The development of hazard perception testing

Prepared for Road Safety Division, Department for Transport

G B Grayson and B F Sexton

TRL Report TRL558

First Published 2002
ISSN 0968-4107

This report has been produced by TRL Limited, under/as part of a contract placed by the Department for Transport. Any views expressed in it are not necessarily those of the Department.

TRL is committed to optimising energy efficiency, reducing waste and promoting recycling and re-use. In support of these environmental goals, this report has been printed on recycled paper, comprising 100% post-consumer waste, manufactured using a TCF (totally chlorine free) process.

CONTENTS

Executive Summary

The decision to add a hazard perception test to the Theory Test that must be passed before learner drivers can take their practical driving test was preceded by an extensive programme of research and development. This report describes this work and its background.

Introduction

It has long been recognised that young drivers are over-represented in accidents. There has been an equally long debate about whether the 'young driver problem' arises because young drivers are immature, or because they are inexperienced. A basic problem is that age and experience are usually highly correlated, in that the majority of inexperienced drivers are also young drivers. The age versus experience issue is not just of academic interest, for there are clear implications for safety countermeasures. At a practical level, maturity cannot be accelerated, while the lessons of experience can - in principle - be taught. A number of TRL studies using multivariate modelling of accidents have established that it is experience that is the main contributor to the higher accident rates of young drivers, with one study (Forsyth e*t al.*, 1995) showing that for young drivers the effect of experience alone over the first three years of driving was some four times that of age.

Background

In the first few years, a new driver is learning not just new skills, but formulating new rules, developing a new repertoire of strategies, and learning new patterns of interaction. However, much of this process takes place in an unstructured and informal way, with no guarantee that what is learned is the most appropriate for the safety of the traffic system. Given the importance of experience in reducing accident liability, it would clearly be desirable if the lessons of experience could be imparted by some formal intervention, rather than being acquired in an uncontrolled learning situation. A candidate for intervention of this sort needs to be a variable that is:

- capable of being measured objectively and reliably;
- related to driving experience;
- amenable to improvement through training; and
- related to accident involvement.

A review of the literature provides strong support for the use of hazard perception. The theoretical basis is sound, and there is good evidence that it is a skill that improves with experience and expertise. Research has shown that hazard perception is capable of being improved through training, and a number of studies have indicated that poor hazard perception skills are associated with elevated accident risk. Hazard perception thus became a candidate for inclusion in the licensing system. The justification for this is that if hazard perception ability is related to accident involvement, and if the ability is amenable to improvement, then introducing hazard perception training on a wide scale could lead to a reduction in accident liability in the early years of driving. Further, the most effective way of ensuring that hazard perception training is undertaken on a voluntary basis is to include a hazard perception test in the procedure for acquiring a licence to drive.

Research

Although a number of research procedures for measuring hazard perception existed, none was suitable for use as a licensing tool, and further development work was necessary. The first stage was a contract with the National Foundation for Educational Research (NFER).

The NFER study set out to demonstrate that it was possible to produce tests that were sufficiently reliable in psychometric terms to be used in the licensing process. In this they were successful; following extensive testing, four tests were produced with high levels of reliability, three of which distinguished between inexperienced and experienced drivers. However, a number of subsequent studies found that the tests were not able to demonstrate an association with accident liability.

A further programme of work was undertaken by TRL to assess the likely effectiveness of introducing hazard perception testing, and one of its first tasks was to evaluate the hazard perception items in the NFER tests that discriminated better on a driver experience scale than others. The 'best' items were found to be those that required good scanning practices and an ability to anticipate potentially hazardous situations. The sub-set of NFER items selected with these characteristics was able to discriminate on the driver experience scale, and appeared to have some association with retrospective accident liability.

To take this further, a broad-based 'blueprint' was developed to define the type of hazard perception items required. Further items were filmed using staged situations as well as opportunistic filming, and produced enough usable items to evaluate in a trial on different driver experience groups. The results of the trial demonstrated that there were enough items that had the necessary psychometric characteristics to allow for the construction of new tests.

While this test development was being undertaken, a specification for suitable hazard perception training material was being developed. This involved reviewing existing driver training material as well as discussions with the driver training industry. The resultant specification was given to DSA who then created a training package. This training package contained three modules – basic, plus two more advanced levels of training. Each module took about an hour to administer and was group-based training using video material with a structured schedule, which could be delivered by non-driver training professionals.

The final task in this research was to run a trial of learners who were ready for their theory test, and to test their hazard perception before and after hazard perception training. Using the bespoke package developed specifically for this project showed that three hours of training resulted in learner drivers obtaining similar hazard perception scores to those of experienced drivers. The conclusion, therefore, is that hazard perception skills can be trained in learner drivers. Improving the hazard perception skills of new drivers should have a beneficial effect on their accident liability, particularly in the first 2 or 3 years of driving which is when they would normally be gaining these skills.

Summary

Research has shown that lack of experience is a major contributor to the high levels of accident liability among new and young drivers, and that hazard perception skills improve with experience. There are strong grounds for thinking that if the lessons of experience could be imparted by formal instruction, rather than in an uncontrolled learning situation, then there could be benefits for road safety.

Research that has been carried out in recent times shows that hazard perception training could contribute to this. Hazard perception tests have been developed that are reliable in a statistical sense, and trials have shown that training can improve the performance of novices to match that of experienced drivers. These findings were key to the decision to incorporate a hazard perception component within the Theory Test at the end of 2002.

Further tests have been developed and calibrated for use in a licensing context, and pass marks for different groups of test candidates have been identified. Details of this implementation work will be available in a report from DSA.

Reference

Forsyth E, Maycock G and Sexton B (1995). *Cohort study of learner novice drivers. Part 3: Accidents, offences and driving experience in the first three years of driving.* Project Report PR111. Crowthorne: TRL Limited.

1 Introduction

Young drivers are an old problem. For example, thirty years ago Goldstein (1972) was able to state ‘that youthful drivers are over-represented in accidents ... has been well known for several decades’, and a major review a few years later concluded that ‘young drivers are over-represented at all time periods and at all levels of severity of accidents’ (OECD, 1975). Young drivers have been a major focus of research and policy in traffic safety in the developed countries of the world for many years.

There has also long been a debate about whether the ‘young driver problem’ arises because young drivers are immature, or because they are inexperienced. A basic problem is that age and experience are usually highly correlated, in that the majority of inexperienced drivers are also young drivers. The age versus experience issue is not just of academic interest, for there are clear implications for safety countermeasures. At a practical level, maturity cannot be accelerated, while the lessons of experience can - in principle - be taught. Attempts to disentangle the separate effects of age and experience have often encountered problems in the past, largely because of the methodological issues involved.

In the late 1980s, a programme of research was started at TRL with the aim of making more explicit the separate effects of age and experience on accident liability. The approach adopted was a statistical one using Generalised Linear Modelling techniques. The first study in the programme was a cross-sectional survey of 13,500 drivers, in which older and younger drivers, and drivers with limited experience were over-sampled (Maycock *et al.*, 1991). In a later study, information was obtained from some 7,000 novice drivers, in which cohorts of drivers reported accidents year by year for the first three years of their driving experience (Forsyth *et al.*, 1995).

Although earlier investigators (e.g. Mayhew and Simpson, 1990) had concluded that age had more effect on accident risk than did experience, the results of the modelling exercise in the cohort study showed that for young drivers the effect of experience alone over the first three years of driving was some four times that of age. There was a 40% reduction in accident liability between the first and second years of driving for 17 - 18 year olds attributable to experience.

The reduction in accident liability in the first years of driving was most marked in the youngest age groups, but was evident at *all* ages, raising the question of just what happens in the first two or three years of driving to turn a high risk novice into a low risk driver.

Analyses of the driving task (e.g. Grayson, 1991) suggest that in the first few years a new driver is learning not just new skills, but formulating new rules, developing a new repertoire of strategies, and learning new patterns of interaction. However, much of this process takes place in an unstructured and informal way, with no guarantee that what is learned is the most appropriate for the safety of the traffic system. Given the importance of experience in reducing accident liability, it would clearly be desirable if the lessons of experience could be imparted by some formal intervention, rather than being acquired in an uncontrolled learning situation.

There have been two main strands in research on young and new drivers that could offer candidates for intervention. The first is concerned with motivation, the second with cognition. Both are active and important fields of research, but the possibilities for intervention are different. The modification of motivational and attitudinal factors at a societal level is a long-term process. This is not to deny that marked changes in social norms can take place (drink-driving being the obvious example), but such changes do not come about quickly or easily. By contrast, cognitive variables offer - in principle - more scope for intervention at an individual level. A candidate for this needs to be able to meet four criteria. It should be:

- capable of being measured objectively and reliably;
- related to driving experience;
- amenable to improvement through training; and
- related to accident involvement.

Several cognitive measures that have been considered in the past, such as selective attention and field dependence, have high face validity but have not shown consistent relationships with accident involvement in well-designed studies. Their lack of success may perhaps be attributed to the fact that the manner in which they were assessed was far removed from the ‘real world’ of driving - by listening to digits through earphones in one case, and examining complex geometric patterns in the other. However, there is one further candidate that is much more closely based on driving, that has high face validity, and has a sound research background. This is hazard perception.

Hazard perception is a surprisingly old concept, dating back nearly 30 years to the work of Currie (1969), who used model cars to assess a subject’s speed of response in identifying potential collisions. He found that accident-involved subjects responded to these more slowly than did the accident-free, although the groups did not differ in simple reaction time. This finding set the scene for further research on the topic in two ways: first, by suggesting that there was some ability over and above simple reaction time that was related to accident involvement, and second by showing that it could be measured using simulation in a laboratory setting.

Pelz and Krupat (1974) developed this line of research further, using cine-film of ‘real road’ situations in a non-interactive driving simulator. They found that students with ‘good’ driving records detected danger more promptly than did students with records of violations or accidents.

Quimby and Watts (1981) made an important contribution by comparing the performance of subjects both on the roads and in a simulator. A variety of measures were taken in the simulator, both continuous and in response to discrete events. Of these two, it was found that only response time to hazard was related - albeit weakly - to accident history.

2 Hazard perception as a research tool

2.1 Developments

In the 1980s and 90s there was growing interest in higher order cognitive skills and their influence on driver safety. The process of perceiving and responding to hazards was regarded as being of particular importance from a theoretical point of view, and this led to a search for tests and measures that could be used for driver assessment.

The early studies of Pelz and Krupat (1964) and Quimby and Watts (1981) had used cine-film in driving simulators. Cine was rapidly being replaced by video, and interest was growing in more portable laboratory-based, as opposed to simulator-based environments. Hazard perception tests using video material were developed for research purposes at Reading University by McKenna and Crick, and in parallel at TRL, where they were used in studies of both driver impairment (Irving and Jones, 1992) and accident liability. At the same time, in the Australian state of Victoria, tests of hazard perception were being developed as part of the licensing procedure for new drivers (Hull and Christie, 1993).

It is of interest to note that many of the early workers in this field seem to have been more concerned with measurement than with definition. Pelz and Krupat (1964) did not use the term ‘hazard perception’, preferring the phrase ‘caution profile’ instead. Quimby and Watts (1981) referred to ‘reaction to hazardous events’, while Quimby *et al.* (1986) introduced more than a hint of circularity in defining hazard perception as ‘the ability to perceive and respond to hazards’. McKenna and Crick (1991) broadened the concept by referring to ‘ability to respond to potentially dangerous traffic situations’ - but without defining what these might be. Despite the somewhat limited contribution of researchers to the definition of hazard perception, at a practical level the concept seems to be easily recognised by trainers and instructors in their use of terms such as ‘appreciation’, ‘anticipation’, and ‘reading the road’.

2.2 Accident liability studies

The Pelz and Krupat and Quimby and Watts studies had shown relationships between hazard perception scores and accident involvement, but were based on relatively small samples and used relatively unsophisticated statistical techniques.

Stronger evidence was provided by the study by Quimby *et al*. (1986) in which hazard perception was included in a large test battery. This study employed multivariate statistical techniques that could deal with the high degree of inter-correlation among variables that had made interpretation of results difficult in the past. Once age and mileage had been accounted for, none of the low-level sensory or psychomotor variables made any contribution to explaining the variability in accident involvement. One of the few factors that did contribute was response time to hazards.

Hull and Christie (1993) reported on the ‘Geelong Trial’, in which the test developed in Victoria was tried out before being introduced into the licensing system. The trial involved more than 3000 participants, and the results indicated a significant difference in hazard perception scores in the expected direction between accident-involved and accident-free subjects over a wide age range.

McKenna and Crick (1991) initially used experience as a surrogate for accident liability, and developed a video-based test that showed a monotonic improvement in hazard perception scores from novice, through experienced, to expert drivers, albeit with relatively modest sample sizes. However, a later paper by McKenna and Horswill (1999) reported that HP scores on the test developed at Reading University significantly discriminated between large samples of accident-involved and accident-free drivers.

2.3 HP training studies

The ‘trainability’ of hazard perception was first demonstrated in an experiment by McKenna and Crick (1994), which was a before-and-after study of drivers participating in a RoSPA advanced driving course. Drivers taking the course showed significant improvements in hazard perception following the course compered with a group of matched controls, even though the course did not deal specifically with hazard perception. A later experiment (McKenna and Crick, 1997) suggested ways in which the content of training could be made more efficient, and found that 2 - 3 hours of classroom training using video material could improve the hazard perception performance of novice drivers to the level of experienced ones.

A study of over 200 newly qualified young drivers (Mills *et al.*, 1998) also showed that scores on HP tests could be improved by training. The authors reported that on-road plus classroom training was the most effective in improving test performance (though the duration of training was greater for those receiving both forms), but classroom training alone was capable of achieving significant improvements.

These findings were important in that there was now evidence that hazard perception was related to experience, it was associated with accident involvement, and it was capable of improvement through training. It thus became a candidate for inclusion in the licensing system.

It should be noted that the primary aim was not simply to restrict entry into the driving population by selecting only those candidates with good hazard perception abilities, but rather to encourage novice drivers to undertake appropriate training. The argument is that if hazard perception ability is related to accident involvement, and if the ability is amenable to improvement, then introducing training on a wide scale could lead to a reduction in accident liability in the early years of driving. Further, the most effective way of ensuring that hazard perception training is undertaken on a voluntary basis is to include a hazard perception test in the procedure for acquiring a licence to drive.

3 From research to policy

3.1 Procedures

The transition from research tool to 'institutionalised' test instrument is a difficult one. The most obvious issue is that of psychometric reliability; highly desirable in a research tool, but essential in a test instrument. An associated topic is face validity, which again is essential if a public test procedure is to be deemed acceptable. Subjects in laboratory experiments will happily submit to incomprehensible test protocols if they are being paid for doing so. By contrast, if they have to do the paying, then the same subjects will expect the benefits of taking and passing the test to be clear to them (as should be the benefits of less competent subjects failing the test).

There is also the question of the experimental setting. Research is carried out in laboratories, with procedures for subject handling that are designed to encourage co-operation. Test instruments are employed in public test centres, where keen experimenters are largely replaced by impersonal machines. In laboratories, instructions can be detailed and comprehensive (and subjects incapable of understanding them can easily be dropped from the sample). By contrast, instructions on official test instruments have to be capable of being understood by as wide a range of the population as possible if the tests are not to be open to the criticism that they are testing intellectual ability rather than the ability to drive in a safe manner. To illustrate this point, the original TRL test, although a complex task, had written instructions that ran to over 800 words. This would clearly be unacceptable in the context of an official test conducted as part of the licensing process.

3.2 The issue of measurement

The measurement of skill is not a simple process, often because skills are not simple to define. With hazard perception, most researchers have taken the line that the skill in question is best measured in terms of response to hazard. The basic principle here is that 'response' is not the same as 'reaction'. This is critical, because performance on a reaction time task becomes poorer with age (and thus largely also with experience), whereas the skills that a hazard perception test is intended to measure should be expected to improve with experience. All previous hazard perception research tools have attempted to show that they measured something over and above simple reaction time; this in turn becomes a requirement for any future test.

There have been a number of approaches to the measurement of hazard perception skills. The TRL test, as used by Irving and Jones (1992), followed earlier TRL work in recording both assessment of hazard and response to hazard. The measures taken included response time to specific hazards, the numbers of hazards detected, the average assessment of hazard, and the variability in this assessment.

The test developed by McKenna *et al.* at Reading University was also video-based, but differed in that it focused on measuring the time of response to hazard. This was done for two reasons. First, it avoided the possibility of confusion between what are held to be the theoretically distinct concepts of hazard assessment and hazard response, and second, because the previous work of Quimby *et al.* (1986) had shown that it was response time to hazard that was most strongly associated with accident involvement. The original Reading test had 35 items, selected on the basis that they distinguished best between novice and experienced drivers in pilot trials. Later work was based on two parallel tests with 13 and 16 scenarios each. The 'score' for the test is the mean response latency to the presented hazards, where latency is defined as the time taken to respond from the start of the hazardous event.

The test that has been incorporated in the licensing system in the Australian state of Victoria is very different. The VicRoads test (as it is known) is a PC-based test using video clips of traffic scenes, and the items in the test are derived from an analysis of road accidents in Victoria in which young drivers are over-represented. In the test, subjects are asked to indicate on a touch screen if and when they would initiate a manoeuvre in a particular situation. The situations include close following, overtaking, curve negotiation, and gap acceptance. The test is relatively short, with only 12 items. The scoring is also simple, in that it is based on the subject giving the correct response to a short video clip lasting 7 seconds (note that the correct response can be to refrain from using the touch screen).

With its emphasis on the initiation of manoeuvres, and the nature of the material used, the VicRoads test is unlike that of the other tests developed in the UK. It would appear therefore that the term 'hazard perception' as used in connection with the Australian tests is a rather different concept from that being used in the UK. Some empirical support for this view comes from recent work by McKenna and Horswill (1997), in which the performance of subjects on four computerised tests was factor analysed. The tests were tests of close following, overtaking, gap acceptance, and a short version of McKenna's hazard perception test. The analysis showed that the four tests loaded on largely orthogonal factors; from this finding the authors argue that close following, overtaking, and gap acceptance should not be regarded as measures of general hazard perception, but rather as separate skills or behaviours.

There are other important practical considerations if a test is to be used as a licensing tool. First, there is the need for a standardised scoring system that is amenable to the imposition of a pass/fail criterion. Second, it is essential to have a number of parallel forms of the test (i.e. matched in content and difficulty). There was only one form of the TRL test, and only two of the Reading University test. Further work was therefore necessary before hazard perception testing could be used in the licensing process. The first stage of this work was a contract with the National Foundation for Educational Research (NFER), and will be described in the next section.

4 The NFER test

4.1 Objectives

The original objective of this work was the development of hazard perception tests that were intended to form one element of the new theory test for drivers and riders to be introduced in the United Kingdom in July 1996. The specification for this work was derived from the statements of required knowledge and understanding given in The 2nd European Council Directive on Driving Licences (Council Directive 91/439/EEC; Amendments made in Council Directives 96/47/EC and 97/26/EC). A decision was taken to give the topic of perception, judgement and decision-making a weighting of 20% of the total test marks. The test as a whole was to be administered by computer, allowing the hazard perception elements to have the possibility of using digitised video to present realistic moving images.

The initial goal was therefore to develop a series of hazard perception sub-tests which could be integrated into the longer theory test. However, in the early stages of the project the then Department of Transport found it necessary for technical reasons to alter the original specification for the theory test. The requirement became for a test which was paper-based and could be administered in around 30 minutes. Moving-image hazard perception items would not be part of the test that was to be introduced, but work would continue on the development of computer-based hazard perception tests that could be used for research purposes.

After discussion, the specification provided to NFER was for the production of four tests of hazard perception. The tests were to take about 15 minutes each to administer, and were to have 13 scenarios and 20 or 22 scored events in each.

4.2 Procedure

The first issue was that of the definition of hazard perception. McKenna and Crick (1994) had proposed ‘ability to identify potentially dangerous traffic situations,’ but without defining those situations. This working definition needed supplementing for individual situations with an empirical demonstration that there was general agreement on the event(s) forming the hazard. It was therefore decided to define and film a range of scenarios. These were then trialled and comparisons made between novices and experts. Situations that gave rise to differences in hazard perception latency between these groups, and also gave a range of times for participants, were determined to be appropriate for testing purposes.

The NFER project drew heavily on the work that had been carried out by McKenna and Crick (1994) at Reading University, in that the principles used by them were incorporated in the selection of scenarios for the NFER test. Using these principles as a guide, a set of 96 scenarios were written. These scenarios were specified to include six types of road covering a range of types of driving circumstances, as follows:

Rural lanes.

Suburban roads - housing, etc.

Urban roads - through shops, etc.

Residential areas.

Single carriageway main routes.

Dual carriageway main routes.

In addition, three categories of hazard were included, sometimes occurring in combination:

- entry of another object into the path of the vehicle;
- unexpected actions of the vehicle in front;
- hazardous actions of oncoming traffic.

These scenarios were then filmed from the bonnet of a car by a professional film company. The most suitable were selected for trialling and digitised so that they could be played through a computer.

Participants in the trials were sampled from people who had booked a practical driving test in the preceding six weeks, and thus were a mix of learner and early novice drivers.

The instructions for participants were shown on-screen. They were as follows:

Screen 1

> You will now be shown some video clips.
>
> Imagine you are the driver of the car.
>
> You should look out for hazards ahead.
>
> A hazard is something that a driver should keep an eye on because it could lead to an accident situation.
>
> 1

Screen 2

> Press the OK button ON THE KEYPAD as quickly as possible when you see a hazard ahead.
>
> There may be more than one hazard in a clip.
>
> You should press the OK button for each hazard.
>
> A ‘stop’ sign will appear every time you press the OK button.
>
> Press the OK button now.
>
> 2

This was then followed by a practice clip.

The most critical issue to be resolved once empirical data were available was the scoring system for the hazard perception items. This was not based on the sum of response times, as in the experimental laboratory work carried out at Reading. Instead, a preliminary scoring scheme was devised which awarded marks for each hazard event successfully detected in a specified time window.

The adoption of this procedure was important in psychometric terms, since it allowed in effect for the standardisation of the individual test items. Different hazards are likely to have different response times, because of variation in the length of time they take to develop; adding times together would give greater weight to longer times at the expense of shorter times. It also avoided the 'cross-contamination' that was possible from an overall mean latency score by using instead an empirically derived score for each item.

To obtain this score, a set procedure was adopted. When the participants took the hazard perception test, all their key presses (i.e. each time they indicated a hazard) were recorded and timed, based on the frame of the video being shown at the time. Since there are 25 frames per second, this has an accuracy of 40 milliseconds. All the participants' responses throughout the clip were then grouped into one-second time intervals and the number in each interval found. An analysis was then performed on the data and clusters of one-second intervals in which over half the participants responded were identified and examined further.

It was decided to devise a differentiated scoring scheme for each event with scores from 0 to 5. In all cases, these were based on the cluster of times, and attempted to cover all genuine responses. Different scoring intervals were used for different items and were split into 5 equal time intervals. Only the first response to an event by a participant within a time window was scored.

The same procedure was followed for all the scenarios tested. In each case, clusters of one-second intervals in which over half the participants responded were identified, a start point specified using the distribution of times and the timings from the video and finally a set of five intervals defined corresponding to scores 5 to 1. All other responses scored zero. Depending on the event, the scoring window could be as narrow as 1000ms (i.e. five bands of 200ms) or as wide as 7000ms (five bands of 1400ms). Each event had the most appropriate interval for itself, rather than a single length interval being adopted for all.

In the initial set of trials, the moving image hazard perception element had been embedded in the computerised theory test. For this reason, only seven scenarios were included in each test. However, as pointed out above, in 1995 a decision was taken to change the specification for the theory test for drivers. The test was to become paper-based and, as a consequence, the moving-image hazard perception element could not be included. However, the project was asked to continue to develop stand-alone hazard perception tests that could be used for research purposes, and to explore the feasibility of their being used in the theory test at a later date. Because the tests could be longer, a further set of trials was carried out using tests that included 12 or 14 scenarios.

Based on the results of the two sets of trials, four proposed tests were specified. In order to maximise reliability and to cover a large number of the available situations, these each included 13 scenarios, giving rise to 20 or 22 marked events. The tests would therefore have a maximum of 100 or 110 marks available. The scenarios were chosen in order to give four parallel tests as far as possible, in terms of location, type of events and road users involved. There was no attempt to ensure that the parallel events had exactly the same psychometric characteristics. It was envisaged that, for each test, a set of norms would be produced which would place the participants onto a common scale whichever test they had taken. Each test would take around 10 to 12 minutes to administer.

The items were selected for inclusion on the basis of a number of criteria. These were:

- the mean score of the event, ideally between 2 and 3 for participants;
- a high positive discrimination between participants;
- a significant difference in scores between experienced and inexperienced drivers (favouring the experienced).

Not all of these criteria could be met for each item; rather, the tests were constructed to utilise the best available items while balancing the differing road conditions and events within them. The structures of the proposed tests are shown in Table 1, which gives a brief account of each scenario and its events.

The four 'final' tests were used with participants in centres around Britain in 1995, when they were asked to attempt one of the hazard perception tests, administered by computer. The venues were London, Birmingham, Liverpool, Newcastle and Bristol. In total, 1057 participants took one of the hazard perception tests. All four tests were used at each venue, except London where only three tests were used; participants were allocated to tests randomly.

Prior to taking the test, participants were asked to complete a brief questionnaire asking for information on:

- gender;
- ethnicity;
- whether their first language was English;
- whether they had problems reading English;
- education and qualifications;
- disabilities.

In addition, participants were asked if they had recently taken their practical driving test, and whether they had passed or failed.

Some data had been collected on experienced drivers in the initial trials. However, the need, for research purposes use of the test, with drivers gaining experience meant that the tests themselves should be capable of discriminating more broadly. Hence it was necessary to obtain further data on experienced drivers. A sample was gathered from two sources. First, TRL used its database of volunteers to recruit participants to attempt the tests in sessions arranged at TRL. Other experienced drivers were recruited from the staff of NFER and took the tests at NFER. In all 172 experienced drivers attempted one or two of the four hazard perception tests.

4.3 Results

Of the 1057 participants who took one of the 'final' tests, the 16-19 years age group accounted for 41% of the sample, which is similar to the proportion for candidates

Table 1 Proposed scenarios in each test

Test 1	*Test 2*	*Test 3*	*Test 4*
A1 Horse ahead	A4 Horse ahead	A3 Jogger ahead	A1 Horse ahead Also in Test 1
A7 Parked van Cyclist	A11 Oncoming van Jogger	A9 Oncoming car	A11 Oncoming van Jogger Also in Test 2
B11 Car reverses	B5 Pedestrian on zebra crossing	B4 Pedestrian on zebra crossing	B14 Car enters on left
B6 Parked lorry Oncoming car	B8 Parked cars Oncoming car	B10 Parked lorry Oncoming car	B7 Vans parked both sides Oncoming car enters gap
C15 Motorcycle from left	C9 Cyclist from left	C10 Car from left	C15 Motorcyclist from left Also in Test 1
C12 Parked car Pedestrian enters	C13 Parked cars Pedestrian enters	C5 Pelican crossing Pedestrian crosses	C1 Pedestrian steps onto zebra crossing
D7 Car from left	D5 Car reverses from drive	D6 Cyclist from left	D12 Cyclist rides into road from pavement
D15 Ball from children	D14 Car ahead stops Pedestrian crosses	D10 Parked cars Pedestrian enters	D15 Ball from children Also in Test 1
D11 Parked cars Pedestrian enters	D8 Following car brakes suddenly	D4 Parked vans Oncoming car	D14 Car ahead stops Pedestrian crosses Also in Test 2
E7 Parked van Car enters	E3 Parked van Bus crosses in front	E6 Parked lorry Car enters	E5 Lorry obscures junction Car enters from junction
E16 Lorry parked	E9 Car from left	E11 Car from right	E9 Car from left Also in Test 2
E12 Traffic lights Car crosses ahead	E13 Traffic lights Motorcycle crosses ahead	E14 Traffic lights - following car Car crosses ahead	E13 Traffic lights Motorcycle ahead Also in Test 2
F7 Car enters DC	F11 Lorry parked Door opens	F10 Car moves to right	F11 Lorry parked. Door opens Also in Test 2

for the practical driving test. Sixty per cent of participants were female, which also matches the pattern for driving test candidates.

There was a divergence from the population as a whole in terms of the ethnic grouping of the sample. The proportions of ethnic minorities were greater than in the general population, although it was not been established how these percentages compared with the characteristics of driving test candidates. The over-representation of ethnic minority groups was probably a consequence of locating the test centres in areas of high population in order to obtain greater numbers. The number of participants who spoke a first language other than English was also high.

Forty per cent of participants were still in full-time education which reflects the high number of young people in the sample. Fifteen per cent of those participating in the trial test sessions had a motorcycle licence, and 61% had taken their practical driving test within the previous couple of months. Of this number, 38% had passed their test.

Table 2 shows the mean score, percentage of maximum, standard deviation of scores, internal consistency (reliability), and the standard error of measurement for the four tests.

As the table shows, the four tests had similar scores lying between 43 and 48% of the maximum. Test HP4 had the highest score; further examination of the data indicated that the difference may have arisen as a result of the

Table 2 Participants' results on hazard perception tests

Test	Mean score	Percentage of maximum	Standard deviation	Reliability	Standard error of measurement	Number of candidates
HP1	44.5	45	16.2	0.85	6.35	299
HP2	43.2	43	14.7	0.82	6.30	285
HP3	48.9	44	17.2	0.86	6.37	257
HP4	53.2	48	17.0	0.86	6.40	216

sample for this test having had participants of a slightly higher ability. The analyses of scores against background information have taken account of this.

The internal consistency reliabilities of the tests are high for tests of this type, which rely on a performance measure rather than being knowledge-based. In each case, it led to a standard error of measurement of around six points of raw score.

One issue that had to be addressed was the possibility that candidates could gain high scores by repeated rapid pressing or by random responses. This was investigated by establishing the distribution of numbers of presses for each scenario and cross-tabulating this with the score attained on the events. This showed that the incidence of large numbers of presses was small but that they did occur occasionally. As a result of this investigation, the scoring system was altered slightly such that a maximum number of presses were allowed. If participants exceeded this number then their score was set to zero, even if they had responded to the events within the designated times. For single event scenarios the maximum number of presses was set to 8. For multiple-event scenarios the maximum was set to 15.

In order to examine the test performance of the various subgroups, the results for the four tests were standardised onto a common scale. This was done by calculating the mean score and standard deviation for the four tests for the venues where all had been used. Since the allocation of participants was essentially random, these could be regarded as equivalent. The mean standardised score for these venues was then set at 50 and the standard deviation to 10.

The results for various groups indicate that, as in the earlier trials, scores on the hazard perception test were related to several background variables. These were:

- Age - the youngest age group scored significantly better than older participants.
- English as a first language - participants who were native English speakers obtained higher scores.
- Problems reading English - participants who had problems with English had lower scores.
- Ethnic group - white participants tended to have higher scores than black or Asian groups.
- Practical test - those who had passed the practical test had higher scores.

However for gender, there were *no* statistically significant differences between men and women.

These background variables are complex and themselves inter-related. Hence, they should not be regarded as indicating causation. There is a possibility that the differences in hazard perception scores which relate to literacy may indicate difficulty in understanding the instructions, but there is no corroborative evidence for this.

Table 3 shows the test results for the experienced drivers taking the four hazard perception tests. The mean scores were around half the marks available, somewhat higher for test HP1 and a little lower for HP4. For three of the tests the internal consistency reliability levels were high - over 0.8. However, for test HP1, the reliability coefficient was somewhat lower, probably because the spread of scores, as measured by the standard deviation, was not as great.

Table 3 Experienced drivers' results on hazard perception tests

Test	Mean score	Percentage of maximum	Standard deviation	Reliability	Standard error of measurement	Number of candidates
HP1	54.9	55	12.0	.69	6.7	53
HP2	49.3	50	14.8	.81	6.5	57
HP3	55.2	50	16.8	.85	6.5	63
HP4	51.8	47	17.2	.83	7.1	61

An important consideration for these tests is the extent to which they discriminated between inexperienced (i.e. trial participants) and experienced drivers. This comparison is shown in Table 4.

Table 4 Results for inexperienced and experienced drivers on hazard perception tests

	Inexperienced		*Experienced*			
	Mean	*sd*	*Mean*	*sd*	*Difference*	*Significance*
HP1	44.5	16.2	54.9	12.0	10.4	p<.001
HP2	43.2	14.7	49.3	14.8	6.1	p = .005
HP3	48.9	17.2	55.2	16.8	6.3	p = .009
HP4	53.2	17.0	51.8	17.2	-1.4	ns

As Table 4 shows, Tests HP1-3 showed large and significant differences. However, Test HP4 had an anomalous result in that the inexperienced drivers gained higher scores than the experienced drivers. This difference was not significant, but was nevertheless disturbing. These results were explored further in two ways. First, since the samples were not equivalent in terms of the age and sex of the drivers, an analysis of covariance was performed for the results of each test with the covariates being age and sex. This had little effect on the results, and it remained the case that the difference for Test HP1 was significant, for Test HP2 and Test HP3 approached significance and for Test HP4 was in the wrong direction, but non-significant.

The second investigation examined individual items in the tests. The results indicated that, within Tests HP1 to HP3, the majority of items had significantly higher scores for experienced drivers than for the inexperienced, and there were very few instances where the inexperienced gained higher scores than experienced drivers. In contrast, for Test HP4, for half of the items, the inexperienced had higher average scores than experienced drivers. In some cases, items which performed appropriately in Tests HP1 or HP2 did not do so for test HP4.

There is no obvious explanation for the anomalous results of Test HP4. It was the last test to be constructed and may have utilised some of the less preferred items. However, all of these had been regarded as acceptable from the results of the earlier trials. This would not explain why items which produced good results in Test HP1 or HP2 did not do so in Test HP4.

Despite the anomalous (if non-significant) results for HP4, the other three NFER tests were successful in that they demonstrated high levels of psychometric reliability, and they were able to discriminate between inexperienced and experienced drivers. Further, since they were based on the McKenna and Crick (1994) work, there is the strong probability that performance on the NFER tests could be improved through training

Thus the NFER tests met two, and by inference, could have met the third of the criteria set out earlier for acceptance as a component in the licensing entry system. However, there remains the fourth, and most important criterion, that of showing that scores on the tests are related to accident involvement.

4.4 Validation

The 'basic' NFER tests (i.e. HP1 - 3) were able to discriminate reliably between inexperienced and experienced drivers. While this provides face validity, there is also the need to demonstrate that the scores that a person achieves on a hazard perception test can be regarded as being in some way predictive of that person's subsequent accident history. A number of investigations were carried out to establish whether this was the case for the NFER tests.

The first was conducted by NFER as part of the analysis of a follow-up survey of 1092 drivers who had taken a theory test. Half of the drivers had also taken a hazard perception test during the trials reported in the previous section. Generalised linear modelling was used to assess the contribution of a number of variables, including scores on the hazard perception test, to the prediction of both accidents and self-reported near misses during the first year of driving.

The results of the modelling exercise showed that, while hazard perception score was a significant predictor of self-reported near misses, it did not reach significance as a predictor in the model for accidents. To pursue this further, some detailed analysis of the data was carried out at TRL, again using generalised linear modelling. This analysis examined each test in turn, both as a whole and at an individual event level. It also investigated whether there was any effect of using hazard latencies as opposed to hazard scores.

Of the 24 models at the whole test level, none showed that the hazard variable made a significant contribution. The results of the analysis at event level were equally discouraging, in that there seemed to be no clear correspondence between events that discriminated between inexperienced and experienced drivers, and those that discriminated between accident involved and non-accident involved drivers. Worse, there was some indication of a contrary effect, in that it was found that the correlation between the differences in mean scores between inexperienced and experienced groups of drivers and between accident involved and non-involved was a negative one.

These analyses indicated that the NFER tests would need further development before they could be used in a licensing context, and that one possible line of approach would be to examine the range and nature of the hazard items that were being presented, and to aim for a more focused test based on the more promising items. The need for further work was supported by evidence from other research groups who had used the tests. At Nottingham, Underwood and Chapman (1998) failed to find differences in hazard scores between novice and experienced drivers (though their sample sizes were only modest), and suggested that differences in hazard types were more influential than differences in the experience of individuals.

The NFER test was also used as part of a large test battery in a study of driver response to risk carried out jointly by TRL and the University of Surrey. The sample of drivers covered a wide range of ages and experience. Preliminary results presented by Field *et al.* (1997) showed no correlation between hazard scores and any measure of accident involvement. A later analysis of the whole sample reported by Grayson and Maycock (1997) also found no association between NFER scores and accident involvement, as shown in Figure 1.

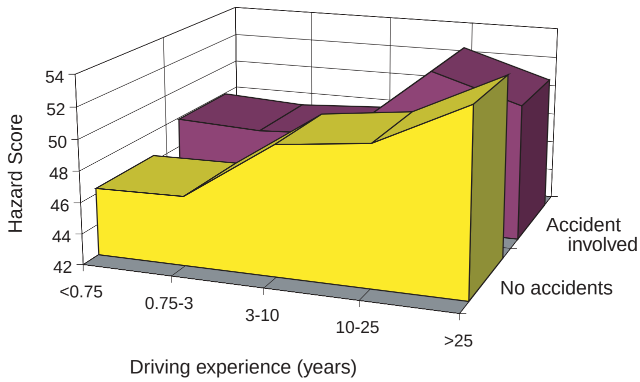

Figure 1 Hazard score, experience, and accidents

The pattern of results shown in the figure is interesting, as it shows that accident-involved drivers have *higher* test scores than the accident-free at the lower experience levels, with a difference of almost two score points. One possible explanation could be that the experience of an accident in some way sensitises the new driver to the hazards on the road, thereby improving test performance. If this is the case, it is both a confounding factor for evaluation studies, and also an incentive to produce materials so that learning can come through training rather than accidents.

In the full report by Grayson *et al.* (2002) it was confirmed that test scores did not discriminate between accident-involved and accident-free drivers, but it was also found that they were reliable indicators of observed speeds in on-road tests, and also of expert judgements of driver ability.

It was the inability to demonstrate a clear relationship with accident liability that raised serious questions about the NFER test. It is a well-designed test that is reliably measuring something, but the 'something' would not appear to be predictive of accident liability. The evidence indicates that the test may be related to broad-based behaviour measures, but that it failed on the ultimate criterion for inclusion in the licensing process. Given the encouraging results from earlier work in the field, and the persuasive theoretical arguments for focusing on skills such as hazard perception, it was recognised that further work was needed. This is reported in detail in the next section.

5 The TRL project

5.1 Objectives

Road Safety Division of the then DETR (Department for Environment, Transport and the Regions) let a contract with TRL in February 1997 to assess the likely effect on road safety of including a hazard perception test in the licensing process.

One early part of the research was to follow up the sample of inexperienced drivers who had taken the NFER hazard perception test, as described in Section 4.2. The analysis of the accident liabilities of those candidates who took the test in October/November 1995 provided evidence for a relationship between hazard perception and experience but did not confirm a relationship with accident liability. This, together with the feeling from other studies mentioned earlier that the NFER test was 'not working', led to the following programme of work:

- To review the NFER hazard perception test material to see which items discriminate best between novice and experienced drivers.
- To conduct a small evaluation trial comparing response mode.
- To draw up a 'blueprint' which describes the type of item required for hazard perception testing of learner drivers.
- To film more material and to construct some new hazard perception tests using the 'blueprint' as a guide for filming.
- To design and develop hazard perception training material.
- To evaluate the trainability of hazard perception skills in learner drivers.
- To make recommendations on the introduction of hazard perception testing into the licensing system and on the likely amount of training required.

5.2 Use of existing test material

The four hazard perception tests developed by NFER under contract to TRL were broad-based by design, in that they covered a range of driving situations and potentially hazardous scenarios. The importance of these four tests was that they provided a pool of hazard perception items and that some of these items discriminated between inexperienced and experienced drivers. However, it was also suspected that the hazard perception items within the NFER tests were too broad-based which resulted in their not distinguishing between accident-involved and accident-free drivers. That is, they covered a comprehensive range of road and road user scenarios but did not always require the scanning and anticipation skills that make up a 'good' hazard perception clip.

As described in section 4.2, the NFER instructions to subjects and definition of what constitutes a hazard were:

You will now be shown some video clips.

Imagine you are the driver of the car.

You should look out for hazards ahead.

A hazard is something that a driver should keep an eye on because it could lead to an accident situation.

Press the OK button ON THE KEYPAD as quickly as possible when you see a hazard ahead.

There may be more than one hazard in a clip.

The definition and instructions used in the TRL project used a slightly modified version:

A hazard is a situation where a driver may have to brake or take avoiding action. A safe driver will be anticipating such situations and will modify his driving such that the potential hazard is avoided.

Press the enter key when you become aware of a situation which might lead to a hazard.

Most important from the perspective of the TRL project was the need to define what is understood to be the important element(s) that low risk experienced drivers have and high risk novice drivers do not. We also know that even within a group of experienced drivers who cover similar mileage in similar conditions, that some will be 'more aware' than others, and we need to measure this 'awareness skill' in a reliable way. It is necessary to try to identify what types of hazards are spotted more quickly by the experienced driver than the novice driver. This is complicated by the fact that novice drivers are often younger than experienced drivers and will have quicker basic reactions.

Hazard perception is a skill that drivers start to acquire during their early driving career - or even before. It is a skill that is difficult to define, but it is not about reacting quickly to a sudden event, and is more about anticipation. Anticipation depends upon good scanning skills and an awareness of developing scenarios. Defining the type of driving situation that typifies such developing scenarios is necessary in order to develop hazard perception tests that measure the underlying skill of interest. The criteria for defining such scenarios have been called the 'blueprint'.

The view taken was that those drivers with 'good' hazard perception would have good scanning skills and good anticipation. 'Good' video clips were thus defined as those that would test these characteristics. It also seemed sensible that a hazard should eventually become obvious to even rank novice drivers, and that there should be a minimum of ‘clutter’ leading up to the hazard so that there should be no doubt as to what a candidate was responding. The following criteria were defined and required for hazard perception items in the test:

- develops into an ‘actual hazard’;
- anticipation is possible for experienced driver or trained novice;
- scanning ahead and/or to the side necessary;
- clear and uncluttered scenario;
- not simply dependent upon reaction time.

Existing video clips from the NFER tests were viewed independently by three ‘experts’ using the above criteria to decide if they met these criteria. A consensus view was reached and thirteen items were selected. The only hazards of interest were dynamic situations. Some hazards distinguished on the ‘experience’ criterion and some did not (or not very well). It was therefore important to try to identify the salient features of those hazard perception events that did result in a ‘good’ clip.

Novice drivers (<2 years of driving experience) and experienced drivers (>10 years of driving experience) were recruited to take a test constructed from the 13 ‘best’ NFER items. Approximately 50 male and 50 female drivers in each experience group were given two tasks: to take a hazard perception test and to complete a questionnaire about their driving history. There were 22 hazardous events within the 13 items that looked promising in that they distinguished between different experience groups of drivers. Of these, 16 had a difference in mean latency greater than 0.1sec (latency is defined as the time taken to respond from the time that the hazardous event first started). The mean overall scores based on these 16 items showed a very significant difference between experienced and novice subjects.

The following conclusions were drawn:

- The mean values of scores proved to be significantly different between novices and experts and in the expected direction - experienced drivers scored higher and had lower response latencies than did the novices.
- The score based on 16 items was predicted by age and annual mileage such that older drivers and higher mileage drivers had significantly higher scores.
- There was a highly significant difference between the average accident liabilities of the two groups of drivers. The average accident liability of novice drivers was about 0.31 accidents per year and that of experienced drivers was 0.06 accidents per year.
- For the data set as a whole, accident liabilities were significantly predicted by age and annual mileage. The coefficients of the age and annual mileage terms were consistent with those found in other accident liability studies.
- Though not significant, there was a strong indication (given the relatively small sample size) that driver response to this particular set of hazard items was predictive of accident liability in the expected sense. High scoring drivers had lower accident liabilities (and more experience). The size of this accident effect was such that those at the 5th percentile (low) score level had 2.4 times the accident liability of those at the 95th percentile (high) score level.

5.3 Test development

New video material following the blueprint description was required in order to develop new hazard perception tests. It also had to include the types of situation that had worked with the ‘best’ NFER developed items. The new film material was required to:

- be filmed from the driver’s normal eye point;
- be filmed from a left-hand drive car with the camera mounted on the right-hand side;
- include a little of the instrument panel to indicate that the view is from within a car;
- only show ‘correct’ driving, i.e. not too close, not too fast etc;
- be filmed in broadcast quality Betacam format;
- be filmed during daylight hours and fairly dry conditions.

A variety of driving situations was desirable. The ‘storyboard’ was to have a developing scenario which experienced drivers would know they should keep ‘an eye on’, but which novices would be more likely to react to later when a situation had developed.

Desirable scenarios included:

a dual carriageway:

joining slip road with traffic on slip road and on nearside lane of dual carriageway;

merging traffic from slip road driven vehicle in nearside lane;

overtaking situations with driven vehicle in nearside lane, possibly boxed in and need to anticipate suitable gap;

overtaking situations with driven vehicle in outside lane and traffic on inside lane which need to also overtake lorry or an obviously slow vehicle.

b country roads:

obstructed corners with oncoming traffic/cycles/horses/motorcycles;

restricted roads due to parked vehicles and oncoming traffic - i.e. need to slow in order to round parked obstruction due to oncoming traffic;

corners where oncoming traffic can be seen ‘through’ the corner and so driver can anticipate;

obstructed junctions with traffic emerging from the minor road and ‘spotted’ prior to joining the major road.

c urban roads:

pedestrians crossing from behind obstructions and using pedestrian/pelican crossing;

pedestrians hesitating at side of road and maybe then crossing;

parked vehicles moving off - showing indicators;

parked vehicles backing out and rear lights showing;

merging traffic from side roads;

traffic slowing for junction ahead, anticipate because; rear lights showing on vehicles ahead.

d suburban roads:

cyclists having to overtake parked vehicles;

cyclists on pavements and then joining road;

children playing (ball games, skate boarding, street roller hockey, etc.);

parked vehicles and on-coming traffic requiring anticipation and slowing down to avoid meeting adjacent to parked vehicle;

ice cream vans and pedestrians/children moving into road.

Filming of new hazard situations was undertaken by the TRL film unit using a left-hand drive car and filming from the driver's viewpoint. The field of view was chosen to give good definition for the image in the centre of the picture but with a reasonable view of the whole driving scene. Some panning of the camera was used on corners and at roundabouts.

Two main types of filming were undertaken. The first was opportunistic on a variety of road types, and yielded some useful video footage that was included in the subsequent trials. However, it was not very effective as a process, especially since the film car was driven in a safe way which reduced the likelihood of hazard involvement. The second type of filming used staged events, and this was generally more efficient. The staged events were based on items in the re-visited NFER test that had discriminated on the experience criterion. However, it was decided not to include events with children because of the risks associated. The DSA training unit provided intensive and helpful assistance in staging a number of events over a three-day filming period. The use of highly-trained drivers in this way was particularly beneficial.

TRL experts and the project steering group selected suitable items for trialling. They applied the blueprint criteria together with their own experience and knowledge of the type of hazard that may be appropriate. As a result of the selection process some 54 suitable video clips were selected. These were mainly about 60 seconds in duration. The 54 items were split into two sets, referred to as tests A and B for the purposes of the trial. One item was selected as an example item in order to give subjects an indication of what was expected.

During filming of staged events, several 'takes' were made of the same scenario. The timing of the event is critical to how well it works, and so it was desirable to film several versions of the same hazard event. In practice, there were several equally good examples of some scenarios and more than one was tried. In order to determine the psychometric properties of the newly filmed hazard perception material, a trial was organized with drivers of different experience levels, including a group of learners. This was essential in order to determine if items could distinguish between drivers of differing experience.

The trial required volunteers to take two hazard perception tests, each with about 27 video clips. Subjects took either test A or test B first, then completed a questionnaire, and then took either test B or A. The order of the first test taken was counter-balanced. Subjects were volunteers and were either learners, novice drivers with less than 2 years experience or experienced drivers who had been driving for at least 10 years. The total target sample was 450 subjects, with 150 from each group. The trial was held at TRL.

Subjects were first given on-screen instructions to read (there was also a paper copy by the computer), and then saw an example video clip. After the example they were told what the hazard was and told that they were expected to respond as soon as they had realised the potentially hazardous situation was developing. They then saw the trial clips in a pre-determined order, shown as a continuous series of clips without a pause between them. Subjects responded by pressing a keyboard key marked with a red triangle. Every response was stored for subsequent analysis.

The scoring method was the same as that developed previously by NFER. The distribution of response times is first generated; this shows where subjects are responding and therefore indicates the time of a potential hazard (or hazards) in the video clip. Looking at the video clip where subjects were responding identified the hazardous event that led to a reaction. This will usually correspond to the staged hazard within the video clip, and confirms that candidates are responding as expected. The scoring window is determined from the earliest point where the hazardous event started to a point where is becomes obvious. Scoring windows can be determined without the empirical data, but if no candidates were responding to the hazard then this would suggest it was a very poor hazardous event that was unlikely to result in a useful item for scoring. Using empirical data and expert judgement to set the scoring window ensures that a useful item is identified that also has good face validity.

The score given to each hazardous event depends upon the first response within a 'time-window'. The earlier within the window that the candidate responds, the higher the score. In practice a 5, 4, 3, 2, 1 score was allocated by dividing the window equally in five sub-time-windows. Candidates who failed to respond in the time-window were given a zero score.

An example of the procedure is shown below. This particular hazard item is in a suburban setting with traffic calming. A cyclist emerges from a right turn and is obscured behind a van, but could have been spotted prior to being obscured by the van. The driver's view of the cyclist just before he disappears behind the van and just after emerging are shown in Figures 2 and 3.

Figure 4 shows the response distribution for this item. The response pattern indicates a cluster of responses around 26 seconds into the clip which was the cyclist. This was confirmed from examining the video. The setting of the scoring window was empirical, in that it depended upon the response distribution. The event of the emerging cyclist was identified on the video, and the time when this event could first be responded to noted, i.e. the point when

Figure 2 Cyclist emerging from side road, about to go behind van

Figure 3 Cyclist emerges from behind the van, into path of oncoming vehicle

the cyclist could first be seen. The length of the scoring window depended on the observed response distribution and the particular developing hazard.

Sixty-five possible events were identified and scored using the methodology described above using empirically derived time-windows. The mean scores for each group of drivers were then compared for every event. A one-way analysis of variance was used to check the statistical significance of group mean score differences. A multiple range test (Student-Newman-Keuls) was applied if there were any overall group differences.

The results for the item above are presented in Table 5.

Table 5 Results of statistical analysis for the item in Figure 4

Group	*Sample*	*Mean*	*Standard error*	*Significant differences*
Learner	149	1.85	0.09	The experienced group is different from the other two groups
Novice	147	2.06	0.10	
Experienced	148	2.50	0.09	

It can be seen from the results in Table 5 that the mean score values in this item monotonically increased with level of experience. This is essential if an item is working in the way required. However, the mean values between groups were not always statistically significantly different. 'Perfect' items would be those that distinguish between high and low accident risk drivers. However, in this context 'perfect' items are described as those that distinguish between experienced drivers and novice drivers, between experienced drivers and learners and between novices and learners. Driver experience is thus being used as a proxy for accident liability.

There were 16 'perfect' events, (within 13 video clips) which could distinguish between all three groups of drivers

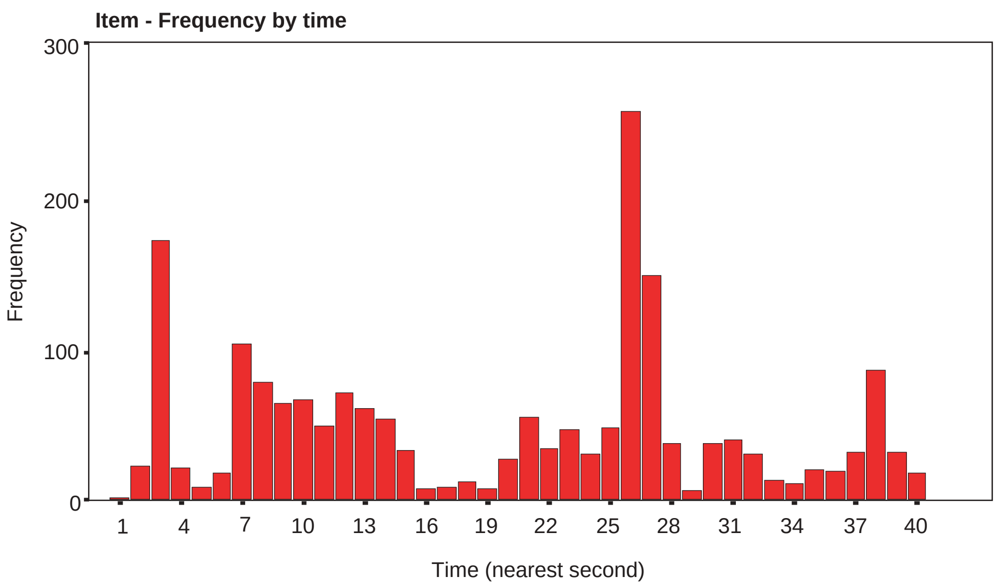

Figure 4 Response distribution for sample item

and 22 events which distinguished between experienced drivers and the others, but not between learners and novices (albeit the scores were monotonically increasing). These 38 events provided a sufficient number from which to construct two new tests. The following criteria were applied:

- the item should, at least, distinguish between the experienced drivers and other drivers;
- the test should take about 15 minutes to complete;
- the test should be internally consistent – as measured by Cronbach's Alpha;
- the two tests should be parallel (i.e. should look similar in terms of the item mix, and be equivalent in difficulty, the number of items, internal consistency, etc.);
- replicates of the same staged events should not be in the same test.

Items were selected from the 'best' 38 as identified from the individual hazard event analysis. All of the 'perfect' items were selected and as many as necessary from the remaining 22 in order to meet the above criteria. The 'new' tests were designated X and Y.

The characteristics of the new tests are shown in Table 6.

Table 6 Test X and Y characteristics

Test	*Statistic*	*Learners*	*Novice (<2yrs)*	*Experienced (>10yrs)*
X	Mean	24.44	29.32	36.50
	Standard error	0.785	0.819	0.890
	Sample size	157	152	153
	Reliability = 0.81			
Y	Mean	22.06	27.91	36.54
	Standard error	0.864	0.899	0.917
	Sample size	157	152	153
	Reliability = 0.84			

The mean values for experienced drivers are very similar, but test Y seems a little more difficult for learner and novice drivers. The internal consistency (reliability) was above 0.8 for both tests. The mean scores are monotonically increasing with experience, and each group mean is different from the others, with a high statistical significance.

Subjects' scores were calculated from their item responses. These scores were analysed with the questionnaire data.

Subjects had also completed a questionnaire between taking the two hazard perception tests being tried out. The questionnaire asked for personal details as well as the subjects' driving experience, offences, accidents and attitudes. There were two versions, one for learners and one for novice and experienced drivers.

The questionnaire data and the test X and test Y scores were analysed using the usual accident modelling approach, using accidents as reported for the past 3 years of driving (or for as many years as they had been driving if less than 3 years). The accident models could only sensibly include the novice and experienced drivers. This was because the learners had not passed their tests, by definition, and some may have not even driven a vehicle. Those learner drivers who had driven would also have been unlikely to have reported any accidents, as accidents while learning are very rare. The accident models follow the well-established functional form used in previous TRL studies. These include age (experience works nearly as well), and the number of miles driven. The contribution in explaining residual accident risk that can be made by the subjects' hazard perception score was then included in the model, but did not quite reach statistical significance.

The original TRL work on hazard perception showed a correlation between hazard perception and accident liability. Other previously reported accident models have found small but significant explanatory power from hazard perception scores once age and mileage have been included in the model. The analysis looked at the predictive power of age and mileage in explaining hazard perception scores. The hazard perception tests are constructed from items that discriminate on the experience axis, and experience is highly correlated with age and related to mileage. Hence, it is not surprising that a multiple-linear equation with age and mileage as independent variables can predict hazard perception scores.

It is also not surprising that when age and mileage are already variables in an accident model, a hazard perception variable may not reach statistical significance (even though the coefficient of the hazard perception term indicates that the higher the hazard perception score the lower the accident liability.) What we cannot tell is what are the causal factors related to accident liability. We know that accident liability reduces with age and mileage and that hazard perception increases with age and mileage. We do not usually have measures of hazard perception when looking at accident data, and so age and mileage have been used as proxy measures for driver skill in avoiding accidents. This skill may be better measured by a hazard perception score. Is the causal factor for accident risk experience and driving exposure, or is it hazard perception? The factors are highly correlated.

5.4 Training material

A prime objective of introducing a hazard perception test is to improve the driving ability of most new drivers to a level that will produce a significant reduction in accident liability. It is assumed that most of the benefit will derive from the training that drivers would take in order to pass the test. One task in this project required the development of a hazard perception training programme.

The first stage was to review earlier work on hazard perception training, as well as driver training materials in general. Discussions with driver training organisations and others suggested that the fundamental behaviours that are thought to be associated with good hazard perception skills are:

- keeping safe driving distances;
- looking well ahead;
- driving at appropriate speed.

It was decided that the training material should have the following characteristics:

- candidates receive a minimum of one session and a maximum of three sessions of training;
- each session lasts a maximum of one hour; candidates receive approximately one session per week;
- a session of one hour duration will be based upon approximately 20 minutes of video material;
- candidates would receive training at the stage when they have mastered basic control skills - the timing of this will depend on the individual, but it is likely to be after about 10 hours of driving lessons;
- the training will be delivered in a 'classroom' type setting in small groups, and will be interspersed with their regular on-road tuition.

The content of the package should facilitate a learning process in which students are required to be *highly active*. One method which has proved effective is to 'freeze' the video, and ask students to predict what might happen next, what they would do next, etc. (this is followed by continuing the video to see what in fact does occur).

A training package was developed by the DSA Training Establishment at Cardington. The package consists of three modules, one for basic training, and two for more advanced training. Each module was designed to last about 60 minutes and to be delivered to groups of between 5 and 10 students.

The modules use a combination of video, video-freeze techniques, some animation techniques to illustrate particular points, and work-cards for trainees to complete. The video illustrates on-road situations that show possible hazardous situations, and freeze-frame techniques are used to help facilitate an interaction with the trainer by asking subjects to identify situations. Work-cards are used in a similar way. Group discussion is encouraged and this can be facilitated with flip-charts. Subjects are given a workbook at the end of the training; this summarizes the training session and could be used as an *aide-memoire*. The more advanced training modules focus more on dynamic situations, building on the initial training module experience. The purpose of the training modules is to improve learner driver awareness of developing road scenarios, and so to make them more like experienced drivers in this respect.

5.5 Training trial

The procedure for evaluating the trainability of hazard perception skills was to recruit learner drivers, test them, train some of them, and re-test all of them. The test was of their hazard perception skills as measured by their performance on the hazard perception tests developed as part of this project.

Subjects were recruited at the time they attended a Theory Test centre to take their Theory Test. They were thus at a stage in obtaining a driving licence where they felt ready to take the Theory Test. This is also likely to be where a test of hazard perception might be introduced into the licensing system, and although they would not have received any hazard perception training, they were at the most appropriate stage for this trial. It is also quite possible that some subjects may not have even driven a vehicle, but most would have done so, and this was one of the questions included in a questionnaire.

Subjects took one of two hazard perception tests, designated test X or test Y. These tests are parallel versions, and if subjects took test X as their first test they took test Y as their second test, and vice versa. Subjects were allocated randomly to either test X or Y on the first test session. This ensured that there were no learning effects that were not controlled for by test order and presentation.

Subjects were recruited to one of three groups, A, B, or C:

Group A took one hazard perception test on recruitment in February and the other hazard perception test in April.

Group B took one hazard perception test on recruitment in February and the other hazard perception test in April, and received 1 hour of basic hazard perception training in March.

Group C took one hazard perception test on recruitment in February and the other hazard perception test in April, and received 1 hour of basic hazard perception training plus 2 hours of more advanced training in March.

Group A was a control group. This was required because during the period between taking the two tests all subjects would probably have still been receiving driver training. It was thus necessary to control for a possible improvement in hazard perception skills due to additional driver training and experience, and also because the taking of a hazard perception test may itself have influenced their score.

The analysis looked at the gains in hazard perception scores over a two month period. The gains in each group will have been influenced by the extra training and experience gained during this two month period. The test taken first or second was balanced across each group, and so any test effect was controlled for. The gain due to either 1 hour of basic training on hazard perception, or the gain in receiving 1 hour of basic training plus 2 hours of more advanced training can be calculated relative to the gain seen in the control group. This net gain indicates the benefits attributable to the hazard perception training received.

Candidates were given a three part questionnaire when they were first recruited. The first part asked for some basic information about their driver training and experience at that time. The questionnaire was kept at the centres, and when candidates came for re-testing they gave information about the interim period. The questionnaires were then returned to TRL, which confirmed that they had returned for re-test and so could receive a payment towards their expenses for helping with the trial.

It was important to obtain a good coverage of test centres to minimise regional biases in the sample of learner drivers recruited. Advice was taken from Theory Test staff and the following ten centres were selected: Brighton, Harlow, Middlesborough, Milton Keynes, Slough, Southampton, Stirling, Stoke-on-Trent, Sutton Coldfield, and Swansea.

The approximate number of candidates booking tests were known for these centres and so the sample sizes required (250 for each group), were known to be available for recruitment. In practice it was estimated that there could be a 20% dropout rate, and so it was hoped to recruit 300 in each group – 900 candidates in total. In the event, a total of 861 candidates agreed to take part in the trial.

The average drop-out rate for the basic training was about 26%, calculated as the percentage of those that did not turn up for their basic training. It excludes those who had previously cancelled. Some candidates were unable to make the training sessions they had booked, and were either re-scheduled to another day or were re-allocated to a different group. In part, this is why there are a higher number in the control group. The actual response rate was lower than anticipated, but acceptable numbers were achieved for analytical purposes.

Table 7 shows the numbers recruited (including those re-allocated because they missed a training session), the numbers where test scores could be matched and the numbers where both the test scores and questionnaires could be matched. There were a few subjects who were originally recruited to one of the three groups but failed to attend the scheduled training for some reason; most of these came back for re-testing as indicated in Table 7. Most of the analysis has been based on the 520 whose test scores were matched and who were in groups A, B, or C.

Table 7 Number recruited by group

Group	*Recruited*	*HP Test score matched*	*HP Test and questionnaire matched*
A – no training	315	191 (60.6%)	188 (59.7%)
B – basic training	273	166 (60.8%)	160 (58.6%)
C – basic + advanced training	252	163 (64.7%)	160 (63.5%)
Others*	21	14 (66.7%)	14 (66.7%)
Total	861	534 (62.0%)	522 (60.6%)

** Those who did not complete training*

The linking mechanism for all candidates was the unique theory test booking reference number. Allocation of candidates to training groups (A: no training, B: basic training, and C: basic + more advanced training), was arranged at the test centres when the candidate was recruited. The centre invigilator was instructed to obtain a mix of candidates per group, although some candidates were constrained by circumstances. The following tables illustrate that there were some differences between candidate profiles within each group, but nothing that appeared important.

Table 8 shows the numbers of candidates who completed the trial for each group by age. The Chi-squared test indicated that the distributions of age were not statistically different between the groups.

Table 9 shows the numbers of candidates who completed the trial for each group by gender. The Chi-squared tests indicated that the distributions of gender were not statistically different between the groups.

Candidates completed a questionnaire in two stages. When recruited, candidates took their theory test and then after a short interval the first hazard perception test. They completed part 1 of the questionnaire at this time. This gave information about their instruction and driving history. When they returned to take their second test, up to 8 weeks later, they completed part 2 and part 3. Part 2 of the questionnaire is very similar to part 1 and asked about their instruction and driving history as well as their success in passing the driver theory test and the practical driving test. Table 10 shows the mean values for some of these questions, including age and number of hours of driving either with a friend or under instruction. The mean values for each group are very similar, although those in group B have had more hours of driving with friends but fewer hours of instruction between taking the two hazard perception tests.

Table 11 shows the percentages relating to the dichotomous response questions. The patterns of responses are fairly similar for each group of candidates, although the percentage who had passed the practical driving test before taking the second hazard perception test are somewhat lower for group B than groups A or C. Interestingly, group B have the highest pass rate for the theory test.

Table 12 shows the mean and standard error of the hazard perception scores, or derived measures, for each group. The experimental design was that candidates were allocated to take either hazard perception test X or test Y on their first test. If they took test X as their first test they took test Y as their second tests, and vice versa. The two tests were reasonably well matched, but were not exactly parallel. This can clearly be seen in Table 12, where test X seems to be about 6 score points easier. This does not

Table 8 Candidates who completed trial by age group and training group

	A: No training		*B: Basic training*		*C: Basic + Advanced*		*Total*	
Age	*Count*	*Column %*	*Count*	*Column %*	*Count*	*Column %*	*Count*	*Column %*
17-18	113	59%	92	55%	91	56%	296	57%
18-19	17	9%	14	8%	16	10%	47	9%
19-20	4	2%	9	5%	6	4%	19	4%
20-25	25	13%	18	11%	18	11%	61	12%
25-30	12	6%	8	5%	9	6%	29	6%
30+	16	8%	17	10%	21	13%	54	10%
N/K	4	2%	8	5%	2	1%	14	3%
Total	191	100%	166	100%	163	100%	520	100%

Table 9 Recruited candidates who completed trial by gender and training group

	A: No training		*B: Basic training*		*C: Basic + Advanced*		*Total*	
Gender	*Count*	*Column %*	*Count*	*Column %*	*Count*	*Column %*	*Count*	*Column %*
Male	91	48%	90	54%	86	53%	267	51%
Female	100	52%	76	46%	77	47%	253	49%
Total	191	100%	166	100%	163	100%	520	100%

Table 10 Mean values for candidates who completed trial

	A: No training		*B: Basic training*		*C: Basic + Advanced*	
Variable description	*Mean*	*se*	*Mean*	*se*	*Mean*	*se*
Age at 2nd February 1999	20.6	0.60	21.4	0.70	21.6	0.67
Hours of driving with friends (at 1st test)	13.0	1.53	14.2	1.82	13.0	1.88
Hours of driving with friends (at 2nd test)	14.9	1.60	18.5	2.97	15.1	1.74
Hours of lessons (at 1st test)	17.5	0.99	18.6	1.05	17.5	0.83
Hours of lessons (at 2nd test)	25.4	1.59	25.5	1.37	26.0	2.64
Number of theory tests taken (at 1st test)	0.42	0.07	0.35	0.06	0.45	0.06
Number of theory tests taken (at 2nd test)	1.46	0.07	1.39	0.06	1.45	0.08

se – standard error of the mean value

Table 11 Percentage of candidates who completed trial

	A: No training		*B: Basic training*		*C: Basic + Advanced*	
Variable description	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Had driving lessons before 1st test	97%	3%	96%	4%	93%	7%
Had driving lessons before 2nd test	97%	3%	98%	2%	97%	3%
Driven with friends before 1st test	58%	42%	55%	45%	53%	47%
Driven with friends before 2nd test	66%	34%	66%	34%	61%	39%
Passed the theory test (after 1st test)	84%	16%	89%	11%	87%	13%
Passed the practical test (after 1st test)	21%	79%	16%	84%	22%	78%

Table 12 Hazard perception test scores

		A: No training		*B: Basic training*		*C: Basic + Advanced*	
Variable description		*Mean*	*se*	*Mean*	*se*	*Mean*	*se*
First test taken	Test X	30.33	1.13	30.94	1.29	31.20	1.11
	Test Y	24.71	1.13	24.83	1.25	22.82	1.14
Second test taken	Test X	31.26	1.01	34.01	1.31	39.79	1.01
	Test Y	26.72	1.08	30.90	1.04	38.88	1.22
Unadjusted score increase between tests		1.49	0.72	4.40	0.84	12.07	0.98
Adjusted score increase between tests, (i.e. controlling for test difficulty)		1.46	0.62	4.63	0.77	12.43	0.92
Standardised score first test, N~(50,10)		50.01	0.73	50.34	0.82	49.58	0.73
Standardised score second test, N~(50,10)		51.36	0.68	54.61	0.77	61.03	0.77
Standardised score increase between tests, N~(50,10)		1.35	0.57	4.27	0.71	11.45	0.85

matter provided there is a balance between the order the tests were taken, which there was by design. However, it is easier to interpret the gains in hazard perception score as if the first tests were identical, as explained below.

The increase in scores between the two test occasions has been analysed in several ways. Table 12 shows the mean scores by test taken and time taken. The unadjusted score increase is simply calculated as the averaged difference between the hazard perception scores of first and second tests taken. The adjusted score increase is calculated by deducting 6.38 score points from each test X score before calculating the difference in test scores (this was the mean difference across all groups between tests X and Y for those taking the first test). The adjusted score increase is thus controlling for the average difference in test difficulty and any imbalance in the design.

A further approach was to standardise both test X and test Y to a distribution with a mean of 50 and standard deviation of 10. The mean standardised score values are given in Table 12, together with the mean difference. The averaged difference between standardised scores is thus controlled for test mean and for any difference in standard deviation.

Group means of the gains in score between the two test occasions were compared. The mean values are shown in Table 12. The results from a one-way analysis of variance found that, whatever measure (unadjusted, adjusted or standardised) for hazard perception test score was used, there was a statistically significant gain from the training. The mean scores for group B were greater than those for group A, the mean scores for group C were greater than those for group A. Hence, the score differences show that basic training does result in an increase in hazard perception score, and additional advanced training results in a further increase. This is clearly illustrated in Figure 5.

It is possible that some factors other than the training may be influencing the gain in scores, for example the gender of the candidate, whether they pass their practical driving test prior to taking the second hazard perception test, and even the centre where they took the hazard perception test (especially since two of the centres had the training given by experienced DSA driver examination staff). These factors were analysed using the GLM (general linear model) module of the SAS (Statistical Analysis System) package, where a nested-model analysis of variance was run. The analyses showed that there were some effects due to gender (p=0.05) and that training centre effects were also close to being significant (p=0.08), but there were no effects due to passing the practical driving test prior to taking the second hazard perception test.

Table 13 shows the mean values for the averaged adjusted score increase across all groups. It shows clearly that, on average, females gained 2.26 more score points than males, and this was statistically significant.

Table 13 Average gains in adjusted scores by gender

Factor		*Mean gain in score*	*Sample size*
Gender	Female	7.01	267
	Male	4.75	253

Table 14 shows the mean adjusted hazard perception scores for males and females within each training group. The difference between males and females is fairly consistent within groups, although there is a suggestion (i.e. an extra score point), that females may have benefited more than males from the more advanced training. There was no statistically significant interaction between gender and training group.

The analysis demonstrates that training does result in an increase in scores the hazard perception test. There is also evidence that females perform better than males, see Table 13. Females may also benefit slightly more from the training since their net gain was 3.12 points with basic training and 11.35 points with the more advanced training, as compared to males who had gains of 3.00 and 10.35 respectively. However, the difference in the gains between males and females is not statistically significant.

The net gain due to training is estimated as 3.17 (se=0.99) score points for the basic training, and 10.97

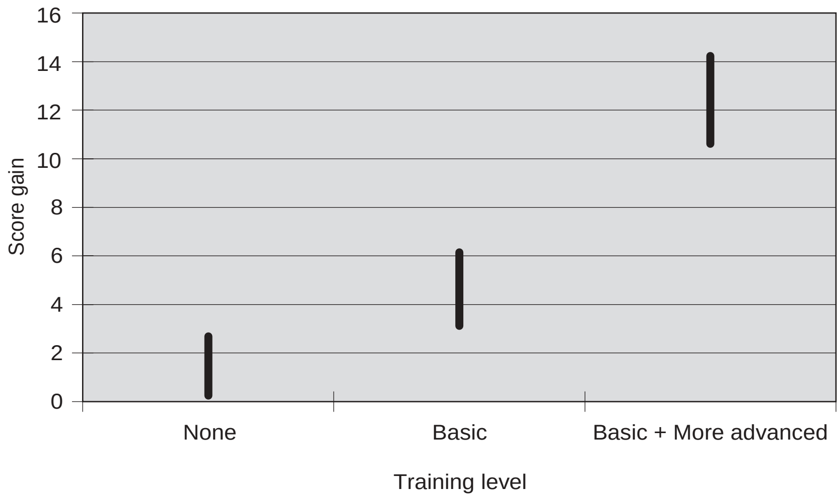

Figure 5 Effect of basic and advanced training on hazard perception scores (mean difference and 95% confidence intervals

Table 14 Average gains in adjusted scores by gender and training group

	A: No training			*B: Basic training*			*C: Basic + More advanced training*		
	Mean	*se*	*Sample*	*Mean*	*se*	*Sample*	*Mean*	*se*	*Sample*
Males	0.70	0.82	100	3.70	1.14	76	11.05	1.39	77
Females	2.30	0.93	91	5.42	1.05	90	13.65	1.23	86
Total	1.46	0.62	191	4.63	0.77	166	12.43	0.92	163

(se=1.11) for the basic + more advanced training. This takes into account the small gain in test scores for the control group. The gains in test score can clearly be seen in Figure 5. It would appear that one hour of training is worth just over 3 hazard perception score points.

6 Implementation

The *Road Safety Strategy* ‘Tomorrow’s roads: safer for everyone’ (DETR, 2000) contained a commitment to introduce a hazard perception test into the driving theory test. Recognising that poor hazard awareness skills are associated with increased accident liability, the intention is that the enhanced theory test will contribute to the road casualty reduction targets.

The *Road Safety Strategy* document identified a date of autumn 2002 for implementation, and work on the implementation process started in 2000. The prime objectives of the TRL project have been to demonstrate that hazard perception tests could be produced that were psychometrically reliable, and on which scores could be improved by training. Once these objectives had been achieved, a necessary first step was to produce a number of equivalent versions of a hazard perception test. This was to ensure that candidates who had to re-take a test would not be presented with the same material again. It also ensured that a wide spectrum of road conditions and road users would be covered in the tests.

Initially a dedicated DSA film unit generated over 650 hazard perception clips. These were reviewed by DSA senior examiners and by TRL to ensure that the ‘blueprint’ guidelines were being followed. Some 390 film clips were selected and went forward for trials. They were given to samples of 4,000 learners, 4,000 drivers with 3-24 months post-test driving experience, and 4,000 with more than 10 years experience. Each clip was seen by about 150 subjects from each of the three groups.

Those items that discriminated between the three experience levels were used to construct 20 equivalent tests. Each test contained a mix of items, such that every test included at least one clip that included an LGV/PCV, a two-wheeled vehicle, pedestrians, adverse weather, rural scenes, suburban scenes, and either a motorway or dual carriageway.

These 20 test versions were then the subject of calibration trials with 150 learner drivers each. Two test versions were also taken by samples of LGV drivers and by ADIs. The psychometric properties of the items and tests were then analysed in order to ensure that the test versions were as equivalent as possible, i.e. they were of equal overall difficulty and had similar internal consistency. The LGV/PCV and ADI trials were used to determine how well these more experienced drivers would perform. This information, together with the data from the initial trials, was used to help to establish pass mark levels for test candidates. This work will be reported in more detail by DSA in due course.

7 Summary and Conclusions

Hazard perception has had a long history as a topic for investigation by research workers, but it is only in relatively recent times that serious consideration has been given to the possibility of using it as a component of the driver licensing system. This report has reviewed that early research history, and then gone on to describe the work that has been carried out to arrive at the present position, where a hazard perception test will become part of the Theory Test that must be passed before new drivers can undertake the practical driving test.

The main conclusions may be summarised as follows:

- The literature review provides strong support for hazard perception testing and training. The theoretical basis is sound, there is good evidence that it is a skill that improves with experience, it is capable of being improved through training, and there are indications that poor hazard perception skills are associated with elevated accident risk.
- The NFER study set out to demonstrate that it was possible to produce tests that were sufficiently reliable in psychometric terms to be used in the licensing process. In this they were successful, in that four tests were produced with high levels of reliability, three of which distinguished between inexperienced and experienced drivers.
- However, a number of studies found that the NFER tests were not able to demonstrate an association with accident liability. To investigate this and other topics, a further programme of work was undertaken by TRL.
- One of the first tasks in the TRL project was to evaluate the hazard perception items that discriminated better than others on a driver experience scale. The ‘best’ items were those that required good scanning practices and an ability to anticipate potentially hazardous situations. The sub-set of NFER items selected with these characteristics was able to discriminate on the driver experience scale and appeared to have some association with retrospective accident liability.

- A broad-based 'blueprint' was developed to define the type of hazard perception items required. Further items were filmed using staged situations as well as opportunistic filming, and produced enough usable items to evaluate in a trial of different driver experience groups. The results of the trial demonstrated that there were enough items that had the necessary psychometric characteristics to allow for the construction of new tests.
- While this test development was being undertaken, a specification for suitable hazard perception training material was being developed. This involved reviewing existing driver training material as well as discussions with the driver training industry. The resultant specification was given to DSA who then created a training package. This training package contained three modules – basic, plus two more advanced levels of training. Each module took about an hour to administer, and was group-based training using video material with a structured schedule, which could be delivered by people who are not professional driver trainers.
- The final task in this research was to run a trial of learners who were ready for their theory test, and to test their hazard perception skills before and after training. Only one hazard perception training package was evaluated, but using this package - developed specifically for this project - showed that 3 hours of training resulted in learner drivers obtaining similar average hazard perception scores to those of experienced drivers. The conclusion, therefore, is that hazard perception skills can be trained in learner drivers. Improving new drivers' hazard perception skills should have a beneficial effect on their accident liability, particularly in the first 2 or 3 years of driving which is when they would normally be gaining these skills.
- In order to encourage new drivers to take hazard perception training they need motivation, and the strongest motivation is a hurdle in the licensing process in the form of a hazard perception test. Any hazard perception test introduced into the licensing system must be psychometrically sound and have face validity. In other words those taking the test must see the value of the test and the test must be a reliable instrument.
- The *Road Safety Strategy* document (DETR, 2000) contained a commitment to introduce a hazard perception test into the driving theory test. Work on the implementation process has led to the construction of 20 equivalent tests covering a wide range of road conditions and road users, and incorporating only those items that have been shown to discriminate between experienced, inexperienced, and learner drivers. The tests have been calibrated, and pass marks for test candidates have been identified.

8 References

Currie L (1969). *The perception of danger in a simulated driving task*. Ergonomics, 12, 841-849.

Department of the Environment, Transport and the Regions (DETR) (2000). Tomorrow's roads: safer for everyone. London: Department of the Environment, Transport and the Regions.

Field D T, Groeger J A and Hammond S M (1997). *Computerised assessment of drivers' skills*. In G B Grayson (Ed.) *Behavioural Research in Road Safety VII.* Crowthorne: TRL Limited.

Forsyth E, Maycock G and Sexton B (1995). *Cohort study of learner and novice drivers: Part 3, Accidents, offences and driving experience in the first three years of driving*. Project Report PR111. Crowthorne: TRL Limited.

Goldstein L G (1972). *Youthful drivers as a special safety problem*. Accident Analysis and Prevention, 4, 153-189.

Grayson G B (1991). *Driver behaviour.* Proceedings of the Safety 91 Conference. Crowthorne: TRL Limited.

Grayson G B and Maycock G (1997). *What do drivers learn through experience - and can it be taught?* Paper presented to NHTSA Symposium on Risk Taking Behavior and Traffic Safety. Cape Cod, October 1997.

Grayson G B, Maycock G, Groeger J A, Hammond S M and Field D T (2002). *Risk, hazard perception and perceived control*. TRL Report TRL560. Crowthorne: TRL Limited. *(In production)*

Hull M A and Christie R J (1993). *The hazard perception test: the Geelong trial and future developments*. VicRoads Report GR 93-13

Irving A and Jones W (1992). *Methods for testing impairment due to drugs*. European Journal of Clinical Pharmacology, 43, 61-66.

Maycock G, Lockwood C R and Lester J F (1991). *The accident liability of car drivers*. Research Report RR315. Crowthorne: TRL Limited.

Mayhew D R and Simpson H M (1990). *New to the Road.* Ottawa: Traffic Injury Research Foundation of Canada.

McKenna F P and Crick J L (1991). *Experience and expertise in hazard perception*. In G B Grayson and J F Lester (Eds.) *Behavioural Research in Road Safety*. Crowthorne: TRL Limited.

McKenna F P and Crick J L (1994). *Hazard perception in drivers: a methodology for testing and training*. Contractor Report CR313. Crowthorne: TRL Limited.

McKenna F P and Crick J L (1997). *Developments in hazard perception*. TRL Report TRL297. Crowthorne: TRL Limited.

McKenna F P and Horswill M S (1997). *Differing conceptions of hazard perception*. In G.B. Grayson (Ed.) *Behavioural Research in Road Safety VII.* Crowthorne: TRL Limited.

McKenna F P and Horswill M S (1999). *Hazard perception and its relevance for driver licensing*. IATSS Research, 23, 36-41.

Mills K L, Hall R D, McDonald M and Rolls G (1998). *The effects of hazard perception training on the development of novice driver skills.* Road Safety Report No.4. London: Department of the Environment, Transport and the Regions.

OECD (1975). Young Driver Accidents. Paris: OECD.

Pelz D C and Krupat E (1974). *Caution profile and driving record of undergraduate males.* Accident Analysis and Prevention, 6, 45-58.

Quimby A R, Maycock G, Carter L D, Dixon R and Wall J G (1986). *Perceptual abilities of accident involved drivers.* Research Report RR27. Crowthorne: TRL Limited.

Quimby A R and Watts G R (1981). *Human factors and driving performance.* Laboratory Report LR1004. Crowthorne: TRL Limited.

Underwood G and Chapman P A (1998). *Eye movements and hazard perception ability*. In G.B. Grayson (Ed.) *Behavioural Research in Road Safety VIII.* Crowthorne: TRL Limited.

Abstract

This report describes the recent history of hazard perception testing using computer-based moving image technology, and sets out the background to the decision to incorporate a hazard perception component into the Theory Test that learner drivers in the UK must pass before taking a practical driving test. Previous research has shown that hazard perception improves with driving experience, is related to accident liability, and is capable of improvement through training. For a test to be suitable for inclusion in the licensing process, it must be psychometrically sound, have a standardised scoring system, and be available in a number of parallel forms. The main body of the report describes the development work that was undertaken to meet these criteria, and also reports on trials showing that the mean scores of learner drivers could be increased to those of experienced drivers by a short programme of focused training.

Related publications

TRL560 *Risk hazard perception and perceived control* by G B Grayson, G Maycock, J A Groeger, S M Hammond and D T Field. 2002 *(In production)*

TRL297 *Developments in hazard perception* by F McKenna and J Crick. 1998 (price £25, code E)

PR111 *Cohort study of learner and novice drivers: Part 3 Accidents, offences and driving experience in the first three years of driving* by E Forsyth, G Maycock and B Sexton. (price £35, code J)

CR313 *Hazard perception in drivers: a methodology for testing and training* by F P McKenna and J L Crick. 1994 (price £35, code H)

RR315 *The accident liability of car drivers* by G Maycock, C R Lockwood and J F Lester. 1991 (price £20, code C)

RR27 *Perceptual abilities of accident involved drivers* by A R Quimby, G Maycock, I D Carter, R Dixon and J G Wall. 1986 (price £20, code C)

Prices current at November 2002

For further details of these and all other TRL publications, telephone Publication Sales on 01344 770783, or visit TRL on the Internet at www.trl.co.uk.