Journal of Traffic and Transportation Engineering 3 (2015) 147-157
doi: 10.17265/2328-2142/2015.03.003

Development and Evaluation of a Competence-Based Exam for Prospective Driving Instructors

Erik C. Roelofs[1], Maria Bolsinova[1], Angela Verschoor[1] and Jan A. M. M. Vissers[2]

1. Measurement and Research Department, Cito, Arnhem 1034, the Netherlands

2. Unit Transport, Royal HaskoningDHV, Amersfoort 1132, the Netherlands

Abstract: According to 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 lesson for driving instruction and coaching. 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 for successive groups of PDIs (prospective driving instructors). Data of 4,741 PDIs, enrolled during the period between January 2010 and October 2012, were used for analysis. The results of psychometric analyses show that the theory tests yielded reliable and fair decisions about instructor certification. The predictive validity of the theory tests for the final performance assessment was low. Implications for the design and on-the-fly maintenance of exam item banks are discussed. Follow-up studies will focus on the question, whether the improved instructor exam produces safer drivers in the end.

Key words: Prospective driving instructors, evidence centered design, validity, competence-based exams.

1. Introduction

1.1 Need for Changed Driving Instructor Exams

A growing consensus among driver trainers and road safety researchers is that driver training should place greater emphasis on higher-order, cognitive and motivational functions underlying driving behavior [6, 10]. This changed conception of driver training has been laid down in the goals for driver education matrix [6]. Recent research seems to support this consensus [2, 7]. Innovative training initiatives appear to counteract overconfidence and address motivational factors such as driving anger, sensation seeking and boredom [8].

Parallel to the doubts raised about the quality of driver training, the quality of driver instructor preparation programs is criticized. The MERIT (Minimum European Requirements for Driving Instructor Training) review study [1] showed that huge variations existed in quality of driver instructor education throughout Europe. The content did not cover higher order skills, such as self-evaluation skills and hazard perception skills. Most programs relied on teacher-focused approaches, which seem to fall short in developing higher order skills.

1.2 Regulating Function of the Exam for Instructor Preparation

In many European countries, the quality of the education of driving instructors is regulated by means of the instructor exam. One may view this as a problem, but on the other side, this also offers opportunities for improvement. An exam that is constructed in such a way that only prospective instructors correctly classified as "proficient" are allowed to enter the profession, may have a positive backwash effect on driver instructor education programs, as in other fields of education. The backwash effect means that teachers teach and students learn what will be tested [3, 4, 15]. As long as the exam content and methods are crucial for the instructor profession, the practice of teaching

Corresponding author: Erik C. Roelofs, Ph.D., research fields: assessment, teaching and driver training. E-mail: Erik.Roelofs@cito.nl.

and learning to the test is desirable rather than harmful.

1.3 Development of a Competence Based Instructor Exam in the Netherlands

In the Netherlands, the exam has been made more congruent with the professional practice during the last 10 years. As part of a new law on driving education in 2003, competence-based outcome standards for prospective driving instructors have been formulated [13]. The most far-reaching change underlying these standards has been the emphasis on performance in critical job-situations with real learner drivers. In addition, supporting knowledge was defined in terms of relevant concepts, principles and decision making skills to be applied in authentic instructional situations.

Based on the standards, a two-stage competence-based exam was designed [16] and put into action in the fall of 2009. Since then, over 6,000 PDIs (prospective driving instructors) have gone through one or more tests.

1.4 Research Questions Regarding the Quality of the Exam

In this research study, the quality of the exam was evaluated. A first major question is whether the assessments have resulted in valid and fair decisions about PDIs. Regarding the tenability of decisions, this paper focuses on the separate theoretical assessments, comprising Stage 1. In addition, their predictive value for instructor performance as demonstrated at the final performance assessment lesson (Stage 2) is studied.

A second major question is whether the exam yields fair results. This question refers to the comparability of different versions of assessments. In the exam under study, items banks are used, from which different sets are drawn to compose exam versions to prevent effects of item exposure and cheating. The question then arises whether one cut-off score implies the same level of required proficiency for different versions. To solve this problem, psychometrical equating methods are common to determine how scores on two different tests can be projected on one (latent) scale [9].

In summary, four research questions are addressed:

(1) To what extent are the individual parts of the exam psychometrically reliable?

(2) To what extent do the different theoretical tests intercorrelate?

(3) To what extent do results on theoretical tests and performance assessment for instructional ability correlate?

(4) Do the used cut-off scores across different versions of the theoretical tests reflect equivalent required levels of proficiency?

2. Design Features of the Competence Based Exam

The exam consists of two stages. The first stage comprises the assessment of the theoretical knowledge base of prospective instructors regarding driving and driving pedagogy. After having passed the first stage, the PDIs receive a provisional instructor license enabling them to enroll in a half year internship at a (certified) professional driving school. In the second stage, after having finished their internships, PDIs are judged on their professional instructional abilities, during a masterpiece lesson involving one of their own learner drivers, whom they have been teaching as an intern. If they pass, the will get a full license for the next 5 years. The design features are described below. A summary of all exam parts is provided in Table 1.

2.1 Design Process

All individual assessments of the exam were designed by means of the evidence-centered design model [11, 12]. The ECD (evidence centered design) model identifies five layers in the design process: domain analysis, domain modeling, conceptual assessment framework, assessment implementation and assessment delivery. These design layers have been gone through successively, whereby a continuous dialogue took place between the assessment designers and different stake holders: a board of instructor

Table 1 Instructor competence profile and used tests for the Dutch prospective driver training exam.

Elaboration of task domains				Assessment method	
				Stage 1 (computer based: knowledge base and cognitive skills)	Stage 2: (on the job: performance assessment)
1. Competence in conscious traffic participation	1.1 Driving responsibly as a first driver The driving instructor is able to drive a vehicle safely, smoothly, socially considerate, and in an eco-friendly way according to Dutch driving standards.	1.2 Verbalizing mental processes of driving The driving instructor is able to verbalize the mental task processes that take place when carrying out driving tasks in different traffic situations.	-	Theory of driving test (60 items): traffic participation rules knowledge, case-based and situational judgment items (Task Domain 1.1).	Performance assessment drive as a first and a second driver (all task domains Cluster 1).
2. Competence in lesson preparation	2.1 Adaptive planning The driving instructor 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 LD (learner driver).	2.2 Elaborating driving pedagogy The driving instructor is able to prepare a driving specific pedagogical learning environment for learner drivers.	2.3 Organizing learning The driving instructor is able to organize lessons in such a way that activities run smooth and without interruptions, ensuring a maximum amount of productive learning time.	Theory of lesson preparation test (60 items): case-based concept application, reasoning and situational judgment items (all task domains Cluster 2).	1. Performance assessment lesson with real learner driver 2. Self-reflection report internship 3. Reflective interview internship (all task domains Clusters 2,3 and 4).
3. Competence in instruction and coaching	3.1 Providing instruction The driving instructor 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 driving instructor is able to monitor learner driver development and guide the LD towards self-regulation in solving driving tasks and driving related tasks.	-	Theory of instruction and coaching test (60 items): case-based concept application, reasoning and situational judgment items (all task domains Cluster 3).	
4. Competence in evaluation, reflection and revision	4.1 Assessing learner progress The driving instructor 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 driving instructor is able to reflect on his own actions and use the results of this reflection for adapting his approach.	-	-	

educators, exam institutes, ICT (information and communication technology) specialists, psychometricians, educational scholars, academic teacher educators, driving examiners and driving instructors.

2.2 Conceptual Assessment Framework

In the exam, the layer of the conceptual assessment framework for assessment task design was of central importance. The conceptual assessment framework helps to sort out the relationships among attributes of a candidate's competence, observations which show competence, and situations which elicit relevant driver performance. The central models for task design are the competence or student model, the task model and the evidence model.

2.2.1 Driving Instructor Competence Model

The competence model encompasses variables representing the aspects of instructor competence that are the targets of inference in the assessment and their inter-relationships. Starting from a literature search on what comprises good teaching in general and more specifically driving instruction, a competence model was constructed. This resulted in the formulation of four domains of competence summarized in Table 1: (1) conscious traffic participation as first and second drivers; (2) lesson preparation; (3) instruction and coaching; (4) evaluation, reflection and revision.

A model of competent task performance formed the basis for two competence models: driving competence [17] and instructional competence [18]. A basic tenet in the model (Fig. 1) is that instructor competence is reflected in the consequences of an instructor's actions. The most important consequences of instructor's actions are students' learning activities, such as listening to an explanation (e.g., "first you scan, then you decide about an action"), practicing an assigned driving task (e.g., merging and left turn) or responding to a question (e.g., "what I did prior to the merging error was that I…"). Consequences relating to driving competence are the results of a driving maneuver as regards safety (e.g., there is just enough space to avoid a conflict) or traffic flow (e.g., other drivers have to wait for the learner driver).

Starting from the consequences, the remaining elements of the instructor competence model can be mapped backwards:

- First, the component "actions" refer to professional activities, e.g., preparing and delivering instruction or coaching to learner drivers;
- Second, any instructor activity is part of a universe of tasks under various task conditions that may be applicable. For instance, instructors will have to plan and adapt their instruction depending on factors like: the learning stage of the learner driver (gaining control of the vehicle, driving in simple situations, driving independently in complex situations), the learning goals to be achieved, the degree of learning progress, the traffic density and weather conditions on the route, or learner characteristics (e.g., self-confidence, motivation and prior experience as biker or moped rider);
- Third, during their instruction, instructors make informed decisions about what to do next. Some decisions are made during preparation of instruction (e.g., planning a route suitable for the LD). Others are made on-the-fly during the driving lesson (e.g., decide to give a hint to the LD or decide to intervene by pushing a pedal or by giving a warning);
- Fourth, when making decisions and performing activities, teachers need to draw from a professional knowledge base. This base relates to: (1) traffic

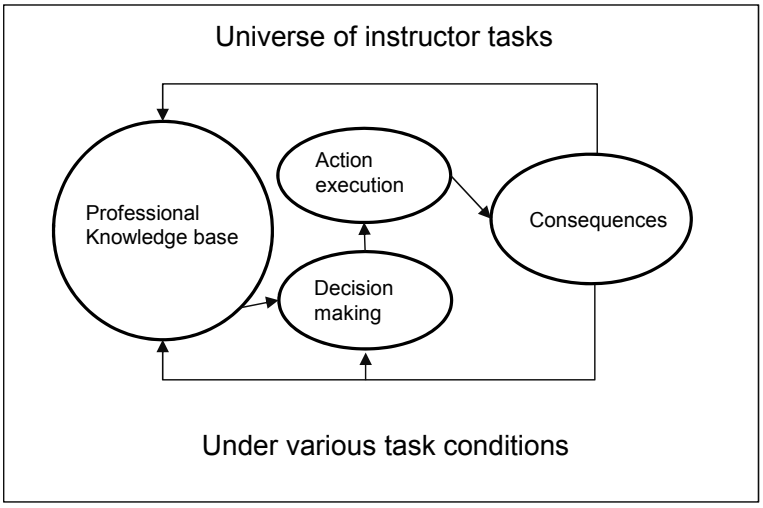

Fig. 1 Model of competent task performance.

psychology, e.g., mental processes that occur during driving, situational awareness during driving; (2) traffic rules and regulations, e.g., meaning of traffic signs, knowledge of speed limits; (3) driving pedagogy, e.g., how to sequence driver learning tasks from simple to complex, how to organize learning and how to provide feedback; (4) assessment of learning progress, e.g., observing driving and judging driving proficiency;

- Finally, proficient driving instructors reflect on their actions: First, the consequences of instructors' actions form input for new decision and action chains; Next, consequences will change the instructors' knowledge base through reflection. This is indicated by the returning arrows in Fig. 1.

2.2.2 Exam Composition: Task and Evidence Models

The tasks employed in the different parts of the exam follow directly from the cognitive activities and interactive activities as mentioned in the competence model. To measure the listed aspects of competence, three theory tests and two performance assessments were employed.

2.2.3 Proficient and Conscious Driving: Knowledge and Performance

To test the PDIs own driving proficiency and their ability to verbalize mental task processes, the PDI had to complete a 60 min 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; (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 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).

2.2.4 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 [14, 19]. 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: (1) the learner will learn less that desirable; (2) the learner will not fully understand your explanation; (3) the learner will have less time to practice new driving tasks; (4) the learner cannot direct his attention to the essential parts of the lesson (correct). A second item type is related to situational judgment items. These items address decision making skills [21]. 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, which 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.

2.2.5 Performance Assessment for Instruction and Coaching Skills

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 through a three-point rubric, representing "counterproductive performance", "beginning productive performance" and "optimal performance". A scoring guide was available for assessors. Examiner agreement scores using Gower's similarity index [5] showed acceptable levels of agreement between examiners, 0.67 for instruction and 0.75 for coaching. The cut-off score for passing the performance assessment was 71 points (out of 102 points).

3. Methods

3.1 Subjects and Data

Test data from 4,741 prospective driving instructors, who enrolled the program between January 1, 2010 and October 1, 2012, were selected. 79% of them were male and 21% female. The mean age was 34.9 years (*SD* (standard deviation) = 10.9): 3,079 (74.4%) of them were born in the Netherlands. The remaining 25.6% originally came from 79 different countries. The majority of them were immigrants from Morocco (*n* (sample size) = 199), Suriname (*n* = 190), Turkey (*n* = 151), Afghanistan (*n* = 112), Iraq (*n* = 89), and Iran (*n* = 46). A total of 4,644 PDIs completed at least one of the theory tests: 2,977 of them passed all their theory tests, from which 1,941 PDIs took part in the performance assessment lesson. From the remaining PDIs, 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. Three hundred and sixty-eight PDIs got dispensation to participate in the performance assessment, although they failed in a theory test. In total, 2,315 PDIs participated at least once in the performance assessment lesson.

3.2 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 $n = 3{,}013$, $n = 2{,}524$ and $n = 2{,}771$ for the tests driving, lesson preparation and instruction and coaching, respectively. The number of items (k) involved in these three tests were $k = 211$, $k = 201$ and $k = 148$, respectively (Table 2).

Using a one parameter logistic IRT (item response theory) model [20], the true ability of PDIs was estimated. As these tests are used for certification

Table 2 Number of test versions, total number of items and sample size chosen.

Test	Number of test versions selected	Minimum number of responses per version	Maximum number of responses per version	Total number of items	Sample size
Theory of driving	14	99	484	211	3,013
Theory of lesson preparation	15	38	586	201	2,524
Theory of instruction and coaching	15	32	551	148	2,771

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, to compare different versions by the level of ability needed to pass the test [9]. 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.

4. Results

Fig. 2 shows a normal distribution of ability scores, (M = 100, SD = 15). In this figure, the cut-off scores for the 14 most frequently administered versions of the theory of driving test are plotted. The dots represent the cut-off scores for each test version expressed in terms of ability that is required to pass the exam (theta). The lines represent the 95% confidence interval for the cut-off score. Two things can be noted. First, the cut-off scores for all test versions fall below the average ability in the total population. The mean cut-off score for the theory of driving test (M = 85.3 in Table 2) is almost one standard deviation below the ability mean of the population. This means that relatively low ability (M = 86.5) was needed to pass the test. Second, there are small differences between the required ability levels for different test versions, but the variation of the cut-off levels (SD = 3.22) across versions is small compared to the standard error.

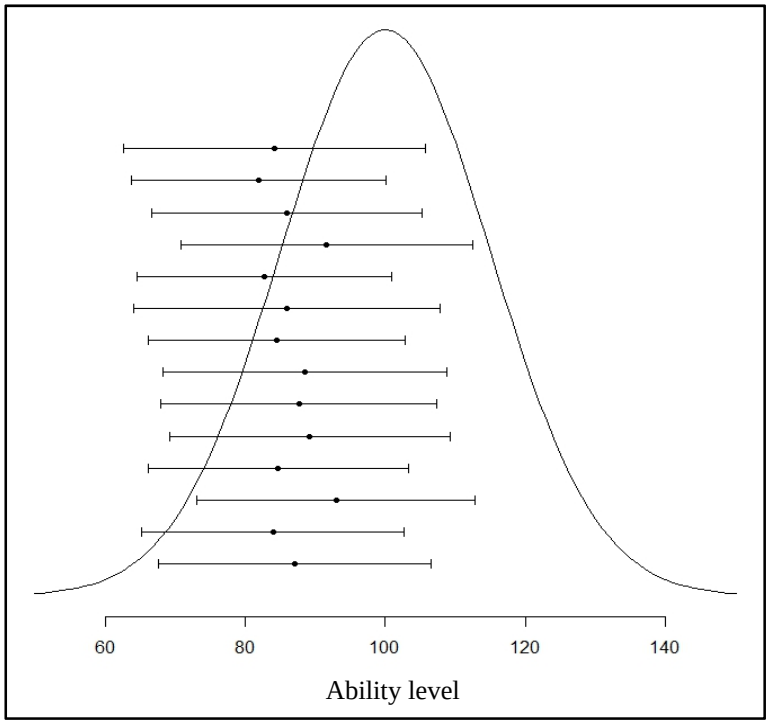

Fig. 2 Cut-off scores and with 95% confidence intervals for 14 versions of the theory of driving test.

Table 3 shows the mean cut-off scores for all three theory tests, in terms of required true ability scores (M = 100, SD = 15). The same things can be noted for all theory tests: First, to pass, for all tests, the required true ability (86.5, 85.3 and 90.2, respectively) is less than what the average prospective driving instructor achieves (M = 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 and 2.5) are compared to the standard errors of the respective cut-off scores (10.0, 9.6 and 7.7). The differences between the test versions fall within acceptable ranges.

Table 4 shows the numbers of misclassifications that arises when pass-fail decisions based on true ability scores are compared with the actual pass-fail decisions. The mean true ability level for all test versions at the cut score was chosen as a "true" cut-off. From there, the numbers of failed and passed PDIs based on true ability could be calculated and compared against the actual pass-fail decision, which was based on the raw test scores. The results show that the percentages of incorrectly passed PDIs amount to 4.9%, 8.2% and 5.9% for the three tests, respectively. These PDIs lack the true ability to pass, although they did pass the test. The percentages of PDIs who had the ability but failed incorrectly amounted between 8.2% and 13.2% for the three respective theory tests. Inspection of the results for the separate versions shows that, in one version of the lesson preparation test, the number of wrongly failed PDIs amounted to 28%. This test version was more difficult than the others, which also means that a given raw score represented a higher true ability score than an identical raw score on another test version.

The correlation coefficients between the ability scores on the three theory tests show a moderate correlation of 0.56 between the theory of driving test and the theory of lesson preparation test. Ability scores on the theory of driving test correlate 0.43 with the ability scores on the theory of instruction and coaching test. Finally, ability scores on the theory of lesson preparation test correlate 0.29 with ability levels on the theory of instruction and coaching test.

Principal component analyses on the item score data 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 (six items, Cronbach's α for reliability = 0.77), diagnosis and task support (eight items, α = 0.79) and instruction (15 items, α = 0.83 (Table 4)).

The motivational support scale correlated 0.46 (probability of null hypothesis that variables have zero correlation, $p < 0.001$) with diagnosis and task support and 0.45 ($p < 0.001$) with instructional skill. Diagnosis and task support correlated 0.66 ($p < 0.001$) with instructional skill.

Table 3 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 4 Misclassifications based on the comparison between the outcome that would hold for true ability at the pass-fail boundary and the actual pass-fail decision.

Decision based on true ability		Actual decision theory of driving		Actual decision lesson preparation		Actual decision instruction and coaching	
		Fail	Pass	Fail	Pass	Fail	Pass
True	Fail	20.5	4.9	24.3	8.2	23.6	5.9
Ability decision	Pass	8.2	66.4	13.2	54.3	9.4	61.1

Table 5 Psychometric report for the performance assessment lesson.

Subscale	*N*	Minimum	Maximum	*Mean*	*SD*	*α*
Coaching: motivational support (six items)	580	9.00	18.0	14.3	2.2	0.77
Coaching: diagnosis and task support (eight items)	580	9.00	24.0	16.9	2.8	0.79
Instructional skill (15 items)	580	21.00	44.0	34.7	4.3	0.77
Exam score (34 items)	580	50.00	99.0	78.2	8.9	0.88

Table 6 Correlations between performance on theory tests and performance assessment lesson.

Subscale final performance assessment lesson	Theory of driving	Theory of lesson preparation	Theory of instruction coaching
Coaching: motivational support (6 items)	0.01	0.06	0.13*
Coaching: diagnosis and task support (8 items)	0.07	0.12*	0.09
Instructional skill(15 items)	0.10	0.12*	0.11*
Overall assessment score (34 items)	0.07	0.12*	0.14**

*Significant, $p < 0.05$; **Significant, $p < 0.01$.

Table 5 shows the correlations between the ability scores on the theory tests and the scores on the performance assessment lesson, for the three subscales and the overall assessment score. These correlations can be regarded as measures of the predictive value of theory tests for coaching and instruction skills shown in practice, also denoted with the term “predictive validity”. The correlation coefficients for theory of driving test do not differ significantly from 0. The ability scores for lesson preparation and instruction/coaching show six low but significant correlations with the subscales and the overall scale for the final performance assessment lesson (between 0.12 and 0.14, $p < 0.05$).

5. 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 PDIs 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. Intercorrelations 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 1 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 PDIs.

Regarding fairness, the 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 theory of coaching and instruction test and the theory of driving test had comparable cut-off scores across versions and were hence equivalent in their ability requirements. For lesson preparation, there were larger differences in required ability across test versions. It appeared that the number of misclassifications, i.e., PDIs 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.

6. Discussion and Practical Implications for Test 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 nine), mental activities (e.g., perception, decision making and action execution) and key situations (e.g., learner characteristics, stage of acquisition and 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 0 or 1). In the current examination practice, poor items were not excluded from the tests after they had been administered, 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 on a representative sample of target candidates before putting them into item banks. This however seems problematic because of the risk of early item exposure. In addition, exam costs would rise.

However, in general, it can be recommended to use exam data to improve the exam on-the-fly. Optimizing the assembly of different versions by using the item statistics that were known at that time, would already have stabilized the number of passed and failed PDIs to a large degree. Originally, the number of passed PDIs varied between 56% and 76% across 14 versions of the theory of instruction and coaching test. An optimization analysis showed that after exchanging items between these versions, the variation in pass rates could have been reduced to a range between 65.8% and 67.8%.

Following the evidence-based design model of Mislevy et al. [11], 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 PDIs to know and to be able to. Certain item characteristics should be traced back to the way the item was designed. In short, using an evidence-based design model, in combination with on-the-fly improvements, can improve our decisions about prospective driving instructors.

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 PDIs 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 [2].

References

[1] Bartl, G., Gregersen, N. P., and Sanders, N. 2005. *EU MERIT Project: Minimum Requirements for Driving Instructor Training*. Final report of The International Commission for Driver Testing. Accessed July 15, 2015. http://www.cieca.eu/template_subpage.asp?pag_id=49&spa_id=63&lng_iso=EN.

[2] Beanland, V., Goode, N., Salmon, P. M., and Lenné, M. G. 2013. "Is There a Case for Driver Training? A Review of the Efficacy of Pre- and Post-licence Driver Training." *Safety Science* 51: 127-37.

[3] Crooks, T. J. 1988. "The Impact of Classroom Evaluation Practices on Students." *Review of Educational Research* 58: 438-81.

[4] Fredericksen, J. R., and Collins, A. 1989. "A Systems Approach to Educational Testing." *Educational Researcher* 18 (9): 27-32.

[5] Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27: 857-71.

[6] Hatakka, M., Keskinen, E., Gregersen, N. P., Glad, A., and 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): 201-15.

[7] Isler, R. B., Starkey, N. J., and Sheppard, P. 2011. "Effects of Higher-Order Driving Skill Training on Young, Inexperienced Drivers' On-road Driving Performance." *Accident Analysis Prevention* 43: 1818-27.

[8] Isler, R. B., Starkey, N. J., and Williamson, A. R. 2009. "Video-Based Road Commentary Training Improves Hazard Perception of Young Drivers in a Dual Task." *Accident Analysis Prevention* 41: 445-52.

[9] Kolen, M. J., and Brennan, R. L. 1995. *Test Equating*. New York: Spring.

[10] Mayhew, D. R., and Simpson, H. M. 2002. "The Safety Value of Driver Education and Training." *Injury Prevention* 8: 3-8.

[11] Mislevy, R. J., Steinberg, L. S., and Almond, R. G. 2003. "On the Structure of Educational Assessments." *Measurement: Interdisciplinary Research and Perspectives* 1: 3-67.

[12] Mislevy, R. J., and Haertel, G. 2006. *Implications of Evidence-Centered Design for Educational Testing.* A PADI (Principled Assessment Designs for Inquiry) technical report.

[13] Nägele, R., Vissers, J., and Roelofs, E. C. 2006. *Revision Law on Motor Vehicle Education: A Model for a Competence Based Exam.* Amersfoort/Arnhem: DHV, Cito.

[14] Norman, G, Swanson, D. B., and Case, S. M. 1996. "Conceptual and Methodological Issues in Studies Comparing Assessment Formats." *Teaching and Learning in Medicine* 8 (4): 208-16.

[15] Pellegrino, J. W. 2014. "Assessment as a Positive Influence on 21st Century Teaching and Learning: A Systems Approach to Progress." *Psicología Educativa* 20: 65-77.

[16] Roelofs, E, and Vissers, J. 2008. *Toets Specificities Theoretisch Deeltoetsen WRM-Examen (Specifications of the Theoretical Exams for Driver Instructor Candidates).* Amersfoort: DHV Group.

[17] Roelofs, E. C., Onna, M., and Van Vissers, J. 2010. *Development of the Driver Performance Assessment: Informing Learner Drivers of their Driving Progress,* edited by Dorn, L. Hampshire: Ashgate Publishing Limited, 37-50.

[18] Roelofs, E., and Sanders, P. 2007. "Towards a Framework for Assessing Teacher Competence." *European Journal for Vocational Training* 40 (1): 123-39.

[19] Schuwirth, L. W. T., Verheggen, M. M., Van der Vleuten, C. P. M., Boshuizen, H. P. A., and Dinant, G. J. 2000. "Validation of Short Case-Based Testing Using a Cognitive Psychological Methodology." *Medical Education* 35: 348-56.

[20] Verhelst, N. D., and Glas, C. A. W. 1995. "The One Parameter Logistic Model." In *Rasch Models: Foundations, Recent Developments and Applications,* edited by Fischer, G. H., and Molenaar, I. W. New York: Springer, 215-39.

[21] Whetzel, D. L, and McDaniel, M. A. 2009. "Situational Judgment Tests: An Overview of Current Research." *Human Resource Management Review* 19: 188-202.