Implementation of Victoria’s new Hazard Perception Test

John Catchpole#, Peter Congdon* and Corinne Leadbeatter~

# Senior Research Scientist, ARRB Transport Research Ltd
* Research Fellow, Australian Council for Educational Research Ltd
~ Manager, Driver Assessment Projects, Registration and Licensing Department, VicRoads

BACKGROUND

In 1999, the Registration and Licensing Department of VicRoads commissioned a team led by ARRB Transport Research Ltd (ARRB TR) to update and extend the Hazard Perception Test (HPT). Professor Gitte Lindgaard and the Australian Council for Educational Research Ltd were sub-contracted by ARRB TR and contributed to a range of tasks within the project.

A paper presented at last year’s Road Safety Research, Policing and Education Conference (Catchpole and Leadbeatter, 2000) described:

?? Victoria’s existing Hazard Perception Test;
?? the objectives for the updated test;
?? the development of a new human-computer interface and new instruction messages;
?? a trial of the user friendliness of the new interface and instructions;
?? the types of accidents in which newly licensed drivers are most likely to be involved; and
?? the development of scripts for test items addressing the most relevant accident types.

The purpose of the present paper is to describe the methods and results of a major trial of the newly developed test items and test construction. The trial formed the basis of the final construction of the updated Hazard Perception Test.

FILMING OF THE TRAFFIC SCENARIOS

As the scripts and other specifications for filming were based on rigorous analysis of traffic hazards and psychometric testing considerations, there were no creative interpretations in the production of video material. There were three main stages involved with the filming of the scenarios. These were video recording; editing to combine driver perspectives to give the impression of a single continuous sequence of driver activity; and conversion of raw material to MPEG format, including compression for optimised delivery in a defined PC environment.

All traffic scenarios required the recording of a driver’s-eye perspective. This included when looking ahead and to the right. Approximately 20 to 30 seconds of material was recorded for each traffic scenario, of which about 15 seconds centred on the specific scenario to be used in construction of eventual test items. To enable the required speed and gap selection judgements to be made, the field of view was such that it results in a realistic impression of speed. That is, the filmed speed gives the impression that the car is moving at the speed shown on the simulated speedometer.

A critical element of this project was the safe and appropriate management of traffic during filming. Safety and traffic management plans were prepared for all filming locations to ensure that actors, film crew and members of the public were not exposed to inappropriate risks during filming. Traffic management requirements varied between locations and times of day, with warning signs, lane closures, temporary diversions and other measures being used as appropriate.

A combination of constraints in terms of time, cost, safety, disruptions to normal traffic flow and availability of sites meant that not all of the scripted traffic scenarios could be filmed. Minor compromises were necessary for some of the scripts that were filmed. These included changes to light conditions, weather conditions, road geometry, roadside vegetation and close approaches by traffic units on a potential collision course.

Of the 150 traffic scenarios originally scripted, 110 were captured on video tape. One scenario was later dropped because the quality of the video was judged to be insufficient to allow test candidates to perceive the main cues. For five of the scenarios, two alternate versions were captured and edited, with a view to selecting the more successful version after trialling. Thus a total of 114 items were available for trialling. The video footage for each item was reviewed jointly by VicRoads and ARRB TR to determine the location of the Correct Response Window (the time during the video clip when the candidate should respond to be awarded a point for the item) and other editing parameters.

HARDWARE AND SOFTWARE FOR THE TRIAL

Trial sessions were conducted using five computers to run the test. On one computer, participants responded to test items by touching the screen. On the remainder, participants responded by clicking the mouse.

The HPT development team created functional specifications for software to implement the updated HPT for trialling new items. The software was developed by a contractor working directly for VicRoads. The software supplied did not meet some of the requirements of the specification documents, including the requirement for an audio track to accompany instructions displayed on the screen and the option for candidates to repeat the block of four practice items before progressing to the test items. Thus the human-computer interface during the trial was slightly different to that which will be used when the test is installed in VicRoads offices for probationary licence testing.

Four test forms were constructed for the trial. A test form is the set of items presented to a single test candidate. Each test form included 57 test items (half of the 114 items to be trialled), plus four practice items selected from the remainder of the item pool. Each test form had 50 per cent overlap (50 per cent of items in common) with two of the other three test forms. The four practice items in each form included representation of all three task types ("click the mouse button when you would slow down/overtake/complete your turn"), items which required a response and items which did not and items filmed from moving and stationary vehicles. Thus the practice items familiarised the participant with all of the item types contained within the test.

The trial software randomly selected the test form for each participant from among the four test forms available. The test items within the form were presented in a pseudo-random order for each participant.

DRIVERS PARTICIPATING IN THE TRIAL

The sample of drivers who participated in the trial included current learner permit holders, probationary licence holders and a small group of highly experienced drivers.

Highly experienced participants comprised commercial driving instructors, VicRoads licence testing officers and highly trained police drivers. They were recruited by telephone contact with their various organisations. Each organisation was informed that suitable drivers were required to have a minimum of 15 years driving experience, to have driven at least 300,000 kilometres in cars and to have been involved in very few accidents.

The target population for the HPT comprises learner permit holders who are ready to take their probationary licence test. Drivers who have only recently acquired a learner permit are unlikely to have accumulated sufficient supervised driving experience to develop their hazard perception skills to the level needed to pass the HPT. Learners participating in the trial were therefore restricted to those who had held the learner permit for at least 6 months at the start of the trial (to ensure they had some driving experience) and who would be aged 17 years 9 months to less than 18 years at the end of the trial (to ensure they could not take the probationary licence test before participating in the trial).

Probationary licence holders were recruited to participate in the trial to provide a more experienced contrast group with whom the learner permit holders could be compared. Verifying that drivers with significant solo experience perform better on the test items than learner permit holders would provide a degree of validation of the test as a measure of hazard perception skill (rather than, for example, merely testing reaction times). To provide a clear contrast with the target population, probationary licence holders participating in the trial were restricted to those who had held a probationary licence for at least two years.

Lists of current learner permit and probationary licence holders were supplied by VicRoads for recruitment of trial participants. A letter was sent to drivers from these lists who met the trial criteria offering a $20 cash inducement. They were asked to return a consent form and supply their telephone number if they were interested in taking part in the trial. Drivers who responded were contacted by telephone to make an appointment for them to participate in the trial. Highly experienced drivers were recruited from among police, commercial driving instructors and VicRoads licence testers.

Trial sessions were conducted in both metropolitan and country areas. Around 75 per cent of participants took part in sessions in metropolitan Melbourne and the remainder took part in sessions in Bendigo and Ballarat. Session times were 9:00 a.m. to 5:00 p.m. on Saturdays and Sundays and 1:00 p.m. to 9:00 p.m. on Tuesdays, Wednesdays and Thursdays to suit the expected availability of potential participants.

A total of 405 drivers participated in the trial. However, 27 of these drivers were unable to complete the full test session due to malfunctions of the software and/or hardware. The databases from the five trial computers were found to contain records for 413 test sessions, of which 50 contained no test item responses. This was apparently a result of sessions being aborted due to software problems and some participants starting again on another computer. Thus the sample of participants from whom test item responses were collected comprised 363 cases.

ANALYSIS OF RESPONSES TO TEST ITEMS

Methods used

A data matrix was constructed that had all the original 114 items represented. Each participant responded to a maximum of 57 items in the matrix. The main aims of the analyses of these data were to generate results on the performance of the items and that of the people. As different forms of the test instrument were used on different people, this meant that differences in the difficulty of the test forms needed to be accounted for when the person ability measures were estimated. The same is true for the item difficulty measures, where the estimates of item difficulty needed to be independent of the abilities of the group of people who responded to them. The Rasch model (Rasch, 1980) was used due to its objective measurement properties.

The item calibration procedure serves the purpose of measuring both the relative difficulty of the items and assessing the extent to which the items work together to represent a single underlying trait, in this case a candidate's driving hazard perception ability. The item difficulties that are produced from the calibration are expressed in logits. The person measures can also be expressed in logits. Both the item and person measures can be expressed on the same logit scale.[1]

The statistical information generated from these analyses was used to flag items that potentially had problems. These problems are usually exposed as items that are confusing, ambiguous or unclear, items that are tapping into traits that are not represented by the majority of the item pool and items that are too hard or too easy for the intended population.

Changes to correct response window locations

The analysis of the data for the purposes of determining item quality was a multi-stage process. The initial analyses were performed by the item task type. That is items of like task type were grouped together to represent the latent trait. There were three task types, slowing down, overtaking and turning. If an item appeared to be performing satisfactorily within its task type then its performance was also analysed within the complete set of items. If it were considered as performing satisfactorily within both sets of items then it was kept in the final pool of items. If item performance was unsatisfactory then the location of the correct response window and the item's ability to discriminate between high, medium and low performers, based on the results of the other items, were both assessed. If it appeared that there were different response patterns from the high, medium and low performers then the item was reassessed to determine whether moving the correct response window could be supported with a substantive reason that was consistent with the intention of the item. When determining the

[1] The difficulty of an item (relative to the other items trialled) and the (hazard perception) ability of a participant are both expressed in logits, a scale ranging from minus infinity to plus infinity. The logit is the natural logarithm of the odds of the event, where the odds of the event is defined as the ratio of the probability that the event will occur to the probability that the event will not occur. The logit scale is used because it is an interval scale. That is, if the difficulty of Item A is 1.0 logits greater than the difficulty of Item B, then the odds of a participant responding correctly to Item B are 2.7 times the odds of the same participant responding correctly to Item A, regardless of whether this person has high or low ability. Similarly, if the ability of Person A is 1.0 logits greater than the ability of Person B, then the odds of Person A responding correctly to an item are 2.7 times the odds of Person B responding correctly to the same item, regardless of how difficult the item is.

location of the correct response window, particular consideration was also given to the responses logged by the 'Expert' group. For a small number of items, a change in the location of the initial correct response window was made, while other items were deleted due to poor performance and/or lack of any defensible reason for moving the correct response window. This procedure was repeated until no further changes or deletions were warranted. This procedure resulted in a final pool of 90 items. Four items from this pool were set aside for exclusive use as practice items.

Comparison of touch screen and mouse response modes

There was a subgroup of the sample whose item responses were collected using the touch screen technology. Seventy two participants were administered the test using the touch screen and 291 participants used a mouse click to register their responses. Comparisons were made to determine whether the relative difficulty of the items was stable across the two modes, whether there were any differences in the performance of participants across the modes and whether there were differences in the response patterns across modes.

The relative difficulty of the items showed to be stable across mode type. The differences in performance of the two groups of participants were not statistically significant at the five per cent level. The variance in responses to items by mode type were similar. There was a small difference in the average response time between the two modes.

Based on these comparisons there appears to be no supporting evidence for this sample that the response mode affects the performance of the items or participants, except for a slight delay (150 milliseconds) in response time on the Touch screen mode. This delay has been adjusted for when analysing the responses of participants who used the Touch screen mode. All responses have been used regardless of mode in all other analyses conducted.

Performance of respondent groups

The performance of various groups within the sample was examined and tested for differences. These data have only been analysed to the level of main effects. The groups compared are based on gender, driving experience, main language spoken at home and frequency of computer use.

The mean performance of males was higher than that of females, with the difference being statistically significant at the five per cent level. This result is consistent with that found by Congdon (1999), where performance was based on the then current Hazard Perception Test.

The mean performance of the expert group was higher than that of the Probation licence group which in turn was higher than that of the Learners permit group. Although these differences were in the expected direction, they were not statistically significant at the five per cent level.

The mean performance of the group reporting English as the main language spoken at home was higher than that of the group reporting other languages. The difference was not statistically significant at the five per cent level.

Performance did not vary significantly at the five per cent level with frequency of computer use. Although the means for the three groups were slightly different, the pattern of difference does not suggest that frequency of computer use affects performance on the test. Only a very small proportion of participants reported using a computer less than once a week, making it unlikely that a significant difference between groups could be obtained.

TEST CONSTRUCTION

A test form is the set of items presented to a single candidate. Fifteen fixed test forms were constructed from the final pool of 90 test items. Each form contains 28 test items and four practice items. All forms are matched in difficulty and in the number of items addressing each type of accident.

To allow the forms to be matched on accident type, the items were classified into eight groups according to the main accident type (as defined in the VicRoads Definitions for Classifying Accidents) addressed by the item. The eight accident type groupings were: on path, off path on curve, off path on straight, pedestrian, rear end, adjacent approaches, opposite approaches and manoeuvring. To allow the forms also to be matched on difficulty, items within each accident type group were further classified into sub-groups by item difficulty. A total of 28 sub-groups were defined across the eight accident type groups because 28 test items were required for each test form. Each test form contains one item from each of the 28 sub-groups. For example, if three items addressing a particular accident type were required in each test form, then three difficulty sub-groups (easy, medium and hard items) were defined within that accident type group. One easy item, one medium and one hard item addressing that accident type were then included in each test form. In determining the number of difficulty sub-groups within each accident type group, and hence the number of items from that accident type group appearing in each test form, consideration was given to the representation of novice drivers in each accident type.

The process of test form construction was both controlled and reviewed, including consideration of the overall usage of individual items, a balance of correct window start locations, task types and window location by task interaction. Consideration of the location of the response window during the item ensured that each form contains a range of correct response window locations (early, middle, late or no correct response window). Thus candidates cannot expect to pass by systematically responding in the same way to every item. Similarity of items was also considered, so that items that are mutually incompatible due to similarity do not appear in the same form. Consideration of mutual incompatibilities meant that it was not possible to use the same set of practice items for every test form; substitute practice items were sometimes required.

Test Reliability

The reliability of the test to separate the trial population based on ability is 0.68. This index is known as a person separation reliability index (Wright and Masters, 1982). This value means that 68 per cent of the variation in participant scores is accounted for by the hazard perception ability of the participants as measured by the test and the remaining 32 per cent of variation in scores is due to measurement error. These are composite measures based on all cases that describe the certainty of differences in the measures of the trial sample. Typically items that contribute most to this type of error are those that have a different meaning to the major subgroups in the sample, those that discriminate poorly and those that perform differently depending on how they are administered. The person separation reliability of the original item pool of 114 items was 0.74. The subsequent removal of the items that have performed least well has resulted in the slight increase in reliability to 0.77. The length of the individual trial test form was 57 items (from the complete item pool) and approximately 45 items were left in each trial form (from the final item pool) at the completion of item analysis (prior to setting aside 4 items to serve as practice items). As test length affects the precision of the individual person measures and therefore reliability, an adjustment has been made to compare the two reliability values. To overcome differences in test length, the Spearman Brown prophecy formula can predict what the reliability of a test would be if it were of a different length. The current HPT has a reliability of 0.27 from 12 items (Congdon, 1999). If it were 28 items in length the reliability would only be 0.46. This value is less than the predicted reliability of 0.68 for a 28 item test based on the final pool of items that have been reported on here.

To compare the precision of the updated HPT with that of the current test, it is useful to consider a candidate whose true ability is 0.5 logits above the ability required to just pass the test. In theory, if the test were perfectly free of measurement error, this candidate would always pass. However, all real world tests are subject to measurement error. Using the current HPT, a candidate whose true ability is 0.5 logits above the ability required to just pass the test would have a 24 per cent chance of failing due to measurement error. Using the updated HPT, the chance of such a candidate failing due to measurement error has been reduced to 14 per cent.

ROLL-OUT INTO VICROADS OFFICES

A complete redevelopment of Victoria’s Computerised Licence Testing (CLT) system is under way. As part of this redevelopment, the new HPT will be integrated into the new CLT to run on standard hardware including the use of mouse interaction rather than touch screen interaction. The new HPT is being translated into 20 community languages and is planned to be rolled out to all Victorian Registration and Licensing offices over a staged six week period commencing October 2001.

CONCLUSIONS

The trial revealed that responses made by touching the screen were systematically delayed by an average of 150 milliseconds when compared with responses to the same items made by clicking a mouse button. However, there was little or no difference in response variability and there was no interaction between items and response modes. The test is expected to function equally well using either response mode, provided that an appropriate adjustment is made to item correct response windows to allow for the systematic delay if touch screen responses are used.

Compared with Victoria's current Hazard Perception Test, the updated test has a larger pool of 86 test items which covers a greater range of traffic situations and contains a greater spread of correct response window locations. The item video sequences are longer than those in the current test and the video quality has been improved. These factors combine to provide a better representation of real world driving experiences. They also discourage candidates from developing strategies for responding based on guessing where the correct response window is likely to be located.

A number of improvements have been carried out to make it easier for candidates to understand and navigate through the test. These include:

?? an increase in the number of practice items for each candidate from one to four;
?? provision of immediate feedback on the correctness of the response after each practice item;
?? redesign of the user interface to provide improved consistency; and
?? rewriting of all instructions and explanations.

The number of test items to be answered by each candidate has been increased from 12 to 28. Items which are not consistent with the majority of the item pool have been eliminated. These factors have combined to increase the precision of the test and to increase its reliability as a tool for discriminating between drivers with adequate and inadequate hazard perception skills from 0.27 to 0.68.

The improvements that have been made in the updated Hazard Perception Test are expected to lead not only to improved public acceptance of the test but also to increased predictive validity. VicRoads plans to assess the validity of the updated Hazard Perception Test as a predictor of accident involvement once the test has been in use in VicRoads offices for at least a year and a sufficient body of data has accumulated.

REFERENCES

Catchpole J and Leadbeatter C (2000). Redevelopment of Victoria's Hazard Perception Test. In: *Road Safety Research, Policing and Education Conference, 26-28 November 2000: Handbook and Proceedings*, pp 327-334. (Queensland Transport: Brisbane, Queensland).

Congdon P (1999). *VicRoads Hazard Perception Test, Can it Predict Accidents?* Contract Report No. CR 99-1. (Australian Council for Educational Research: Camberwell, Victoria).

Rasch G (1980). *Probabilistic models for some intelligence and attainment tests.* (University of Chicago Press: Chicago, USA).

Wright BD and Masters GN (1982). *Rating scale analysis: Rasch measurement.* (MESA Press: Chicago, USA).