Getting Rasch about assessment: Using a psychometric approach to select and confirm items in the new Victorian licence test

Harrison, W.A.
Eastern Professional Services Pty Ltd
Email: wh@eastprof.com.au

Abstract

Selecting assessment tasks and items for the new Victorian licence test relied in part on reducing a large number of potential assessment items to a smaller set of items with a high level of internal consistency. This made use of a psychometric method commonly used in the development of ability tests – Rasch modelling – that does not appear to have been applied to on-road tests. The Rasch approach allows test developers to assess the performance and calibration of items and test takers against a common underlying ability dimension.

The Rasch approach was used to select potential items for the first, screening stage of the test from a pool of items derived from the current Victorian test, the Californian DPE, and the first stage of the New Zealand full licence test. A trial involving 352 learner drivers provided data for the analysis. The analysis identified a small set of internally-consistent assessment items that were used as the basis for the screening stage of the test.

The same approach was used to select potential items for the main part of the new test from a large pool of assessment items originally developed with a focus on construct validity (see McDonald and Harrison's paper at this conference). Data for the analysis were collected in a trial involving over 400 learner drivers. The Rasch analysis identified a subset of internally-consistent items that were then considered for inclusion in the new test.

A third trial with over 500 experienced learner drivers was used to assess performance of the selected test items. A Rasch analysis confirmed that the items in Stages 1 and 2 have a high level of internal consistency and are appropriately calibrated against the ability level of experienced learner drivers.

This paper summarises the results of the trials and their implications for the Victorian licence test, and discusses the potential value of the Rasch approach for licence test development in future.

Keywords

Licence Test, Driving Test, Test Development, Psychometric Assessment

Introduction

Most Victorian driver-licence applicants after 1 July 2008 must have at least 120 hours of supervised driving experience. This improvement to graduated licensing necessitated the development of a new on-road, practical driving test to replace the current test (the POLA). The rationale for developing a new drive test was as follows (see Figure 1):

- The effect of the new experience requirement was expected to be a shifting and narrowing of the distribution of learner experience, as depicted at the top of Figure 1.
- The relationship between experience and skill acquisition is well understood – so the increase in average experience levels is expected to result in an increase in the aptitude of licence applicants.
- This expected increase in the aptitude of licence applicants has consequences for the POLA. The POLA was developed and has been used in a context where licence applicants had relatively little learner driving experience. The ability of a driving test to identify more- and less-able learner drivers relies on the calibration of the test items against the skill level of the learner

drivers. The POLA was presumably well-calibrated against the skill level of learner drivers before the new experience requirement. It is unlikely to be well-calibrated against the improved skill levels expected under the new requirement. Identifying relatively skilled and unskilled drivers will require changes to the test to match the improved skill levels of drivers.

- This calibration issue necessitates the development of a new test rather than simply changing the scoring criteria of the POLA because the calibration of a driving test depends on item difficulty and on the match between the type of skills assessed in the test and the skills of the driver. Although a test calibrated for low levels of driving experience might assess skills relating to simple car control to identify more- and less-skilled drivers (such as hill starts using a handbrake), a test calibrated for learners with high levels of experience might better assess skills relating to the perception of and response to hazards in busy traffic.

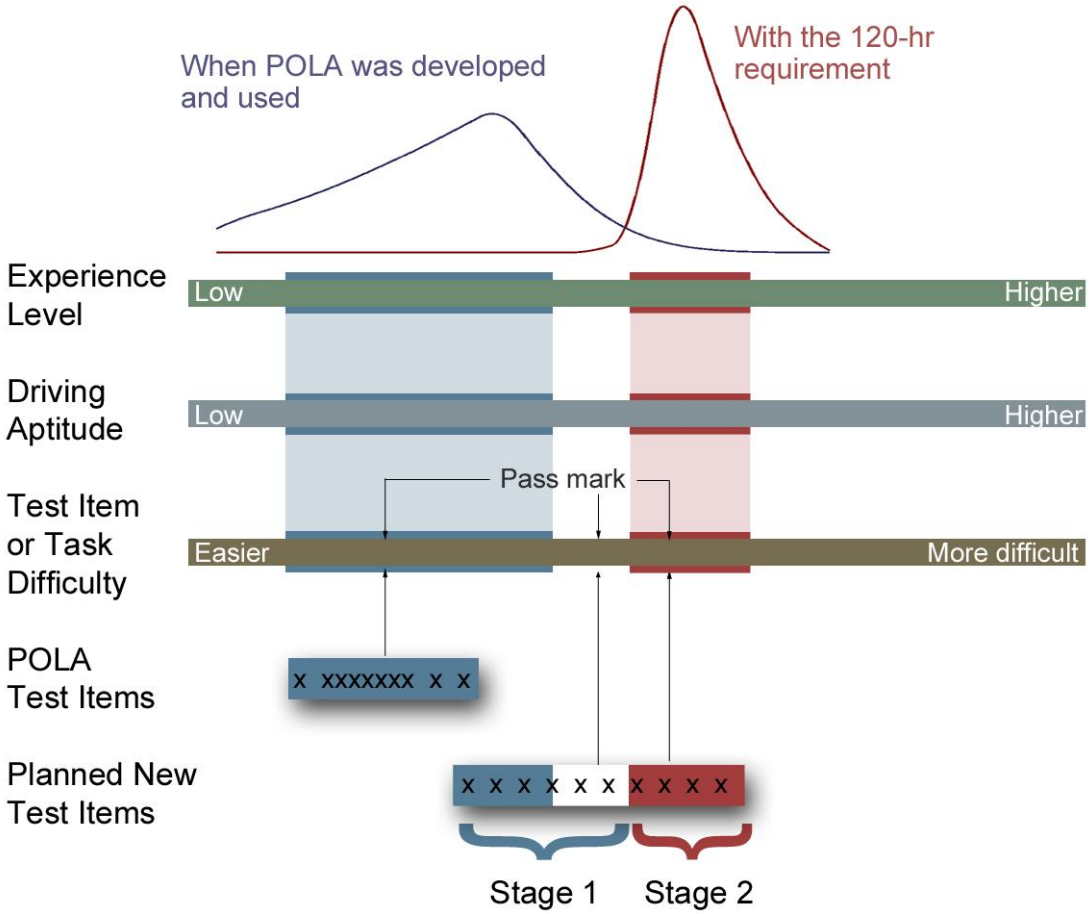

Figure 1: Experience, driver aptitude, and calibration of items in the POLA and the new test

VicRoads chose to develop a new drive test, in part, to provide a well-calibrated assessment of driving skills that would identify safer and less-safe licence applicants with the relatively high experience levels expected under the new requirements. Test development and the trials were managed by VicRoads, assisted by a group of external consultants with expertise in relevant areas. The test development team commenced with a recommendation that the test should assess driving in relatively busy, real-traffic situations. As this could be dangerous for licence applicants without basic skills, an initial screening stage was recommended as part of the test.

The result was a plan for a two-stage test as shown at the bottom of Figure 1, with the first stage in a less-challenging driving context to ensure licence applicants are ready for the busier driving contexts proposed for the second stage. The test development team was therefore required to develop and trial two sets of test items (one set for each stage) and then to trial the combined test stages as a single test.

The development of two large pools of potential test items and the broad rationale behind the test development project are discussed in other papers prepared for this conference. This paper summarises the method used to select potential test items from the large pools of potential test items for stages 1 and 2 of the test, and then to verify that the selected items performed well as a single group of test items.

The key innovation here was the application of Rasch modelling techniques (described in Bond and Fox, 2001). These techniques are widely used in the development of other psychometric tools, such as ability tests, and are well suited to the identification of an internally consistent, well-calibrated set of licence test items. The project team is unaware of any other instance where these methods have been applied to the development of an on-road licence test. The Rasch approach was used as the basis for the overall method for the project:

- Potential tasks, items, or test requirements were selected or developed assuming that they relate to the specific, single ability dimension of interest – safe driving aptitude. Potential test items or requirements that might be "interesting" but unrelated to safe driving aptitude were not included in the trials. The logic of this restriction is that the test should aim to measure a single ability dimension. As the main reason for having a driving test relates to safety, it seems reasonable to focus on this dimension when developing the test. The Rasch approach includes methods that allow an assessment of each item's "fit" to the underlying single ability dimension to be assessed.

- The test items aimed to assess and score positive behaviours rather than counting errors. Extrapolation of Groeger's (Groeger & Brady, 2004) use of an error-counting measure suggested that this approach may be less useful for more-experienced learners. The error-curve (or learning curve) appears to flatten for this type of measure well before learners have accrued 80 hours of driving, suggesting that observable errors may not discriminate reliably at higher experience levels.

- The selection of potential test items was based on consideration of the young driver crash research and broader human factors and skill acquisition research. These were used to generate expectations about the behaviour of experienced learners that were then be translated into present / not present observations suitable for inclusion in a driving test. The test developers used a behavioural sampling approach – where specific safe driving behaviours were assessed when undertaking an assessment task on a test route that was in turn designed to join a series of assessment tasks.

- These larger pools of potential test items were then trialled with large samples of inexperienced and experienced learner drivers to assess the reliability of the items, their validity (ie their ability to discriminate between experienced and inexperienced learners), and their fit with other items along the safe driving aptitude dimension. The use of the Rasch approach at this stage ensured that the trials provided information about:
    - The placement both of learner drivers and items along the assessed aptitude dimension.
    - The fit of items to the assessed dimension, allowing items with a poor fit to be discarded.
    - The reliability of items at specific ability levels, allowing weak items to be improved or discarded.

Method

Figure 2 shows the general approach taken to the development of the new test. The Rasch modelling methods were applied to data provided by the three trials.

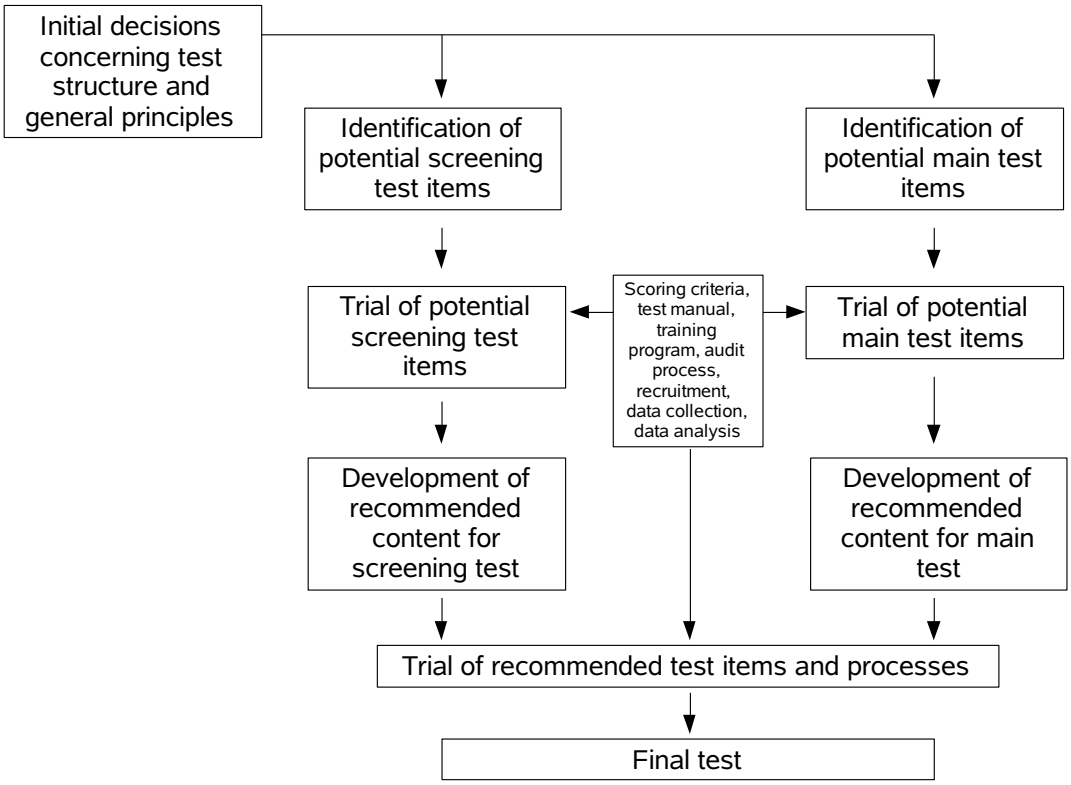

Figure 2: Test development process

The Rasch approach was used in the first two trials to select test items from larger pools of potential test items. In each trial, participants undertook trial assessment drives that included many more test items than were likely to be used in the test. Items were selected for inclusion in each Rasch analysis if performance was correlated with other safety-related measures (these are discussed later). The Rasch modelling method was then used in each trial to identify those items that were well calibrated against the ability level of trial participants and that together resulted in a measurement tool with a high level of internal consistency.

The items that survived the Rasch modelling were then recommended as the basis for the two stages of the test. Further consideration by the project team resulted in some minor changes to these item sets, and they were then used in the third trial to assess the operation of the test with relatively experienced learner drivers. The Rasch modelling approach was then applied to all the items used in the third trial, and further recommendations made to delete any items that did not conform to the Rasch assumptions.

The Rasch modelling method was applied in each trial using the approach outlined in Bond and Fox (2001), making use of the Winsteps software package (Linacre, 2007) after initial data processing and analysis using the Statistica package (Statsoft, 2004).

Trial 1 – the Screening Stage

Potential items for the screening stage of the new test were drawn from the POLA, the Californian DPE, and the first stage of the New Zealand full licence test. The trial involved 352 participants assessed at one of three test locations (Morwell, Frankston or Carlton) by specially trained licence testing officers. Participants were between 16 and 18 years old, and had a mean of 76 hours of driving experience accrued over a mean of 66 weeks as learner drivers.

In all three trials participants completed a questionnaire prior to their assessment drive that collected basic demographic data and information about driving experience (amount and context). It was initially planned to select items for the test based on their relationship with driving experience. It was proposed that the first stage (the screening stage) should aim to identify drivers with and without sufficient experience to drive safely in the second stage of the test, and that the second stage would identify drivers with and without the mandatory 120 hours of experience. In both cases, this relied on identifying a correlation between item performance and driving experience to select items for inclusion in the Rasch analysis.

It was not possible to select potential Stage 1 items based on their correlations with driving experience. Performance on very few items in the first trial was correlated with driving experience, and attempts to identify subgroups of items that together predicted self reported experience were unsuccessful.

An alternative approach was used whereby a potential set of screening items was identified based on the correlations between item scores and the number of interventions by the driving instructor present in the front passenger seat during the assessment drive. Driving instructors involved in the trial would presumably only intervene where they perceived there to be a threat to safety. Assessment items that were correlated with a high number of interventions would therefore be indirectly correlated with unsafe driving behaviours.

The potential items identified using this approach were then used in a Rasch analysis. A summary of the analysis outcome is shown graphically in Figure 3. The Rasch method allows items and participants to be located along a common underlying difficulty/aptitude dimension, assesses the reliability of this estimate of item difficulty or participant aptitude, and assesses the consistency of the item and participant data in relation to the assumption that the items together assess one underlying ability construct. In an ideal test, there would be substantial overlap between items and participants on the underlying dimension (the test would be well calibrated), the location of items and participants would generally be known with some reliability, and items and participants would behave consistent with the assumption that there is one underlying construct assessed by the items. The results of the Rasch analysis can be used to exclude items that do not perform well.

The data in Figure 3 show items (purple) and participants (light blue) located on the underlying dimension identified in the analysis. The diameter of each data point indicates the reliability of the estimate of the item's difficulty or participant's aptitude, and the infit statistic on the vertical axis reflects the consistency of the item or participant with the assumption about a single underlying assessed dimension.

The data in Figure 3 suggest the following:

- Some of the items identified as having a correlation with driving instructor interventions in the trial were too easy for trial participants (items on the left of the Figure).
- A small number of items, although well calibrated, do not appear to assess the same underlying dimension as most items (the items with infit z-statistics greater than ±2).
- The reliability of item and participant estimates was generally good, with those at the extremes of the ability/difficulty dimension having lower reliability levels (greater inaccuracy about their location on the underlying dimension). This is a common outcome in Rasch modelling and reflects the limited amount of data available for participants and items at the extremes.

Items that overlapped with participant aptitude (ie, those that were well calibrated) and items with infit statistics of $|z| < 2$ were recommended for the screening stage. The recommended subset of items for the screening stage included assessment items from the DPE, the POLA, and the NZ full licence test – thus the recommended screening stage content would be an amalgam of items from the three tests shown in the trial to be correlated with a safety-related measure and shown to be well calibrated against the driving ability of learner drivers with an average of about 80 hours of experience. The internal consistency of this small subset of items was good – with a Cronbach alpha of 0.62 and an average item-total correlation of $r = .24$.

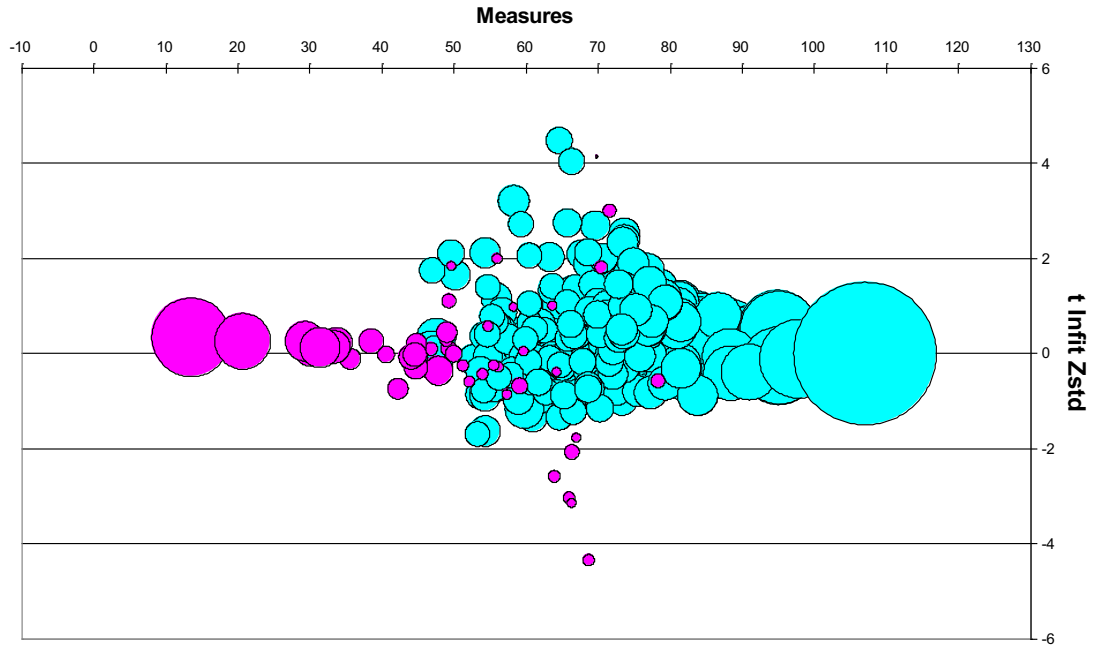

Figure 3: Rasch analysis of items correlated with interventions – estimated locations of items (purple) and participants (blue) on an underlying dimension identified by the modelling.

Trial 2 – The main assessment stage

Potential items for the main assessment stage (Stage 2) of the new test were developed as described in other papers at this conference, with an emphasis on construct validity. It was intended that items should reflect our current understanding of factors associated with unsafe driving amongst young drivers. There were no known tests that were thought to be adequate as models for Stage 2.

The second trial used this large pool of potential assessment items and a sample of 434 learner drivers with a mean of 91.7 hours of driving experience, 2.9 professional driving lessons, and a mean age of 17.6 years. Assessment drives were conducted at the same three locations using the same licence testing officers as in the first trial. The licence testing officers participated in a training program to prepare them for using the potential test items on the test routes developed for each location.

Again it was not possible to identify potential test items that were correlated with total self-reported driving experience, so a number of alternative approaches were used to identify potential groups of assessment items that were predictive of safety-related measures that could then be used in Rasch analyses designed to assess their calibration and internal consistency. The following approaches were used to identify potential sets of items for further analysis. In each case the items identified were subjected to a Rasch analysis to eliminate items that were poorly calibrated with the driving performance or skill level of the participants.

- *Amount of experience in higher speed zones and in wet weather*. Supervising drivers are unlikely to allow their learner driver to drive in challenging driving situations (such as in wet weather or in higher speed zones and on freeways) until they are convinced that they are ready to do so – presumably by the learner's demonstrated driving skill. The amount of driving experience accrued in these challenging situations is therefore a measure of the supervising driver's perception of the learner's safety. Those learners with high levels of wet-weather and high-speed-zone driving may be safer than those with lower levels of this experience.

This may also be true for another reason – learners who have had a breadth of driving experience, including driving in challenging situations, may be safer than other learners because they have had an opportunity to develop driving skills in different contexts. This breadth of experience may be important for the development of driving ability.

Items that were correlated with experience in these challenging driving situations were identified, and those that were well-calibrated against participants' ability levels were included in a pool of potential items for further analysis.

- *Errors and interventions by the instructor during the drive*. As in the trial of potential Stage 1 items, it was possible to identify items that were correlated with the number of errors and interventions during the test drive. Presumably the number of errors and interventions reflects the safe driving ability of participants, so items that are correlated with the number of errors and interventions should be able to identify safer and less-safe drivers.

  Items that were correlated with errors and interventions were identified, but none were included in the pool of potential items because they were all too easy for trial participants – they were poorly calibrated against the ability of drivers in the trial, consistent with the Rasch modelling results in Trial 1.

- *Performance on a distraction task while driving*. One of the tasks included in the trial was a number-recognition task used as a distractor while assessing other items. Performance on the distractor is likely to be influenced by experience – as drivers accrue experience they should become better at controlling the focus of their attention. In some situations this may help their performance on a distractor, and in some situations they may be better-able to ignore distractors and focus on their driving. Test items that are correlated with performance on the distraction task may therefore identify safer and less-safe learner drivers.

  Items that were correlated with performance on the number-recognition task were identified, and some of these that were well-calibrated against participants' ability levels were included in the pool of potential items for further analysis.

- *Self-assessment of safety-related driving skill*. Participants were asked to self-assess their likely performance on a new driving test before they were taken on the assessment drive. Although the young drivers routinely over-estimate their driving skill, within this over-representation it is reasonable to assume that learners with better-developed safety-related skills would provide a higher self-assessment than those with less-developed skills. Their self-assessment should, in part, reflect an assessment of their safety. Items that are correlated with this self-assessment should, therefore, identify safer and less-safe learners.

  Items that were correlated with self-assessment were identified, and some that were well-calibrated against the ability level of trial participants were included in the pool of potential items for further analysis.

The forty-one potential items identified in the analyses described above were then analysed together using Rasch modelling to ensure that the items performed well as part of the group and that they were well calibrated against ability level. Three items were excluded based on a poor fit to the underlying construct identified by the Rasch analysis, and six items were excluded because they were poorly calibrated against participant ability – they were too easy. The remaining items were generally well calibrated against the trial sample and were internally consistent with a Cronbach alpha of 0.92.

Item difficulty did not vary with any meaningful pattern between the three test locations, and there were no consistent sex differences in performance on the identified subset of items. The items remaining in the pool after this analysis, and the driving tasks used for each, were then recommended as items for inclusion in Stage 2 in the final trial.

Trial 3 – Confirmation of the Test Items

After incorporating further considerations raised by VicRoads and the project team, two sets of items (Stage 1 and Stage 2) were recommended for the test. These were trialled at six locations (Morwell, Frankston, Carlton, Bendigo, Bundoora, & Geelong) using a team of licence testing officers who participated in a trial of the training program planned for all VicRoads licence testing officers.

The trial involved having a total of 523 learner drivers undertake the whole of the recommended drive test. Participants had a mean age of 18.6 years, mean self-reported driving experience of 120.6 hours, a mean of 1.9 years as a learner driver, and a mean of 7.7 professional driving lessons. Sampling was designed to target drivers who were likely to be ready or almost ready for their test.

The item-total correlations for items in Stage 1 ranged from r = .21 to r = .78 and the Stage 1 items had a Cronbach alpha of 0.66. The items accounted for 51 percent of the variance in performance, and principle components analysis of the residuals suggested the remaining variance was random measurement error. Similarly for the Stage 2 items – item-total correlations ranged from r = .23 to r = .91 and the items had a Cronbach alpha of 0.83. The 30 percent of performance variance unaccounted for by the items was subjected to a principle components analysis that suggested it was random measurement error.

Figure 4 summarises the results of the Rasch analysis of the Stage 1 items. For each assessment item it shows outfit (a measure of the items' conformity to the Rasch assumption of a single underlying dimension), modelled position against the underlying construct or "measure", and accuracy of the estimated position (standard error – the size of each circle). To aid in interpretation, the modelled measures were transformed to match the range of YES scores obtained in Stage 1. Only one item (Stop Position for a turning task) is of concern. This may reflect a smaller amount of data for this item as there were relatively few turns at Stop signs included in the test routes.

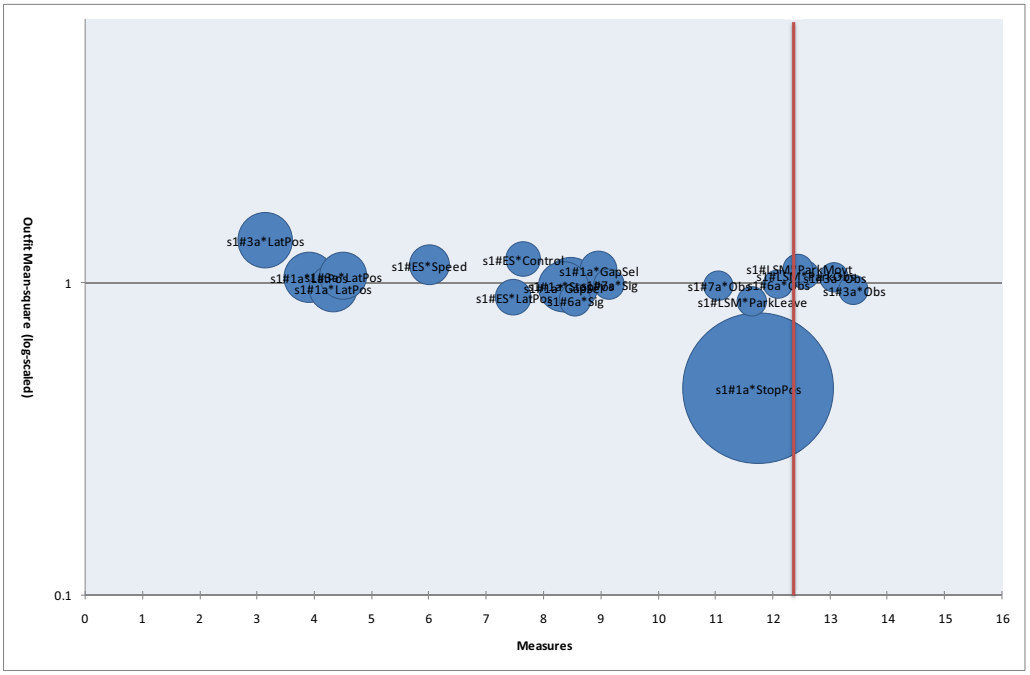

Figure 4: Pathway chart for Stage 1 items. The red line represents the mean performance level of participants against the measure.

Note that the red line in the above graph is the average level of performance of trial participants in Stage 1. The items in Stage 1 were relatively easy for participants, as expected. No items were too difficult. Some items were very easy compared to the ability level of most participants, but no items were so easy that all participants would be expected to meet the assessment criteria.

Figure 5 summarises the results of the Rasch analysis for Stage 2 of the proposed drive test. It shows outfit (a measure of the items conformity to the Rasch assumption of a single underlying dimension), modelled position against the "measure", and accuracy of the estimated position (standard error). Again, the modelled measures were transformed to match the range of total scores obtained in Stage 2.

A number of items at the extreme ends of the range of test performance have large standard errors, and some of these have such small "outfit" statistics that they add little to the information provided by the other items. These could be excluded from the test. One item (a Signalling item) had a large outfit statistic (4.9), suggesting it is a relatively poor fit to the other items. Its infit statistic was within the expected range (0.5 to 1.5), however, so the item could be left in the test pending the collection of further data once the test is live.

The red line in Figure 5 is the average level of performance of trial participants. The items were generally relatively easy compared to the modelled ability level of participants. Some items were so easy that their modelled position on the underlying ability scale is negative. Two items were very difficult for all participants, and at least ten items were relatively easy. The easiest items were those relating to signalling and stop position. The most challenging items were generally those relating to Observation and Lateral Position. There are few items calibrated to the upper end of the aptitude range. Given that the role of the test is to identify drivers with relatively poor safe driving skills, this bias towards the lower end of the aptitude range is unlikely to be a problem.

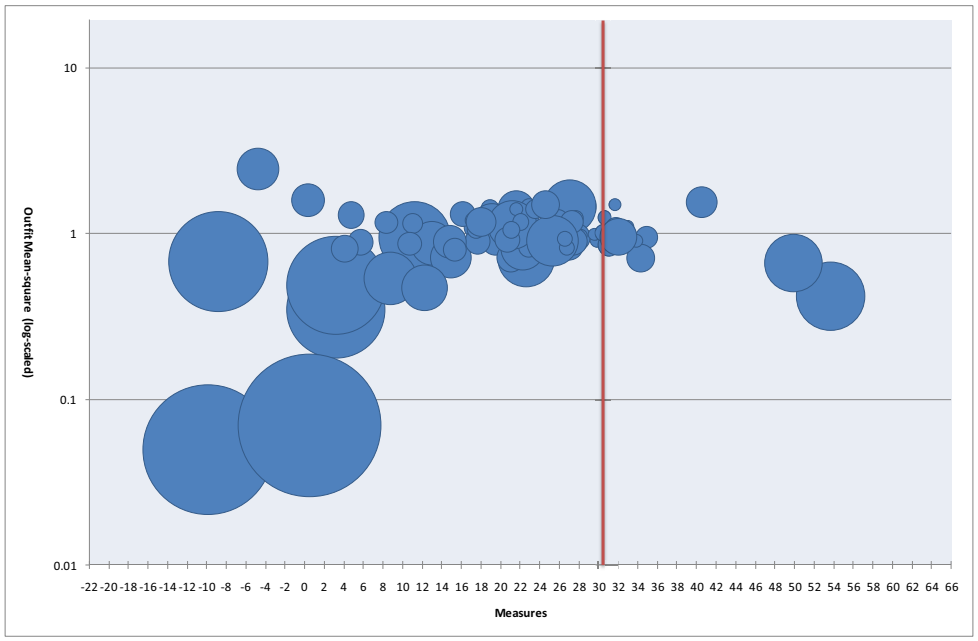

Figure 5: Pathway chart for Stage 2 items. The red line represents the mean performance level of participants against the measure.

The Rasch analyses were used to recommend that some items be excluded from the test. These results, and further recommendations by the project team and VicRoads resulted in the final set of test tasks and assessment items.

Discussion

Rasch modelling methods are routinely applied in the development and analysis of ability tests, surveys, and psychometric instruments across the behavioural sciences, but appear not to have been applied in the development of on-road driving tests. They provide some clear advantages over other psychometric and test development methods. In particular:

- They allow analysis of items and trial participants to assess the performance of the test.
- They allow item difficulty and participant ability to be calibrated against a single underlying dimension or construct, providing better information about test calibration and item reliability.
- They avoid making some of the invalid assumptions about the measurement scales used in the licence test items that are made by other psychometric and test development methods.
- They focus on the development of item sets that have a high level of internal consistency.
- They encourage efficient test development and trialling methods that make effective use of limited numbers of participants.

The Rasch approach was applied successfully in the development of the new Victorian drive test with a dual focus on selecting a small number of internally-consistent items from larger pools of potential items (trials 1 and 2), and confirming the performance of the selected test items (trial 3). The outcome of this approach was that the two test stages are composed of items that appear to perform well together and that are well-calibrated to the ability level of learner drivers with relatively high levels of driving experience.

There was a high level of consistency between the results in trials 1 and 2 and the results in trial 3. The internal consistency measures (the Cronback alphas) were high and consistent between the two development trials and the final confirmatory trial. The consistency of the results suggests that the test should continue to perform well in this respect across different testing conditions.

A number of issues remain to be addressed. The key issue for Victoria is the performance of the test items in a real test situation. The trials used volunteer learner drivers, and although the learners in the last trial were paid a small bonus for "passing" the test, their motivations were not the same as those of licence applicants taking the test. The motivational differences between trial participants and real licence applicants, and the differences in test preparation, are expected to influence the performance of the test and may influence its reliability, calibration, and internal consistency. These issues will be assessed with data from the first few months of the test's operation, and it is anticipated that some further fine tuning may be required.

The consistent failure to identify differences between total learner driving experience and performance on the test items is of some interest and may need to be investigated further. The most likely explanation is a combination of reliability problems in using self-reported estimates of experience and, perhaps more importantly, the incorrect assumption that experience and item performance would necessarily be correlated. This assumption is based on an implicit assumption that a particular amount of experience should be associated with a particular level of performance across all participants. There is no sound reason to believe that each additional hour of accrued experience will have the same effect on the driving skill of all learner drivers. It may have been optimistic to expect to detect a correlation between an unreliable estimate of driving experience and performance on a binary test item that has a variable relationship with experience across different learner drivers.

The Rasch approach to test development and assessment may be worth pursuing in other jurisdictions and in the development of other licensing tests. It would be particularly interesting to use this approach to assess the calibration and reliability of test items used in jurisdictions with different levels of mandatory experience for learner drivers – especially where new tests have been developed and implemented without a thorough, sound approach to assessing the psychometric qualities of the test.

Acknowledgments

The analyses reported here were performed by the author. The project team responsible for the development of test items and making recommendations about the trial and the final content of the test consisted primarily of John Catchpole, Antonietta Cavallo (VicRoads), Ron Christie, Warren Harrison, Darryl Johnston, Wendy Macdonald, Sophie Oh (VicRoads), and Tom Triggs. Other people assisted at various times.

References

Bond, T.G. & Fox, C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences Mahwah, NJ: Erlbaum

Groeger, J.A, & Brady, S.J (2004) Differential effects of formal and informal driver training Road Safety Research Report No. 42. London: Department for Transport

Linacre, J.M. (2007) A User's Guide to Winsteps Ministep Rasch-Model Computer Programs.

StatSoft, Inc. (2004). STATISTICA (data analysis software system), version 6. www.statsoft.com.