Title
Redevelopment of Victoria's Car Driver Learner Permit Knowledge Test

Names of authors
Butler, A., Australian Council for Educational Research, Congdon, P., Australian Council for Educational Research, Imberger, K., VicRoads.

Name of company where the work was conducted
Australian Council for Educational Research Limited

Abstract
It is widely acknowledged that young and inexperienced drivers are over represented in traffic accidents compared with more experienced drivers. It is important that a licensing system assesses the critical aspects of an applicant's knowledge of road rules and driving performance. Some research suggests that a licence testing program directed at critical knowledge requirements is capable of reducing the likelihood that drivers would be involved in crashes for which they are responsible.

In Victoria, a Graduated Licensing Scheme was introduced in 1989. The first stage of the Scheme is for an applicant to obtain a learner permit by sitting a computerised Knowledge Test. The preparation for this test is based on the *Road to Solo Driving* handbook. This new handbook contains information on the complexity of the driving task, hazard perception, how to best acquire the necessary skills for solo driving, and the road rules.

Due to the introduction of the new handbook and new road laws, a new Learner Permit Knowledge Test has been developed. This paper provides an overview of the new Test's development process including its ergonomic design, the design of test items, trialing of the test items and interface, and psychometric properties.

Keywords
Ergonomics, knowledge test, learner permit, novice drivers, psychometric, testing

Introduction
It is widely acknowledged that young and inexperienced drivers are over represented in traffic accidents compared with more experienced drivers. In Australia, 16 to 24 year olds comprise about 15 percent of the driving population, but account for around 35 percent of fatal and 50 percent of injury crashes (Macdonald, 1994 sited in 1). It is important that a licensing system assesses the critical aspects of an applicant's knowledge of road rules and driving performance. Some research suggests that a licence testing program directed at critical knowledge requirements is capable of reducing the likelihood that drivers would be involved in crashes for which they are responsible (AAMVA, 1997 cited in 1). Thus, the purpose in giving knowledge tests before learners are permitted to drive on the road is to ensure that driver licence applicants possess the information required to operate vehicles in a way that is consistent with the safety and mobility of the public.

In January 2001, VicRoads commissioned the Australian Council for Educational Research (ACER) to redevelop the Victorian (car driver) Learner Permit Knowledge Test by the end of the 2001 financial year. Monash University Accident Research Centre (MUARC) was subcontracted by ACER to assist with road safety advice. The aim of the project was to develop a new item pool with a minimum of 300 reliable items based on the new *Road to Solo Driving* handbook (which has replaced the old *Victorian Traffic Handbook*), as well as a new ergonomic design for the test items and associated test instructions. The items and design developed were then trialed to enable a valid and psychometrically sound test to be constructed.

Methods
Preparation for Test Development
An investigation of current international learner and probationary driver knowledge tests from Sweden, USA, Canada, Europe and New Zealand was undertaken to provide direction for the redevelopment of the test. The content, item format, wording of items, scoring standards, alternate test forms, item analysis, test administration, reliability and validity of tests, and problems and benefits of tests were investigated (1).

Development of Test Items
In order for the test to have good content validity an estimation was made of the proportion of test items to be included from each of the five chapters contained in the *Road to Solo Driving* handbook. These proportions were based on judgements made by the MUARC team regarding the relative importance of these information

categories from a road safety viewpoint with consideration of the amount of material in each chapter. The proposed percentage of items for development from each chapter of *Road to Solo Driving* are shown in Table 1.

Table 1 – Proposed Percentage of Items from Each Chapter of Road to Solo Driving

Chapter of *Road to Solo Driving*	**Percentage of Items**
1) The Challenges of Driving	4%
2) Learning to Drive	12%
3) Managing Risk	11%
4) Rules and Responsibilities	72%
5) Your Driving Log	1%

As per VicRoads requirements, it was aimed to develop a pool of 300 items of which 60 per cent of items were to contain a graphic.

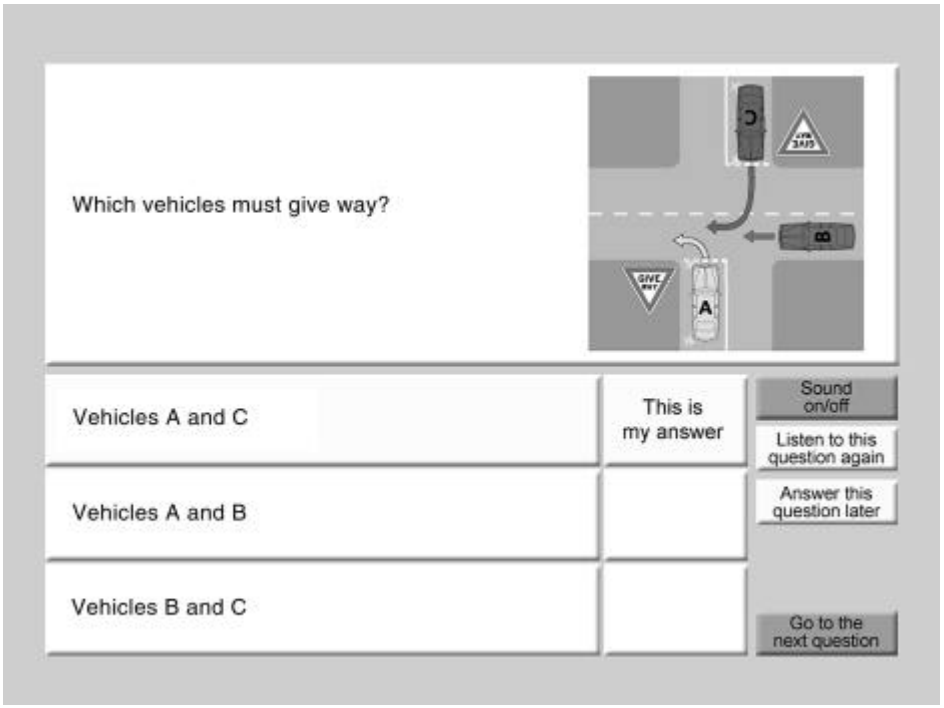
When the items were designed, specific guidelines were followed, such as multiple choice format with one correct answer and two plausible incorrect answers and distractors to contain relevant subject matter. Items were also written for a twelve year old reading age, checked to ensure cognitive load was kept to a minimum and that candidates from different socio-economic or cultural backgrounds were not disadvantaged. In addition the need for each item to be translated into approximately 20 languages was recognised, thus a linguistic expert checked each item.

All developed items were scrutinised during a panelling session. The panel examined the tenor and soundness of the items for example in terms of precision and clarity of the phrasing and reasoning or logic that gives an answer. The objective of the panel session was to challenge the material. Participants worked through the material and the questions before the session (without knowing the proposed answers) and in the panel they compared answers. In some cases this involved challenging the answer proposed and offering an amendment or alternative, or even challenging the basis of the question itself. As a result of the panel session items were re-written as necessary. Items were then sent to MUARC for a review of their suitability in road safety terms. Finally VicRoads checked the items for overall suitability in terms of item construction, road safety and road law (2).

Development of Ergonomic Test Items and Interface

The new interface and item design created was based on the current existing design of the knowledge tests contained in the VicRoads Computerised Licence Testing System and from the input of ergonomic experts. It has been designed to maximise usability and minimise the effect of a candidate's different levels of computer familiarity. This was important as VicRoads is changing from a touchscreen interaction to that of mouse interaction. The use of a single format for all items, with and without graphic stimulus, reduces the difficulty that less experienced computer users will have in familiarising themselves with the assessment. The item graphic user interface (GUI) also uses clear task-differentiated neutral colours, which remain consistent throughout the introduction, practice items, and assessment items. Coloured boxes are used to clearly separate functional areas of the item GUI. These functional areas are the question, answer and navigation/control areas. See Figure 1 for the GUI design. The architecture of the assessment and design of the GUI also sought to reduce the impact of language/culture skills by keeping the functionality to a minimum, allowing candidates to focus on key functions clearly explained in the introductory screens and practice items. At all times the numbers of choices a candidate has are minimal and opportunities for checking and additional instruction are built into the architecture prior to the commencement of the assessment (3).

Figure 1 – The Graphic User Interface



Trial

Sample. Participants were individuals who were about to apply for their learner permit or had recently done so. To ensure maximum participation, every secondary school in Victoria was sent background information about the redevelopment of the test by VicRoads and their students were invited to participate in the trial. Schools who were interested in participating completed a fax form and returned this to ACER. These schools were then sent a letter by ACER providing detailed information about the trial including the demographic data that students would be required to complete, internet address for the trial, username and password. One hundred and fifty-seven schools responded to the invitation out of a total of 447. This included government (107) and private and catholic schools (50). To include participants who had left school early and non-English speaking adults every Adult Community and Further Education centre and Adult Migrant Educational Service centre was contacted by email or letter inviting their students to participate in the trial. A small number of TAFE institutes that offered VCE or special programs for youth were also contacted. A total of 21 centres across Victoria agreed to participate.

Materials and Procedure. The trial of the test items (a total of 508 items contained within 21 overlapping test forms) and design was conducted via the internet. As proposed in the real test users were provided with the option to listen to each test question and answer. However, the use of audio was optional. The trial also included a pre test survey and a post test survey. These surveys were used to gather demographic data and the evaluation of various aspects of the test, including test design, instructions, and perceived difficulty of the test. A database, developed by ACER IT staff, housed the 21 different test forms. This database kept a record of the number of times each form was used. As each candidate logged in, the system was able to present the candidate with the form with the least number of hits. This ensured that the forms were evenly distributed during the trial. This also had the advantage that when a school had a whole class completing the trial, they were presented with different trial items (4).

Results

Over 6000 individual hits were recorded during the trial period. However, only limited data was logged before many of the participants exited the site. This was not unexpected given the voluntary nature of the data collection process. These cases have been filtered out of the description of the achieved sample as they supplied very little information towards the functioning of the items and test delivery method. Those people who responded to the trial test and completed at least 5 out of the 31 items made up the achieved sample. Only those people who logged on to the web site using a code to indicate that they were part of the targeted sample have

been included in the achieved sample. After making the above modifications to the data set there were 2775 remaining cases included in the sample.

The demographic data from the internet pre test survey revealed that the majority of the sample, (two thirds of the respondents) reported that they had not studied the *Road to Solo Driving* handbook prior to sitting the trial test. The individuals who had recently received their learner permit had studied the *Victorian Traffic Handbook* rather than the *Road to Solo Driving* handbook. Slightly more males than females were in the sample (the difference was 1.3%). The majority did not have a drivers licence, while 24% held a Learner Permit. The vast majority were within the age range of 15 to 18 with the actual range being 7 to 100 years. The vast majority also reported that English was the main language spoken at home, with less than 9% indicating that English was not the main language spoken at home. The level of computer usage was high with less than 7% reporting that they used a computer either monthly or less. There was an initial concern that people would refer to the *Road to Solo Driving* handbook during the trial test in an attempt to answer the items correctly. As this would not represent standard testing situations, people were asked if they would agree to not refer to the material. Forty seven people disagreed and were removed from the achieved sample.

The post test survey revealed that the majority of candidates believed they either just passed (56%) or passed the test easily (33%). The majority believed they would have done about the same if the test had been a paper and pen version (74%). The same applied with respect to doing better or worse if they knew more about computers (91% believed they would perform the same). One of three possible options were provided for describing the most difficult part of the test. Participants responded that the most difficult part of the test was understanding the different possible answers (37%) followed by remembering the facts from *Road to Solo Driving* (35%) and then understanding the question (28%). The majority found the design of the screen to be clear and simple (84%) and the same applied for moving from one question to the next (86%). Most felt that the practice questions and instructions were not required (69% and 62% respectively). Only 23% of candidates listened to the audio instructions.

Candidates were presented with an open ended question ('any other comments') at the end of the post test survey. Of approximately 800 responses gathered from all respondents, most registered a "no comment". There were also quite a few with little clear meaning. Of the remainder, by far the most common comments were that the pages were slow to load over the Internet (104 responses), and that the respondents were disappointed that they did not receive feedback on their responses (82 responses) [4].

Discussion

Item Calibration and Construction of Final Item Pool

Items were calibrated with the purpose of measuring both their difficulty and assessing the extent to which they work together to represent a single underlying trait, in this case, a candidate's knowledge of the *Road to Solo Driving* handbook. The person measures that are produced from the calibration converts the ordinal raw responses into an interval measure expressed in logits. The item difficulties are also expressed in logits. Both the item and person measures can be expressed on the same scale.

A data matrix was constructed that had all the original 508 items represented. The main objectives of the analyses of the person item responses were to generate results on the performance of the items and that of the sample population. 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 applies for the item difficulty measures, where the estimates of item difficulty needed to be independent of the group of people's abilities who responded to them. The Rasch model, due to its objective measurement properties, was used to determine these measures.

To measure how well the items work together to represent a single underlying trait a number of fit statistics were generated. These fit statistics represent the difference between the modelled responses and the observed responses. The mean square fit statistics provide a measure of the magnitude of item coherence to the underlying trait. For example if the mean square fit statistics were within the range of 0.77 to 1.3 for all items except one item which had a value of 2.0, it would indicate that the single item was an outlier and probably was measuring a different underlying trait. The items in the final pool had mean square residuals within the range of 0.68 to 1.34. These values are weighted fit statistics. The weighting is applied to the observed responses close to the average difficulty level of the item. When these observations are unweighted it is not uncommon for a small number of unexpected observations from people around the extremes of the scale to inflate the statistics. This would occur when someone of high ability gets an easy question wrong, or someone of low ability gets a hard question

correct. It is important to note that these fit statistics are sample dependent and as such will vary with both different groups of cases and different sets of items.

By using the weighted mean square fit statistics, the magnitude of coherence to the underlying trait, or lack of, can be expressed as the change in variation of observed responses required for the item to be consistent with the other items in the scale. The mean square fit statistics have a mean of 1.0 and range from zero to infinity. When the observed responses fit the modelled responses the weighted mean square will be 1.0. The modelled responses are defined by the average of the responses from all of the items. For all items there is a variance between observed response and modelled responses. For an item with a fit statistic of 1.3, the 0.3 indicates that the variance between the observed and modelled response is thirty per cent more than the average. Similarly a fit statistic of 0.77 indicates there is 30 per cent more variance required between the observed modelled responses for this item to fit the same as the average of other items.

The item difficulty estimates represent the difficulty of an item relative to the other items in the scale. The location of the items on a scale involves using an arbitrary origin. The origin used in these results is the mean of the items' difficulty estimates, which is set at 0.0 logits. Items with a positive logit value are more difficult items and those with a negative logit value are easier items. There are no predetermined acceptable maximum or minimum logit values. However if all candidates get an item incorrect or all candidates get it correct then this item is automatically deleted. Two items were responded to correctly by all those who responded to the items. As a result it was not possible to correctly estimate how easy the items were, as it is only when someone gets an item incorrect and someone else gets it correct that is possible to locate an item's difficulty estimate. The two items have not been included for possible selection into a test form as they are of an unknown difficulty. The difficulty of items in the final pool of 312 ranges from –3.13 to 3.91.

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 measuring 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. The item parameters that are used to indicate these types of problems are low point-biserial correlations for the correct response together with high infit mean square value for the item and extreme item difficulty estimates. There are no set values for determining the exclusion of items. The point-biserial and infit values are used to flag items that need to be examined for potential removal. For these items further analyses on each of the item distracters were undertaken to ensure that the incorrect responses were not attracting the higher performing people.

In summary, 196 items were removed from the initial pool leaving a final item pool of 312 items, with 57% containing graphics. These items are determined to be representative of a candidate's knowledge of the *Road to Solo Driving* handbook, are clear, and unambiguous and have acceptable psychometric properties (5).

Test Form Construction

Each test form consists of 32 items and the content representation of each test form is fixed. In every test, one question will be drawn from Chapter 1 – "The challenges of driving" and one from Chapter 5 – "Your driving log". Four questions will be based on the contents of Chapter 2 – "Learning to drive" and three from Chapter 3 – "Managing risk". The largest proportion of items, a total of 23, will be from Chapter 4 – "Rules and responsibilities". The item pool is summarised in Table 2.

Table 2 – Breakdown of Item Pool by Chapter in the Road to Solo Driving Handbook

Chapter Number	No. of Items in Pool	No. of Test Items	% of Pool (Target)	% of Pool (Actual)	Box Size (Items)	Prob. of Selection
1	12	1	4%	3.8%	12	8.3%
2	34	4	12%	10.9%	8.5	11.8%
3	38	3	11%	12.1%	12.7	7.9%
4	224	23	72%	71.8%	9.8	10.2%
5	4	1	1%	1.3%	4	25.0%
Total	312	32	-	-	-	0.002%

The method of test form construction is designed to minimise the variance in test form difficulty and to control the distribution of items based on their relative difficulty. This is important as difficulty of an administered test can vary in two ways. One way is by sampling the harder or easier items only. The other way is by having

different distributions of item difficulty, which can impact on the most probable raw score achievable, given a fixed candidate ability. For instance, for a person of fixed ability, a bi-modal distribution of item difficulties can produce a different raw score than for a test of the same average difficulty but with a uniform distribution of items. These potential problems are overcome to a degree by the test construction method outlined below. In addition, an essential part of dealing with these problems has been the production of sample independent measures of item difficulty using the Rasch model (as described earlier).

To minimise the differences in test form difficulty and to allow each chapter to be represented in a predetermined ratio, each item has been assigned to a box based on chapter and then relative difficulty. There are 32 boxes as there are 32 items in the test. The content representation of the test, as categorised by chapter number, has been controlled by assigning only items from a common chapter into the same box and by having a predetermined number of boxes per chapter. This information is shown in Table 2. It can be seen that the number of items in each box is not the same for each chapter. This is because the proportion of the number of items in the pool to the number of items in the test is not the same for each chapter. Given that Chapter 1 has 12 items in the pool and one item from this chapter is to be in the test, the box size is 12. In contrast 23 items will be drawn from Chapter 4. To keep the box sizes equal within Chapter 4, each box would hold 9.8 questions. As the number of items in a box needs to be an integer value, some boxes will have 10 items and other boxes will hold 9 items. The effect of this is that some items will be used slightly more than other items from within the same chapter when summed across all test forms. The column labelled 'Prob. of selection' shows the degree to which this effect will occur when comparing item usage between chapters. Any single item from Chapter 5 has a 25 per cent chance of being selected into a test form where as any single item from Chapter 3 has a 7.9 per cent chance of being selected into a test form. This means that the individual items from Chapter 5 will appear more regularly than the individual items from Chapter 3. (5)

Conclusions and Recommendations

This paper discussed the development of the new Victorian Car Learner Permit Knowledge Test. Both the developed item pool, and design of the user interface and items was trialed. The item pool and resulting test has been constructed using latest psychometric test development procedures. Various monitoring strategies are recommended to maintain proper functioning of the test, for example undertaking a comparison of item difficulty estimates between the target population and trial sample, and undertaking any modifications to ensure that difficulty of test forms generated is held constant and the pass rate is maintained. It is recommended that monitoring occur biannually to ensure proper psychometric functioning of the test (6).

References

1. Australian Council for Educational Research. *Report Number 1: Redevelopment of the VicRoads Knowledge Test*. January 2001. Pages 9 and 11.

2. Australian Council for Educational Research. *Report Number 3: Development of Test Items*. February 2001. Pages 1 to 4.

3. Australian Council for Educational Research. *Report Number 4: Review of HCI/Ergonomic Issues*. March 2001. Pages 1 to 4.

4. Australian Council for Educational Research. *Report Number 5: Trialing and Analysis of Tests.* May 2001. Pages 1 to 10, 22 to 30, and 50.

5. Australian Council for Educational Research. *Report Number 7: Construction, Amendment and Functionality of Tests.* June 2001. Pages 3 to 10.

6. Australian Council for Educational Research. *Report Number 8: Strategy for Ongoing Monitoring of Tests.* June 2001. Pages 2 and 3.