Predicting Fitness to Drive Using the Visual Recognition Slide Test (USyd)

Lynnette G. Kay, Anita C. Bundy, Lindy M. Clemson

KEY WORDS
• driving
• fitness to drive
• rehabilitation

OBJECTIVE. The authors examined the construct and predictive validity and internal reliability of the Visual Recognition Slide Test developed at the University of Sydney (VRST–USyd).

METHOD. A historical cohort study using retrospective descriptive analysis of VRST–USyd scores and on-road driving performance for 838 drivers with impairments was conducted.

RESULTS. Rasch analysis provided evidence for the construct validity and internal reliability of the VRST–USyd. Goodness-of-fit statistics for all items were acceptable. The test had high participant and item reliability indexes and separated the participants into four groups with varying levels of skill. Using a cutoff score of 95/164, the sensitivity of the test was 81%, and the specificity was 90%. However, when coupled with clinicians’ judgment of participants’ awareness of their driving performance during the on-road assessment, this score improved.

CONCLUSION. There is evidence for reliability and construct and predictive validity of the VRST–USyd. The measurement of awareness requires further research.


Driving a motor vehicle is an integral part of daily life in modern Western society, enabling people to participate in work, leisure, and family activities. Driving therefore represents not only independence in mobility but also freedom to fulfill life roles, resulting in improved quality of life. Older people who no longer drive report decreased satisfaction with life, loss of independence and personal identity, and increased levels of depression (Stutts & Wilkins, 2003). However, driving is a complex task requiring the integration of basic competence in motor, sensory, perceptual, and cognitive skills in a rapidly changing environment (Coleman et al., 2002). Other factors, such as experience and attitudes, also contribute to safe driving performance (Lundqvist & Ronnberg, 2001).

Medical conditions, disabilities, accidents, and the aging process can cause changes in the requisite skills that potentially affect a person’s ability to drive (van Zomeren, Brouwer, & Minderhoud, 1987). Because testing the actual driving performance of every at-risk driver is both expensive and potentially dangerous (Klavora, Heslegrave, & Young, 2000), researchers have, for the past 20 years, examined a variety of clinical tests to identify a screening test that predicts on-road driving performance (Mazer, Gelinas, & Benoit, 2004). Some researchers have used accident statistics and traffic violations as the measure of safe driving performance (Ball et al., 2006; Bouillon, Mazer, & Gelinas, 2006; Coleman et al., 2002; Katz et al., 1990). Because accidents and violations represent a failure to drive safely and are relatively infrequent events, however, most researchers have preferred to use an assessment of actual driving performance (Bouillon et al., 2006; Korner-Bitensky et al., 2000; McKenna, Jefferies, Dobson, & Frude, 2004). Despite some difficulties...
(such as unpredictability of traffic flow and road conditions), assessments conducted by a trained driving assessor in real traffic, rather than on a closed circuit, are considered to resemble everyday driving performance most closely (Withaar, Brouwer, & van Zomeren, 2000). This study considered only studies that used an on-road assessment in real traffic as the criterion measure.

Several clinical tests have yielded promising results in predicting on-road driving performance. They include the Stroke Drivers’ Screening Assessment (Nouri & Lincoln, 1992; Radford & Lincoln, 2004), the Cognitive Behavioral Driver’s Inventory (CBDI; Bouillon et al., 2006; Engum, Cron, Hulse, Pendergrass, & Lambert, 1988; Klavora et al., 2000), the Motor-Free Visual Perception Test (MVPT; Korner-Bitensky et al., 2000), the Trail-Making Test, Part B (Korner-Bitensky et al., 2000; Mazer, Korner-Bitensky, & Sofer, 1998) alone or as a component of the Gross Impairments Screening Battery for General Physical and Mental Abilities (Austroads, 2004), the Useful Field of View test (UFOV; Myers, Ball, Kalina, Roth, & Goode, 2000), a Maze Task (Snellgrove, 2005), and a battery of 12 neuropsychological tests (McKenna et al., 2004). Although it is difficult to compare results across studies because the studies varied in sample size and heterogeneity, methodologies, and statistical analyses, the predictive validity of the tests is best compared using statistical measures of sensitivity and specificity.

**Sensitivity** refers to the test’s ability to identify a problem when a problem truly exists (true positives), or in this case, a failed test result and unsafe driving or a failed on-road assessment. **Specificity** refers to the test’s ability to obtain a negative result when there is no problem (true negatives), or in this case, a pass on the test and safe driving performance (Portney & Watkins, 2000). Unfortunately, not all studies presented results in terms of sensitivity and specificity, nor can these measures be calculated from the reported results. Those studies that did report sensitivity and specificity are summarized in Table 1.

Analysis of Table 1 reveals that few tests have acceptable levels of both sensitivity and specificity. Tests that accurately identify unsafe drivers, such as the MVPT, also falsely identify many capable drivers as unsafe. Conversely, those tests that accurately identify safe drivers, such as the UFOV, Maze Task, and CBDI, also classify an unacceptably high number of incapable drivers as safe. Some tests, including the Trail-Making Test, Part B, do not accurately discriminate safe from unsafe drivers. Administration time for the tests varies depending on their complexity. Several of the tests are simple pen-and-paper tasks that can be completed within 5 min, such as the Trail-Making Test, Part B, and the MVPT. Other, more complex tests take 30 min to complete (UFOV and the battery of 12 tests), and the CBDI takes up to an hour to administer. The CBDI includes a series of four computerized tests of visual processing, subtests of the Wechsler Adult Intelligence Scale (Revised Picture Completion and Digit Symbol subtests); the Trail-Making Test, Parts A and B; a brake reaction time test; and an evaluation of visual fields (Bouillon et al., 2006). Although many of those tests have been used extensively in driving research, others have been used in only a single study, raising the need for replication (McKenna et al., 2004; Snellgrove, 2005). Most tests have been used with

### Table 1. Predictive Power of Off-Road Screening Tests

<table>
<thead>
<tr>
<th></th>
<th>CBDI&lt;sup&gt;a&lt;/sup&gt;</th>
<th>MVPT&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Trail-Making Test, Part B</th>
<th>UFOV</th>
<th>Battery of 12 Tests&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>As part of GRIMPS&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Alone&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CVA&lt;sup&gt;c&lt;/sup&gt; &gt;80 years&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mixed&lt;sup&gt;c&lt;/sup&gt; &gt;80 years&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MCI&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Maze Task&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>142&lt;sup&gt;c&lt;/sup&gt; + controls</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Sample size         | 172              | 84               | 84                        | 100  | 43                          |
| Diagnoses of sample |                  |                  |                          |      |                             |
| Sensitivity (fail test + fail on-road assessment; %) | Mixed | CVA | Mixed |
| Specificity (pass test + pass on-road assessment; %) | 62 | Mixed | 78 |
|                  | 61               | CVA              | Mixed                    |
|                  | 46               | >80 years        | >80 years                |
|                  | 67               | Mixed            | MCI                      |
|                  | 82               | Maze Task        | 82                       |
|                  | 71               |                   | 71                       |

Note. CBDI = Cognitive Behavioral Driver’s Inventory; MVPT = Motor-Free Visual Perception Test; GRIMPS = Gross Impairments Screening Battery for General Physical and Mental Abilities; UFOV = Useful Field of View test; CVA = cerebrovascular accident; MCI = mild cognitive impairment.

<sup>a</sup>Bouillon et al. (2006).
<sup>b</sup>AustRoads (2004).
<sup>c</sup>Myers et al. (2000).
<sup>d</sup>Snellgrove (2005).
<sup>e</sup>McKenna et al. (2004).
samples including clients with varied diagnoses, but the Stroke Drivers’ Screening Assessment was developed for use with clients following a stroke; attempts to validate the test with clients with traumatic brain injury yielded disappointing results (Radford, Lincoln, & Murray-Leslie, 2004). For the purpose of a screening test, many researchers advocate the use of a single test that can be used to identify at-risk drivers (Coleman, Bryant, Rapport, & Hanks, 2006), but when more comprehensive assessment is required, there may be the need for specific tests for particular diagnostic groups (Mazer et al., 2004).

Many of the most recent studies have used a combination of several tests to determine whether this improves predictive validity. The UFOV has been used with four neuropsychological tests in a sample of older drivers referred for a driving assessment (De Raedt & Pona Jærst–Kristoffersen, 2000). It also was used with a simulator task in a sample of community-dwelling seniors (Hoffman, McDowd, Atchley, & Dubinsky, 2005). Finally, UFOV was used with the Trail-Making Test, Part B, and a maze test in a small sample of clients with dementia (Whelihan, DiCarlo, & Paul, 2005). These combinations of tests resulted in little improvement in predictive validity. Similarly, in a large retrospective sample of clients referred for a driving assessment, the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975); the Trail-Making Test, Part B; and a reaction time score accurately identified unsafe drivers (sensitivity of 85%), but many capable drivers were falsely classified as unsafe (specificity of 39%), which was not an improvement in predictive validity (Kantor, Mauger, Richardson, & Unroe, 2004). An extensive battery of neuropsychological tests including the Trail-Making Test, Part B, was used in a study of clients with Alzheimer’s disease or Parkinson’s disease and matched controls (Grace et al., 2005). Although the Trail-Making Test, Part B, was significantly associated with unsafe driving performance, the relationship between the aggregate score and driving performance was not reported, and therefore it is not possible to determine whether the combination of tests improves predictive validity. These studies reveal that combining tests does not improve predictive validity to more clinically useful levels than the tests alone, a somewhat counterintuitive conclusion.

In summary, the research to date indicates that there is no single screening test that accurately predicts on-road driving performance. On the one hand, short, simple tests do not have the discriminative power required for a screening test, but on the other hand, more complex tests with better discriminative power are too time consuming for use as screening tests. Additionally, none of the tests demonstrates face validity for driving, with the possible exception of the Stroke Drivers’ Screening Assessment, which includes a road sign recognition task with some face validity (Nouri & Lincoln, 1992). This lack of face validity is a problem when at-risk drivers may be required to relinquish their drivers’ licenses on the basis of a test that appears to have little to do with driving.

In the absence of a simple and predictive screening test, most researchers have recommended a comprehensive assessment of fitness to drive including both off-road screening tests and an on-road driving assessment (Bouillon et al., 2006). In most Western countries, specialist occupational therapists who conduct driving assessments use a comprehensive approach. This is the case in Australia, where clients complete off-road screening tests of their visual and physical function; a pen-and-paper test of knowledge of road rules; and a test to assess their awareness of the driving environment, namely, the Visual Recognition Slide Test developed at the University of Sydney (VRST–USyd).

In assessing global awareness of the driving environment, the VRST–USyd is conceptually different from other tests currently being used and evaluated that assess component visual processing or cognitive skills. Developed more than 12 years ago, the VRST–USyd is widely used in Australia. This test provides useful information about a client’s ability to drive and anecdotally has good face validity. However, its psychometric properties have not been empirically analyzed. If this test accurately predicts those who are not fit to drive, then the cost and potential risk of on-road assessment could be avoided for many clients. Thus, the purpose of this study was to examine the construct and predictive validity and internal reliability of the VRST–USyd.

Method

Study Design

In this historical cohort study, the medical files of all eligible participants were reviewed. Demographic information including age, gender, and diagnosis and the results from the VRST–USyd, the on-road assessment, and the clinicians conducting the assessment were recorded.

Participants

Data were from all clients referred over a 10-year period to two major driving rehabilitation centers in a large metropolitan area in Australia—Calvary Rehabilitation and Geriatric Services and Driver Rehabilitation and Fleet Safety Services at the University of Sydney—and were analyzed retrospectively (n = 838). The majority (55.5%) of the participants were assessed at the University of Sydney’s Driver Rehabilitation and Fleet Safety Services. Data were available for 598 (71.4%) men and 240 (28.6%) women, with an age
range of 16 to 93 (\(M = 53, SD = 19.8\)). A wide spectrum of diagnoses was represented, as illustrated in Table 2.

To be included in the study, participants had to have completed an off-road assessment, including the VRST–USyd, and an on-road assessment. Data from those clients who used an interpreter for the assessment or who had receptive or expressive aphasia were excluded from the study because they were unable to complete the standard measures. Originally, data from 888 participants were included in the analysis; however, data from 50 were subsequently excluded because the raters, who tested clients early on and before procedures were described fully, were shown to be unreliable.

**Measures**

The VRST–USyd consists of 15 images of the same rotary, or roundabout (an alternative to a four-way stop), projected on a screen to simulate the view through a windshield, in which the number and position of pedestrians and vehicles vary (see Figure 1). Participants are asked to observe each image for 3 sec and, when the image has been removed from the screen, to report details about the position and direction of travel of each pedestrian and vehicle in the slide. The images vary in complexity, requiring participants to report from 4 to 16 elements. The participants complete three practice images to ensure that they understand the instructions. Performance is recorded as a score out of 164. The test takes 20 min to administer, and verbal responses are recorded by the clinician. If participants use hand gestures in providing answers, they are requested to respond verbally; if they are unable to do so, the clinician accepts the responses but notes that hand gestures were used. In the reported study, the test was administered according to the standard instructions and using the standard scoring sheet.

A 60-min on-road driving assessment was completed by the same clinician within 1 week of the clinical assessment. The vehicle had automatic transmission, power steering, and dual brakes. A professional driving instructor sitting in the passenger seat gave directions and monitored safety, and a registered driver-trained occupational therapist sitting behind the instructor recorded the participant’s driving performance. The use of two assessors in the vehicle is standard clinical practice in Australia and elsewhere (Fox, Bowden, & Smith, 1998). The on-road assessment began on quiet suburban streets to allow the participant to become familiar with the vehicle controls before progressing to more demanding driving environments. Each center used a standard route unless vehicle modifications (such as hand controls) were required, in which case the assessment continued in quiet suburban streets. The occupational therapist recorded and categorized participant performance under the headings of observation, speed control, planning and judgment, vehicle positioning, vehicle control, and reaction time. Any driving instructor interventions, such as using the dual brake or taking the steering wheel to prevent an accident, were recorded.

Consistent with standard clinical practice in Australia, the outcome of the assessment was agreed on by the driving instructor and the occupational therapist. The outcome was categorized as

- **Pass**: Safe and legal driving and no further intervention required;
- **Conditional pass**: Safe and legal driving with restrictions on license (e.g., automatic vehicle only, limiting driving distance or time);
- **Downgraded to a learner’s license**: Participant to undertake a series of driving lessons to learn to use modifications; or
- **Fail**: Failed to meet criteria for safe and legal driving and judged not to have the potential for improvement.

The criteria for failure were errors in all areas of driving or substantial errors in two or three areas, driving instructor intervention required to avoid a collision, or both. Participants who, after a series of driving lessons failed a second on-road assessment, were categorized as a **conceded fail**, and those who passed were categorized as a **conditional pass**. For the purposes

---

**Table 2. Participant Diagnoses (N = 838)**

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthopedic or spinal injury</td>
<td>320</td>
<td>38.2</td>
</tr>
<tr>
<td>Neurological conditions (including cerebrovascular accident, Parkinson’s disease, cerebral palsy)</td>
<td>149</td>
<td>17.8</td>
</tr>
<tr>
<td>Multiple diagnoses (diagnosis included both orthopedic and neurological or cognitive impairment)</td>
<td>145</td>
<td>17.3</td>
</tr>
<tr>
<td>Cognitive impairment (including dementia and Alzheimer’s disease)</td>
<td>104</td>
<td>12.4</td>
</tr>
<tr>
<td>Traumatic brain injury</td>
<td>54</td>
<td>6.4</td>
</tr>
<tr>
<td>Miscellaneous diagnoses (diagnosis does not fit into any other category; e.g., driving anxiety)</td>
<td>43</td>
<td>5.1</td>
</tr>
<tr>
<td>Vision impairment</td>
<td>23</td>
<td>2.7</td>
</tr>
</tbody>
</table>

---

**Figure 1. Example of Visual Recognition Slide Test–USyd item.**
of statistical analysis, the categories were collapsed into a pass (pass, conditional pass, or downgrade to learner’s license) or fail (fail or conceded fail) category. Those categorized as pass were able to continue driving legally.

A total of nine occupational therapists conducted the assessments, two at Calvary Rehabilitation and Geriatric Services and seven at Driver Rehabilitation and Fleet Safety Services. The therapists worked consistently with five driving instructors, two of whom worked at both centers. All occupational therapists had completed graduate training in driving assessment and rehabilitation and were registered with the state licensing authority to conduct driving assessments. Similarly, the driving instructors were qualified and registered and had completed additional training in rehabilitation techniques.

Data Analysis

Construct validity and internal reliability of VRST–USyd were examined using Rasch modeling (Bond & Fox, 2001) with the winssteps 3.58 program (P.O. Box 811322 Chicago, IL 60681-1322; Linacre, 2005). Rasch modeling constructs a linear measure from ordinal scores (by converting raw scores into scaled scores) and assesses the goodness of fit of both items and participants along a continuum (in this case describing awareness of the driving environment). An item and participant map is generated in which the items are arranged in order of difficulty and participants are arranged in order of competence.

The program generates two pairs of goodness-of-fit statistics, infit and outfit, expressed as mean square (Mnsq) and standardized fit statistics, which indicate how well the data from each item and participant conform to the assumptions of the Rasch model (i.e., easy items are easy for all people and more competent people will perform better on all items). For adequate fit to the model, Mnsq values of 1 ± 0.4 and standardized values of −2 to 2 were taken as acceptable for items and participants (Bond & Fox, 2001). Fit statistics below and above this level indicate too little or too much variation, respectively. Items with fit statistics in excess of these acceptable limits should be considered for removal from the test (Garratt, 2003). The desired fit is 95% and provides evidence of unidimensionality. Characteristics of participants who did not fit the model were investigated to determine whether there were any patterns that might explain why the test might be working differently for the participants. Additionally, point-measure correlation coefficients should be positive and large enough (> .50) to show a strong relationship between the item and the construct (Streiner & Norman, 1995).

A principal-components analysis of residuals was also conducted. When used with Rasch modeling, the purpose of principal-components analysis is to examine unidimensionality of the scale rather than to construct variables (as in standard factor analysis). If the empirical variance closely matches the modeled variance, then the test fits the expectations of the Rasch model; if the unexplained variance from the first factor is less than 3 eigenvalue units, then the result provides additional evidence that the test is unidimensional (Linacre, 2005).

Rasch modeling also produces reliability estimates for participants and items. A separation statistic provides evidence of internal reliability or the ability of the instrument to separate groups of participants into levels of ability. To conclude that differences in the measure are the result of real differences in the extent to which participants possess the trait (i.e., awareness of the driving environment) and not to error of measurement, the separation statistic should be 2.00 or greater. The participant reliability index (a Cronbach’s alpha equivalent) and the item reliability index (i.e., the replicability of placement of participants or items along the continuum) should be .80 or better.

To ensure that the test items function similarly for participants with different relevant characteristics (e.g., gender), we computed a differential item function (DIF) analysis using winssteps (Badia, Prieto, & Linacre, 2002). Items that were functioning significantly differently (p ≤ .05) for participants with specific characteristics were identified and could be considered for removal from the test. When more than one rater is scoring the test, it is possible to use DIF analysis of raters to determine whether any unexpectedly large differences exist in the way in which raters score each item on the test. This analysis provides evidence for interrater reliability.

Finally, we examined the predictive validity of the VRST–USyd. Data from all participants were randomly allocated to two groups. A discrimination score was determined for one group using descriptive statistics (i.e., specificity, sensitivity, positive and negative predictive values). The accuracy of the score was then tested in the second group. The most useful test will have a positive predictive value approaching 100%, minimizing the number of drivers who failed the clinical test but passed the on-road test (false positives).

Results

Construct Validity

All items had fit statistics within the acceptable range (Table 3), indicating that the items conformed to the assumptions of the Rasch model. Table 3 also documents the items arranged from hardest to easiest on the basis of the measure
Table 3. Item Statistics, With Items Arranged From Hardest to Easiest

<table>
<thead>
<tr>
<th>Item</th>
<th>Details to Be Recalled</th>
<th>Measure Score</th>
<th>Infit MnSq</th>
<th>Outfit MnSq</th>
<th>Point-Measure Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12</td>
<td>58.3</td>
<td>1.25</td>
<td>1.28</td>
<td>.82</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>54.7</td>
<td>0.99</td>
<td>0.99</td>
<td>.96</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>53.4</td>
<td>0.84</td>
<td>0.85</td>
<td>.84</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>53.2</td>
<td>0.90</td>
<td>0.90</td>
<td>.77</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>53.1</td>
<td>0.85</td>
<td>0.85</td>
<td>.87</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>53.1</td>
<td>0.96</td>
<td>0.96</td>
<td>.86</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>52.5</td>
<td>0.93</td>
<td>0.96</td>
<td>.78</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>51.2</td>
<td>0.92</td>
<td>0.93</td>
<td>.85</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>49.7</td>
<td>1.05</td>
<td>1.19</td>
<td>.74</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>45.9</td>
<td>1.06</td>
<td>1.04</td>
<td>.72</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>45.6</td>
<td>0.83</td>
<td>0.91</td>
<td>.73</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>45.2</td>
<td>0.94</td>
<td>0.91</td>
<td>.75</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>44.6</td>
<td>0.83</td>
<td>0.83</td>
<td>.64</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>42.5</td>
<td>1.31</td>
<td>1.11</td>
<td>.68</td>
</tr>
</tbody>
</table>

*Note. MnSq = mean square.*

Table 3 illustrates that all items had point measure correlation coefficients between .64 and .87. The map of items and drivers (Figure 2) illustrates the spread of driver ability from more capable to less capable in standard deviation units (shown as 1 and 2 standard deviations above and below the mean). Items were similarly located on the line from hardest to easiest as they compared with the drivers’ ability levels. The map demonstrates that the spread of participants was greater than the spread of items, indicating that the test does not assess the full range of driving competence. Specifically, the most competent participants were insufficiently assessed. However, those participants below the mean or who were least competent were assessed well, and this is the group of greatest concern. Some items did not contribute to the test’s ability to differentiate between participants with different levels of skill. That is, four groups of items appear at exactly the same level, and all seemed to contribute in similar ways to the construct. Therefore, one or more items in each grouping could be considered for removal from the test to reduce administration time.

Data from approximately 91% of participants conformed to the expectations of the Rasch model. Those whose data failed to fit the model were examined further to determine whether there were any patterns that would explain the results. Of the participants, 5% had unacceptably high fit statistics, indicating erratic responses, and 4% had unacceptably low fit statistics, indicating overly predictable responses (i.e., too Guttman-like) or more probably that their scores were within a narrow range. Although participants with low fit statistics are of less concern than those with unacceptably high fit statistics, all participants with responses that failed to conform to the expectations of the Rasch model shared some similar characteristics. They were younger (>60% were <65), demonstrated good awareness of the driving environment (>65% scored >110/164 on VRST–USyd), and drove safely (>80% passed the on-road assessment).

Principal-components analysis of residuals demonstrated that the empirical variance (81.6%) closely matched the modeled variance (81.3%). Moreover, the unexplained variance by the first factor was only 1.6 eigenvalue units. These provide further evidence that the test is unidimensional.

Reliability

The test separated the participants into four groups (a model separation of 4.61) with a participant reliability index (Cronbach’s alpha equivalent) of .95. The item reliability index was 1.00. These results provide evidence for the internal reliability of the VRST–USyd.

The DIF analysis revealed that there was no significant difference (p > .05) in the performance on test items associated with gender. When the DIF analysis was applied to raters, there was no significant difference (p > .05) in how the nine raters scored participants on any test item. Thus, the DIF analysis provided evidence for the interrater reliability of this instrument when used by trained raters according to standardized procedures.

Predictive Validity

The outcome of the on-road assessment was a *pass* for 363 participants (43%), a *conditional pass* for 143 participants (17%), a *downgraded license* for 195 participants (23%), and a *fail* for 137 participants (16%). When all participants who did not fail were included in the *pass* category, 701 participants (84%) passed.

A raw score cutoff of ≤95/164 (equivalent to a measure score of 54.6 generated WINSTEPS) yielded a sensitivity of 85% (unsafe drivers who failed VRST–USyd), a specificity of 87% (safe drivers who passed VRST–USyd), a positive predictive value of 55% (predicting those drivers who failed the on-road test), and a negative predictive value of 97% (predicting those drivers who passed the on-road test) in the first randomly allocated group (n = 447). When this cutoff score was tested on the second group (n = 391), sensitivity was 77%; specificity was 92%; and positive and negative predictive values were 67% and 95%, respectively. Because this cutoff score did not predict unsafe driving behavior as accurately as was expected on the basis of clinical experience, a content analysis of the participants’ files was conducted.

Decreased awareness of driving ability was a factor reported in the files of most participants who failed the on-road driving assessment. That is, clinicians documented that participants were unaware of driving errors when they were
provided with feedback or that, in response to questioning, these participants reported that they had driven well despite driving instructor interventions. This was interpreted as being indicative of reduced awareness of driving performance. Even though this judgment was made with the knowledge of on-road performance, we coded participants’ awareness as “intact” or “decreased” and added that to the predictive equation together with the VRST–USyd score.
The descriptive statistics for several cutoff scores were compared, including both the ordinal or raw scores and the linear or measure scores in scaled units (Table 4). A cutoff raw score of ≤110/164, together with decreased awareness, yielded the best statistics: sensitivity of 89%, specificity of 100%, and positive and negative predictive value of 100% and 98%, respectively, in the first randomly allocated group. In the second group, these values were 80%, 99%, 95%, and 96%, respectively. Only 3 participants (2.5%) predicted to fail the on-road driving test actually passed.

Discussion

The purpose of this study was to examine the construct and predictive validity and internal reliability of the VRST–USyd to determine how well test scores predict drivers who are unfit to drive. The findings yielded strong evidence for construct validity and internal reliability, indicating that the VRST–USyd measures a theoretical construct related to driving and necessary for safe driving, namely, awareness of the driving environment. Several redundant items (items at the same level of difficulty; see Figure 2) could be deleted, however, to create a more efficient test without losing any discriminative power. This analysis indicated no need for sophisticated modifications to the measure (Smith, 1996).

One of the strengths of the VRST–USyd is its clear face validity, which is the suitability of an assessment in a practical situation and the confidence the users have in its efficacy (Anastasi, 1988). Many of the cognitive tests that have been used in previous research do not have face validity for testing driving, compromising the use of test results as a basis for cancellation of driving licenses. Nonetheless, asking participants who lack awareness of their driving ability to accept a decision about license status without taking an on-road assessment may create problems.

This sample of drivers with medical conditions provided a suitable population for examining the VRST–USyd using Rasch analysis; Rasch analysis assumes that easy test items will be easy for all participants, regardless of diagnosis or other factors (i.e., the test is sample free) and that the most competent participants will perform the best on all items (Bond & Fox, 2001). The inclusion of participants with orthopedic and spinal conditions and no known cognitive or neurological deficits provided a group of participants expected to be competent and, thus, to perform better on all test items. This outcome was indeed the case; the most competent participants on the hierarchy were largely those without cognitive or neurological deficits. Additionally, it was important to examine the VRST–USyd’s construct validity with participants who had a variety of conditions for it to be demonstrated to be sample free.

Used alone, the VRST–USyd predicted drivers who would pass an on-road assessment but was not as accurate at predicting drivers who would fail the on-road assessment. That is, 40% (n = 75) of drivers who failed the VRST–USyd and were therefore predicted to fail the on-road assessment actually passed it. With a cutoff score of 95/164, the VRST–USyd has a sensitivity of 81% (those who fail VRST–USyd and the on-road assessment) and a specificity of 90% (those who pass VRST–USyd and the on-road assessment), results that compare favorably with published results for other tests used for screening at-risk drivers (see Table 1). With good face validity and predictive accuracy, the VRST–USyd could be used to screen at-risk drivers internationally. The images of a rotary are comparable to a four-way stop, and with minor photographic modifications, the difference in the side of the road used for driving could be addressed.

What emerged in this study is the presence of another factor contributing to safe driving (which is obvious after observing individuals’ driving performance); namely, awareness of driving performance. In this retrospective study, awareness was subjectively reported after the driving assessment. Nonetheless, when a judgment of participants’ awareness was used with a discrimination score of 110/164 on the VRST–USyd, the test’s predictive accuracy increased, and 98% of participants who were predicted to fail the on-road assessment did fail. The accuracy of this prediction raises the question of whether it is possible to measure awareness objectively before observing driving performance.

Table 4. Descriptive Statistics for Cutoff Scores on Visual Recognition Slide Test–USyd With Reduced Awareness

<table>
<thead>
<tr>
<th>Cutoff Score (Raw Score)</th>
<th>100</th>
<th>105</th>
<th>110</th>
<th>115</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cutoff Score (Measure Score in Scaled Units)</td>
<td>56</td>
<td>57.3</td>
<td>58.8</td>
<td>60.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptive Statistic</th>
<th>Group 1 (n = 447)</th>
<th>Group 2 (n = 391)</th>
<th>Total (n = 838)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>88</td>
<td>71</td>
<td>80</td>
</tr>
<tr>
<td>Specificity</td>
<td>100</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>100</td>
<td>98</td>
<td>95</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>98</td>
<td>94</td>
<td>95</td>
</tr>
</tbody>
</table>

Note: Values are percentages.
The importance of awareness of driving performance for safe driving has been documented in the literature for healthy older drivers (Marottoli & Richardson, 1998), drivers with dementia (Cotrell & Wild, 1999; Wild & Cotrell, 2003), and drivers with traumatic brain injury (Brooks & Hawley, 2005; Huchler et al., 2001), although the precise nature of the relationship needs further study. From a theoretical perspective, awareness underpins the widely accepted hierarchical model of driving performance (Michon, 1985). Drivers are unable to make strategic adjustments to their driving, such as avoidance of complex traffic situations, if they are unaware of the need to do so (Ball et al., 1998). Similarly, a recently documented, theoretically based, multifactorial model suggested that safe driving behavior requires accurate self-monitoring of cognitive, sensory, and physical function (Anstey, Wood, Lord, & Walker, 2005). Figure 3 depicts a flowchart that yields an accurate prediction of on-road performance, developed on the basis of this study’s results together with previous models of safe driving performance. The proposed model requires further research for verification.

Study Limitations

Retrospective studies reflect clinical reality when it is not always possible to control for all the potential variables, and limitations common to retrospective studies in general were true of this study (Korner-Bitensky et al., 2000). Moreover, selection bias is inherent in the sample because the sample represents only participants who were referred for driving assessment. This group does represent at-risk drivers identified because of medical conditions; however, no normative sample was available with which to compare the participants undergoing driving assessment. Additionally, the on-road assessors were not “blind” to the participants’ performance on VRST–USyd because in the clinical setting, the purpose of completing the off-road assessments is to inform assessors of deficits that may affect on-road performance. Because the study was retrospective, however, the assessors were unaware of the purpose of the study (or, indeed, that any study would be conducted) at the time they conducted the assessment (Bouillon et al., 2006). Lack of information concerning the reliability of the outcome measure (i.e., on-road driving performance) was a further limitation of the study. The retrospective nature of this study, however, enabled a large sample size, a characteristic rarely present in studies of driving performance.

Another important limitation of this study was that the clinicians reported the level of participants’ awareness on the basis of their performance on both the off-road and the on-road tests rather than on the basis of their performance in the off-road assessment alone. If awareness is to be used along with the VRST–USyd as a predictor of driving performance, then an accurate method of measuring the construct before the on-road assessment is required (Howorth & Saper, 2003; Sherer, Hart, & Nick, 2003). The subjectivity of the reporting of awareness in this study is problematic, but the fact that it was such a clinically striking factor is indicative of the need for further objective research.

Because data from VRST–USyd have not previously been analyzed, no traditional interrater reliability statistics (e.g., intraclass correlation coefficients) are available for the test, and it was not possible in this retrospective analysis to address this limitation using standard statistics. DIF analysis within Rasch modeling, however, provided more detailed information than would be obtained through traditional measurement. That is, we were able to discern that raters did not differ significantly in their use of any item. In using intraclass correlation coefficients or other similar measures, we would know only how overall ratings correlated.

Several limitations of the test itself could not be addressed in this study. The VRST–USyd is heavily reliant on language skills. If clients must use an interpreter, then the test is not likely to be valid. Similarly, if clients have receptive or expressive language disorders, then the test cannot be administered according to the standard instructions. Future modifications of the test will need to eliminate the need for verbal responses. The influence of socioeconomic or educational status on test performance has not been investigated, although it is presumed that these factors varied widely in the current sample.
Conclusion

Rasch analysis of the VRST–USyd provided evidence for the test’s construct validity and internal reliability. The test predicted those drivers who were unsafe to drive with a level of accuracy rivaling that of other available tests. Coupled with a clinician’s judgment of participants’ awareness of their driving ability, made after the on-road test, the predictive accuracy increased. Until a formal measure of awareness is developed that can be used before the on-road assessment, however, use of the test as a sole basis for license cancellation is not supported. The predictive validity of the test prospectively and a suitable measure of awareness require further research. ▲

Acknowledgments

The authors thank the clinicians at Driver Rehabilitation and Fleet Safety Services and Calvary Rehabilitation and Geriatric Services for their support, Alexandra Wu and Virginia Deal for their assistance in data collection and entry, and Tim Luckett for his assistance with manuscript preparation.

The research described in this article was presented in part as a research poster at the 14th Congress of the World Federation of Occupational Therapists, July 2006, Sydney, New South Wales, Australia, and in part as a paper at the Australian Consortium for Social and Political Research Incorporated Conference, December 2006, Sydney, New South Wales, Australia.

References


