enables us to locate all different versions of the theory test on the same latent ability scale, and therefore compare different versions by the level of ability needed to pass the test. For each test version, a level of true ability at the cut-off score was estimated. The resulting latent estimates from different test versions are directly comparable. In addition, after knowing the needed true ability to pass the test on the one hand and the actual applied cut-off score on the other hand, it is possible to calculate misclassifications. These are participants who had been classified incorrectly as 'failed' or 'passed', due to measurement errors. To determine the reliability of the final Performance Assessment Lesson, principal component analysis and alpha reliability analyses were carried out to obtain a limited number of interpretable reliable criterion variables. Correlations between three latent abilities for the theory tests on the one hand and the scores on the resulting scales for instruction and coaching on the other hand were computed to determine the degree of predictive validity.

RESULTS

Table 1 shows the mean cut-off scores for all three theory tests, in terms of required true ability scores (mean=100; SD =15). Two things can be noted. First, to pass, for all tests the required true ability (86.5, 85.3, 90.2 respectively) is less than what the average prospective driving instructor achieves (mean 100). Second, there are differences between the required ability levels across versions. However, these are relatively small, when the standard deviations of the respective cut-off scores (3.1, 5.5, 2.5) is compared to the standard errors of the respective cut-off scores (10.0, 9.6, 7.7). The differences between the test versions fall within acceptable ranges.

Table 1. Cut-off scores for the three theoretical tests expressed in true ability

Theory tests	Average required ability	SD	Min	Max	Mean ability in population	Standard error cut-off score
Theory of Driving	86.5	3.2	81.9	92.9	100	10.0
Lesson Preparation	85.3	5.5	77.3	93.1	100	9.6
Instruction and Coaching	90.2	2.5	86.9	95.5	100	7.7

Table 2. Misclassifications based on the comparison between the outcome that would hold for on true ability at the pass-fail boundary and the actual pass-fail decision

		Actual decision Theory of Driving		Actual Decision Lesson Preparation			Actual decision Instruction and Coaching		
		Fail	Pass		Fail	Pass		Fail	Pass
True ability decision	Fail	20.5	4.9	Fail	24.3	8.2	Fail	23.6	5.9
	Pass	8.2	66.4	Pass	13.2	54.3	Pass	9.4	61.1

Table 2 shows the numbers of misclassifications that arise when pass-fail decisions based on true scores are compared with the actual pass-fail decisions. The mean true score for the all test versions at the pass-fail boundary was chosen as a ‘true’ cut-off. From there, the numbers of failed and passed PDI’s based on true ability could be calculated and compared against the actual pass-fail decision. The results show that the percentages of incorrectly passed PDI’s amount to 4.9%, 8.2% and 5.9% for the three tests respectively. These PDI’s lack the true ability to pass, although they passed. The percentage of PDI’s who had the ability but were failed incorrectly amounted between 8.2% and 13.2%. Inspection of the results for the separate versions shows that in some versions of the Lesson Preparation Test the number of wrongly failed PDI’s went up to 28%. This test version was more difficult than the others, which means that a given raw score represented a higher true ability score than an identical raw score on another test version. Principal component analyses on the items of the Performance Assessment Lesson resulted in three clearly interpretable factors. Three fairly reliable scale scores could be composed, representing aspects of coaching: Motivational Support (6 items, alpha= .77), Diagnosis and Task Support (8 items, alpha= .79) and Instruction (15 items, alpha= .83). The Motivational Support Scale correlated .46 ($p<.001$) with Diagnosis and Task Support and .45 ($p<.001$) with Instructional skill. Diagnosis and Task Support correlated .66 ($p<.001$) with Instructional Skill. Correlations between the ability scores for the theory tests on the one hand and the scores on the Performance Assessment Lesson on the other hand, were low. The correlation coefficients between the Theory of Driving Test and the subscales for Performance Assessment Lesson do not differ significantly from zero. The ability scores for Lesson Preparation and

Instruction/Coaching show low but significant positive correlations with the subscales of the Performance Assessment Lesson, ranging between .12 and .14 ($p<.05$).

DISCUSSION

Conclusions

The central question in this study was whether decisions made about prospective driving instructors, as they follow from the results on theory tests of the innovated exam are valid and fair for the PDI's involved.
First it can be concluded that the overall reliability of estimated ability scores on the theory tests shows acceptable levels. The reliability around the cut-off scores was also acceptable, which seems most important, because here the pass/fail decisions are made. Second, the IRT models showed an acceptable fit, suggesting that the tests represent separable one-dimensional abilities. In addition, the theory tests show discriminative validity. Inter correlations showed that knowledge of the traffic task is an important but not sufficient predictor for knowledge regarding lesson preparation. Third, the predictive value of theory test performance for in-car instructional and coaching performance was very low. The ability scores for lesson planning and instruction and coaching only show very low, although significant, correlations with the in car-performance for coaching and instruction. An explanation for this finding may be that only those who passed the stage one theory exams are allowed to go through the final assessment, after their internship. In addition, the effect of half a year of internship may have washed out initial differences between PDI's. The fairness question was whether the different versions of the theory tests required the same level of true ability to pass. A first finding was that the cut-off scores for pass-fail decisions for the theory tests were well below the average ability level of the population, implicating relatively low ability requirements. The versions used for the Theory of Coaching and Instruction Test and the Theory of Driving Test were equivalent in their ability requirements. For Lesson Preparation there were larger differences in across test versions. It appeared that the number of misclassifications, i.e. PDI's who were wrongly classified as failed or passed, based on their true abilities, differed across test versions. This implies that it was challenging for the exam constructors to assemble truly equivalent test versions

Implications for exam design

As far as the construction and delivery process of the innovated exam concerned, some problems need further attention. Many of these are related to the way the test versions are delivered. The exam is computer-based and takes place at an exam office, where different versions are drawn from items banks, to counter the effects of public item exposure.
Each individual test version needs to represent all sub domains (for each test at least 9), mental activities (E.g. perception, decision making, action execution), and key situations (E.g. learner characteristics, stage of acquisition, traffic situation). The relatively small size of the item bank (300 items) resulted in the frequent reuse of items, which may have led to overexposure of the items, which may result in the lowering of item difficulties.
In addition, inspection of item parameters showed that a part of the items had poor quality, e.g. low or highly negative item-test correlation coefficients, and extreme p-values (near zero or one). In the current examination practice poor items were not excluded ad posteriori from the tests, because shorter test versions would not have been accepted by stakeholders. However, it would

have been defendable to estimate ability levels based on a smaller 'cleaned' subset of items, yielding a more reliable and still representative score. An optimal approach to warrant acceptable item quality is to pre-test all items before putting them into item banks. This however seems problematic because of the risk of early item exposure. In addition, exam costs would rise. Nevertheless, in general it can be recommended to use exam data to improve the exam on the fly. Following the evidence-based design model (Mislevy, Steinberg, & Almond (2003), many design requirements can be investigated on the fly: the competence model reflected in the IRT model should show fit. If not, adjustments are needed. The cut-off scores should represent what we want PDI's to know and to be able to. Item parameters can be traced back to the way the item was designed. In follow-up research, we intend to take a closer look at other parts of the exam, the functioning of different item types, the way items are presented, the stimuli used in items, the responses that are asked and the way these are related to estimates of PDI's abilities. To evaluate the long term effects of the innovated exam for instructional practice, learner driver gain and crash involvement, longitudinal research will be necessary. In such a study one should take into account the quality of all subsequent educational interventions and related driver activities to determine whether there is a case for driver training (Beanland et al., 2013).

REFERENCES

Bartl, G., Gregersen, N.P., & Sanders, N. (2005). *EU MERIT Project: Minimum Requirements for Driving Instructor Training.* Final Report. Vienna: Institut Gute Fahrt.

Beanland, V., Goode, N., Salmon, P.M., Lenné, M.G. (2013). Is there a case for driver training? A review of the efficacy of pre- and post-license driver training. *Safety Science*, *51*, 127-137.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438-481.

Gower, J.C. (1971). A General Coefficient of Similarity and some of its Properties, *Biometrics*, *27*, 857-871.

Hatakka, M., Keskinen, E., Gregersen, N.P., Glad, A., Hernetkoski, K. (2002). From control of the vehicle to personal self-control; broadening the perspectives of driver education. *Transportation Research Part F: Psychology and Behaviour*, *5*(3), 201–215.

Isler, R.B., Starkey, N.J., Sheppard, P. (2011). Effects of higher-order driving skill training on young, inexperienced drivers' on-road driving performance. *Accident Analysis Prevention*, *43*, 1818–1827.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In: G.H. Fischer & I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments and applications* (pp. 215-239). New York: Springer.