Evaluation of a roadside impairment test device using alcohol

Philip R. Dixona,*, Tony Clarka, Brian Tipladyb

a Home Office Scientific Development Branch, Sandridge, St Albans, AL4 9HQ, United Kingdom
b Department of Anaesthesia, Critical Care, and Pain Medicine, University of Edinburgh, EH16 4SA, United Kingdom

Abstract

The main objective of this study was to compare the performance of a portable impairment test device known as roadside impairment testing apparatus (RITA) with the field impairment tests (FIT) that are used at the roadside by UK police. One hundred and twenty two healthy volunteers aged 18–70 years took part in this two-period crossover evaluation. The volunteers received a dose of alcohol and placebo, in the form of a drink, on separate days. Doses were calculated to produce blood alcohol concentrations of 90 mg/100 ml and RITA and FIT testing was carried out between 30 and 75 min post-drink. FIT was found to have a diagnostic accuracy of 62.7%. However, there was a substantial age effect for FIT scores, with volunteers aged over 40 showing failure rates on placebo similar to the failure rates on alcohol of younger volunteers. The accuracy of RITA was between 66 and 70%, not significantly higher than that of FIT. However, RITA did not show a marked age effect. Advantageously, this could result in fewer false positives being recorded if RITA were deployed at the roadside. Horizontal gaze nystagmus (HGN) was also investigated and posted an accuracy of 74%. The inclusion of HGN as one component of a UK roadside impairment test battery warrants further exploration with other drugs.

1. Introduction

Driving under the influence of drugs is an ever increasing problem in modern society. If allowed to go unchecked, it may become as prevalent as ‘drink-driving’ which, in the UK, is a factor in almost 20% of road fatalities each year (Tunbridge et al., 2000). At present there is no reliable method equivalent to the alcohol breathalyser to detect drug-driving. The UK police currently use field impairment tests (FIT) at the roadside to assess drivers suspected of driving under the influence of drugs. The FIT battery (Fleming and Stewart, 1998; Stark et al., 2002) is derived from the US standardised field sobriety tests (SFST) developed and clinically validated by Burns and Moskowitz (1977). The only previously reported scientific evaluation of FIT was conducted at Glasgow University by Oliver et al. (2006), and showed an overall diagnostic accuracy of 66%.

In 2002 the Home Office Scientific Development Branch (HOSDB) began work to develop a roadside screening device capable of giving an objective and quantitative measure of impairment due to drugs. The project was organised into three phases. In the first phase, 11 impairment tests were developed to run on a variety of handheld computers. These were evaluated in small-scale pilot studies the results of which have been reported previously (Boyle et al., 2004; Degia et al., 2006; Dixon et al., 2004; Pearce et al., 2004; Tiplady et al., 2005).

In phase two the six most promising tests identified in the pilot studies were programmed onto a single handheld computer. The resulting prototype became known as RITA, an acronym for roadside impairment testing apparatus. Although RITA was designed to be capable of measuring impairment due to all drugs its first major evaluation, the subject of this paper, was conducted using alcohol. The rationale was that if RITA was unable to demonstrate the well-documented impairment due to alcohol, then it would be unlikely to be any more discriminating for other depressant drugs. The principal objectives of the study were to:

(a) evaluate the ability of RITA to discriminate between individuals who had taken an impairing dose of alcohol with individuals given a placebo treatment;
(b) find the best combination of three tests that would combine to give a practical roadside battery of no more than 12 min duration;
(c) compare the sensitivity, reliability, accuracy of RITA with those of FIT.

2. Methodology

2.1. Performance tests

2.1.1. Field impairment tests

The FIT battery takes between 10 and 12 min to perform and comprises five tests:
Pupillary examination: Pupillary diameters are measured using a gauge of calibrated pupil sizes held up to each side of the subject’s face in turn. A pupil size between 3 and 6.5 mm is considered normal. Reddening and watering of the eyes are also recorded.

Romberg test: The subject stands upright with head tilted backwards and eyes closed whilst estimating the duration of 30 s. Swaying, stepping, raising of the arm(s), opening the eyes and inability to follow the instructions are also recorded.

Walk and turn: The subject takes nine steps heel to toe along a line on the ground, then turns and walks back, while counting steps out loud. Lack of balance during instructions, starting too soon, stepping off the line, stopping, raising arms to balance, taking the wrong number of steps, incorrect counting and improper turns are recorded.

One leg stand: The subject stands on one leg with the foot of the other leg raised 6–8 in. off the ground for 30 s while counting “one thousand and one, one thousand and two.” The procedure is then repeated with the other leg. Swaying, hopping, putting a foot down, raising of arms and incorrect counting are recorded.

Finger to nose test: With head tilted back and eyes shut, the subject touches the tip of the nose with the tip of the right or left index finger in accordance with spoken instructions. Missing the nose, using the wrong finger, swaying and hesitation are recorded.

All research staff who conducted FIT tests during the trial had been trained by an accredited national FIT instructor and so administered the tests in accordance with standard UK police procedures. There are no definitive pass/fail criteria for FIT and whether a subject is considered to be fit or unfit to drive comes down to the best judgement of the officer conducting the tests. For validation purposes, and to ensure consistency of testing amongst different staff, a random sample of 10% of the FIT tests administered during the trial were independently witnessed by either a trained police officer or the same accredited FIT instructor. A more detailed description of the FIT battery has been given by Fleming and Stewart (1998) and Collier (2004).

2.1.2. Nystagmus

Nystagmus is the rapid oscillation (saccades) of the pupil which occurs at a certain angle from the straight ahead position. It is induced by a number of drugs, including alcohol (Good and Augsburger, 1986). Nystagmus is not currently part of FIT but is incorporated in the standardised field sobriety tests used by US police officers (McKnight et al., 2002). The subject is asked to follow the lateral and vertical motion of a pen, or similar object, held about 12 in. in front of the face. Any resulting horizontal gaze nystagmus (HGN) or vertical gaze nystagmus (VGN) is then recorded.

In this study, the nystagmus and pupillary examinations were conducted under controlled artificial lighting. A lightmeter was used to ensure consistency of illumination between test sessions.

2.1.3. Roadside impairment testing apparatus

The RITA system was implemented on a Sony VAIO U71 handheld computer with a 127 mm diagonal touch-sensitive screen running Windows XP Tablet Edition. All the impairment tests and data collection software were programmed ‘in-house’ at HOSDB. The suite of tests takes approximately 25 min, and includes the following six tests:

Critical tracking task (CTT): This combines tracking and a reaction time task. The volunteer tracks a spot on the screen with a stylus. The spot moves along a single vertical dimension, according to a composite sinusoidal function. The motion of the spot becomes increasingly more complex as the test proceeds. The reaction time task requires the volunteer to respond to symbols appearing on the screen by pressing a button (Degia et al., 2006).

The task period is divided into three equal periods, labelled A, B, and C, for which the following performance parameters are recorded:

- **Tracking error:** Root mean square deviation from target over the entire period.
- **Post-stimulus tracking error:** A single error measurement is recorded 10 ms after the response or 10 ms after the stimulus disappeared (it is displayed for 2 s if not responded to).
- **Reaction time:** Mean reaction time for responses to the reaction-time task stimulus.

Length estimation (LE): In each trial, a vertical line is presented, with a gap to its right. The volunteer presses the right button (Yes) to indicate that the line fits through the gap, the left button (No) if not. Feedback is given for correct or incorrect ‘Yes’ responses, but not for ‘No’ responses. The size of the gap, the difference between the line length and the gap length, and the distance between the line and the gap all vary between trials (Tippladty et al., 2005).

The performance measures are as follows:

- **Overall reaction time:** Mean reaction time of all correct responses.
- **Total number correct:** Number of correct responses to both “fit” and “no-fit” stimuli.
- **Total no-fit number correct:** Number of correct responses to “no-fit” stimuli.

Paired associate learning (PAL): Two shapes appear on the screen; one on the left, the other on the right. A series of these shapes then appears in the centre of the screen, and the volunteer presses the left or right button as quickly as possible to indicate the side on which the shape initially appeared. If an incorrect response is made, the pair of shapes appears again. After eight trials using two shapes, a second pair of shapes appears, and then single shapes now drawn from the set of four appear in the centre of the screen. The volunteer continues to respond in the same way. This continues until eight shapes are in the response set (Tippladty et al., 2005).

The performance measures are as follows:

- **Overall reaction time:** Mean reaction time of all correct responses.
- **Total errors:** Number of incorrect responses (responses cannot be missed in this task).

Sustained attention to response test (SART): A series of 243 stimuli in the form of road sign images are presented on the screen at a rate of one every 1105 ms. Volunteers are instructed to respond to each stimulus as quickly as possible with a button press unless presented with a “No Stopping” sign in which case they should not respond. Feedback is given to incorrect responses (Degia et al., 2006).

The performance measures are as follows:

- **Overall reaction time:** Mean reaction time of all correct responses.
- **Number incorrect:** Number of responses to “no stopping” stimulus (i.e. false positives).
- **Number missed:** Number of failures to respond to target stimuli.

Choice reaction time (CRT): The screen display has six response areas arranged in an arc, with two rows of lights above them. The outer row of lights are green prompt lights, and indicate which of the inner red stimulus lights may appear. The subject is instructed to hold a stylus on the “Rest” button until a red light appears, then move to tap the corresponding button as quickly as possible. Forty-eight responses are recorded, with the number of prompt lights varying from 1 (simple reaction time) to 6 (Pearce et al., 2004).
The performance measures are as follows:

- **Total reaction time**: Mean time elapsed between the appearance of the stimulus and the tap on the correct target button.
- **Off reaction time**: Mean time elapsed between the appearance of the stimulus and lifting of the stylus from the start button.

Reaction times were calculated with the slowest response in each category, removed.

**Arrow flanker task (AFT)**: Five symbols appear on the screen. The central symbol is an arrow, pointing either right or left, and the task is to press the button corresponding to the direction of the central arrow as quickly as possible. The other symbols can be either congruent—arrows pointing in the same direction as the central arrow; non-congruent—arrows pointing in the opposite direction to the central arrow; or neutral-squares (Tiplady et al., 2005).

The performance measures are as follows:

- **Overall reaction time**: Mean reaction time of all correct responses.
- **Total errors**: Number of incorrect or missed responses.

### 2.2. Overall study design

The study used a two-period crossover design, comparing alcohol with placebo. Following a dose-finding session, volunteers took part in two testing sessions, at least 2 days apart. Treatments were administered double-blind in a randomised order, stratified by age and gender.

### 2.3. Volunteers

A total of 60 male and 62 female volunteers were recruited. At least nine volunteers (protocol target 10) of each sex were recruited in each of the age ranges 18–24; 25–30; 31–40; 41–50; 51–60, and 61–70. The overall age range was 18–68 for both males and females. All were in good health, held a full driving licence and were neither teetotalers nor excessive consumers of alcohol.

Volunteers were paid a small compensatory sum for their participation in the trial. They all gave written consent to take part in the study, which was approved by the Lothian Research Ethics Committee.

### 2.4. Study treatment

All subjects first took part in a dose-finding session in which they received a dose of alcohol calculated using a modified version of the Widmark equation (Tiplady et al., 2005). Linear modelling was used to calculate doses for the main test sessions for each subject from the peak blood alcohol concentrations (BAC) recorded in the dose-finding sessions. The aim was to achieve a target BAC of 90 mg/100 ml with reduced variance in the main test sessions. A target BAC of 90 mg/100 ml was chosen as it was considered to be a level at which most subjects would be unfit to drive and could be achieved by administering an ethically acceptable dose of alcohol. This level is just above the UK legal limit for drink-driving of 80 mg/100 ml.

### 2.5. Blinding procedure

The active drink consisted of vodka mixed with an equal volume of orange juice. The placebo drink used an equivalent volume of water, also made up with an equal volume of orange juice. The taste of the drink was masked by giving a lozenge containing benzocaine (Tyrozet) to suck before the drink, and by spraying the drink with a peppermint breath freshener.

### 2.6. Treatment plan

On each test day volunteers initially performed a shortened version of the test battery for familiarisation purposes. They then consumed the drink, 10 min being allowed for this. The test battery and FIT were then performed twice, between 30 and 75 min after the start of the drink, and between 90 and 135 min post-drink. Volunteers were breathalysed at each of the above time points using a Lion SD–400 breathalyser. A multiplier of 2300 was used to convert breath alcohol concentration (BrAC) to BAC. The test order was randomised between subjects.

### 2.7. Statistical analysis

#### 2.7.1. Analysis of Variance

Following exploratory descriptive statistical analysis, scores from the individual test measures from Run 1 were subjected to a two-way within-subjects Analysis of Variance (ANOVA). A model including Treatment (alcohol or placebo), Day (session I or session II) and the Treatment × Day interaction was developed to identify any asymmetric transfer effects. This potential problem is discussed further in the Results Section 3.5. Multivariate Analysis of Variance (MANOVA) was also performed to investigate the best test combinations based upon differences in means. A $p$-value of 0.05 or less was deemed to be significant.

#### 2.7.2. Linear discriminant function analysis

The time available for roadside impairment testing is limited in practice to about 10–12 min, which allows a maximum of three tests to be included. It was also necessary to combine test scores into a single measure to allow an ‘impaired/unimpaired’ decision, analogous to that of FIT that would correctly classify as many individuals as possible. Linear discriminant function analysis (LDFA) was used in this study as it was considered to be the most appropriate method to determine the optimum test selection and weighting. The method is mathematically equivalent to MANOVA but in LDFA the emphasis is on combining predictors so that cases can be classified as accurately as possible into groups. A comprehensive description of the technique has been given by Tabachnick and Fidell (2001). LDFA is known to be particularly sensitive to the presence of outliers and the violation of homogeneity of variance–covariance matrices and so these were initially checked by exploratory data analysis. Data displaying heterogeneity were transformed using standard functions as described in Tabachnick and Fidell (2001).

LDFA was performed on all 20 possible 3-test combinations from the 6 in the initial RITA battery. The two classification groups were defined as ‘impaired’ and ‘unimpaired’ according to treatment. Generally, a combination of error scores and reaction times from each test were used in forming the discriminant function for each test combination. These can be regarded as being analogous to regression equations, with each coefficient found by maximising the difference between groups relative to within groups. A subject’s discriminant score was found by multiplying each test variable score by its respective coefficient and summing over all terms. By averaging over all discriminant scores within a group, the mean or centroid was calculated for reduced space. For two equally sized groups, the optimal cutting point for classification is midway between the two centroids. Positive scores represented an impaired classification while negative scores were classified as unimpaired. The overall success of the classification scheme was then evaluated in a confusion matrix where a percentage correct classification was produced. Unsurprisingly, LDFA often yields optimistic results because the same data are used both in model construction and classification. Therefore, as a further check of robustness, internal cross-validation was also performed for the most successful test combinations. This
Table 1
Proportions of pass/fail FIT scores after alcohol and placebo (n = 122).

<table>
<thead>
<tr>
<th>FIT score</th>
<th>All Pass</th>
<th>Fail</th>
<th>Male Pass</th>
<th>Fail</th>
<th>Female Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>65</td>
<td>57</td>
<td>35</td>
<td>25</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>Alcohol</td>
<td>34</td>
<td>88</td>
<td>20</td>
<td>40</td>
<td>14</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>145</td>
<td>55</td>
<td>65</td>
<td>44</td>
<td>80</td>
</tr>
<tr>
<td>Accuracy (%) [95% CI]</td>
<td>62.7 [57.7–67.7]</td>
<td>62.5 [54.8–70.2]</td>
<td>62.9 [56.4–69.4]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. Time course of breathalyser readings for the dose-finding and main test sessions.

was done by randomly dividing the sample into two equally sized smaller samples. The first sample was used to develop the function while the second was fed back into the classification function to assess correct classification. This process was repeated three times and a mean correct classification was recorded.

All procedures were completed in SPSSv12 or SASv9.1.

3. Results

3.1. Blood alcohol concentrations

Fig. 1 shows the time course of breathalyser readings from the dose-finding and main test sessions. BAC levels were highest between 30 and 75 min after the drink which is why this time band was used for the first period of FIT and RITA testing. The mean peak BAC in the main test sessions, from breathalyser readings at 30 and 75 min, was 86.7 mg/100 ml (S.D. 18.4), very close to the target of 90 mg/100 ml. The difference in BAC between gender and the different age groups was small and insignificant. Unfortunately, the dose adjustment based on the initial dose finding session, did not result in an overall reduction in variance in recorded BAC levels.

3.2. Field impairment testing

A pass/fail judgement was made for each FIT test based on the observations described in Section 2.1.1. These pass/fail judgements corresponded to whether the investigator considered the subject would have been deemed ‘fit’ or ‘unfit’ to drive had they been tested at the roadside. The results, broken down by sex, are shown in Table 1.

3.3. Nystagmus

Table 2 presents cases in which HGN was observed. Only HGN data are shown as VGN was relatively uncommon, and was never seen without HGN. There was little variation in the true positive rates for HGN across the six age groups. However, only 6 of the 25 false positives were aged over 40 years.

3.4. Individual test scores for RITA

The results from the two-way within-subjects ANOVAs are shown in Table 3. Every test had at least one measure showing significant effects of alcohol, as expected. The highest F-ratios were recorded for SART (number incorrect) and CRT (Off reaction time) indicating that these two components were most sensitive to the treatment.

3.5. Asymmetrical crossover effects

One potential problem with a within-subjects crossover trial is that the observed treatment effects may be dependent on the order of administration. Such asymmetrical transfer between treatment conditions can result in misleading and sometimes incorrect conclusions (Millar, 1983). Asymmetry problems were apparent in two tests: PAL and, to a lesser extent, LE where there was a significant Day × Treatment interaction (Table 3, Fig. 2). In both cases performance was impaired by alcohol relative to placebo on Day 1, but on Day 2 the alcohol performance was more similar to placebo. Placebo performance was relatively stable for both measures.

Several of the measures in the Critical Tracking Task showed order (Day) effects but not a Day × Treatment interaction. This is illustrated in Fig. 3 for tracking error in the last third of the test. Similar checks were conducted for each of the five FIT tests but no significant asymmetrical transfer effects were found.

3.6. LDFA results

Initially, multivariate analysis was performed using Day 1 data in order to avoid the problems with asymmetry discussed in Section 3.5. The best resulting LDFA classifications (accuracies) are presented in Table 4 alongside their corresponding internal N-1 cross-validations. The small drop in classification accuracy after
Table 3
ANOVA results for individual test scores.

<table>
<thead>
<tr>
<th>Test/measure</th>
<th>Treatment</th>
<th></th>
<th></th>
<th></th>
<th>Day</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Interaction</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fp</td>
<td>p</td>
<td>Fp</td>
<td>p</td>
<td>Fp</td>
<td>p</td>
<td>Fp</td>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arrow flankers (AFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall reaction time</td>
<td>7.81</td>
<td>0.0061</td>
<td>2.75</td>
<td>0.1000</td>
<td>2.77</td>
<td>0.0985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total errors</td>
<td>9.01</td>
<td>0.0033</td>
<td>1.00</td>
<td>0.3191</td>
<td>0.73</td>
<td>0.3961</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choice reaction time (CRT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total reaction time</td>
<td>2.09</td>
<td>0.1510</td>
<td>2.25</td>
<td>0.1361</td>
<td>2.98</td>
<td>0.0870</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off reaction time</td>
<td>56.35</td>
<td>&lt;0.0001</td>
<td>0.11</td>
<td>0.7460</td>
<td>0.98</td>
<td>0.3235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length estimation (LE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall reaction time</td>
<td>6.42</td>
<td>0.0126</td>
<td>15.10</td>
<td>0.0002</td>
<td>5.20</td>
<td>0.0244</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number correct</td>
<td>7.05</td>
<td>0.0090</td>
<td>8.32</td>
<td>0.0046</td>
<td>11.22</td>
<td>0.0011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total no-fit number correct</td>
<td>0.07</td>
<td>0.7860</td>
<td>0.87</td>
<td>0.4160</td>
<td>3.78</td>
<td>0.0543</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paired associate learning (PAL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall reaction time</td>
<td>4.21</td>
<td>0.0424</td>
<td>52.27</td>
<td>&lt;0.0001</td>
<td>0.63</td>
<td>0.4292</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total errors</td>
<td>21.24</td>
<td>&lt;0.0001</td>
<td>11.68</td>
<td>0.0090</td>
<td>11.84</td>
<td>0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sustained attention to response test (SART)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall reaction time</td>
<td>3.02</td>
<td>0.0849</td>
<td>3.25</td>
<td>0.0742</td>
<td>0.11</td>
<td>0.7451</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number incorrect</td>
<td>66.39</td>
<td>&lt;0.0001</td>
<td>0.64</td>
<td>0.4260</td>
<td>0.30</td>
<td>0.5850</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number missed</td>
<td>0.32</td>
<td>0.5731</td>
<td>1.14</td>
<td>0.2885</td>
<td>0.68</td>
<td>0.4106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical tracking task (CTT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Tracking error</td>
<td>0.83</td>
<td>0.3650</td>
<td>0.27</td>
<td>0.6053</td>
<td>1.18</td>
<td>0.2793</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Post-stimulus tracking error</td>
<td>0.05</td>
<td>0.8305</td>
<td>2.44</td>
<td>0.1209</td>
<td>0.05</td>
<td>0.8273</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Reaction time</td>
<td>9.11</td>
<td>0.0031</td>
<td>21.14</td>
<td>&lt;0.0001</td>
<td>1.11</td>
<td>0.2937</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Tracking error</td>
<td>5.07</td>
<td>0.0262</td>
<td>0.53</td>
<td>0.4700</td>
<td>0.06</td>
<td>0.8022</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Post-stimulus tracking error</td>
<td>2.61</td>
<td>0.1091</td>
<td>0.08</td>
<td>0.7759</td>
<td>1.19</td>
<td>0.2774</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Reaction time</td>
<td>22.77</td>
<td>&lt;0.0001</td>
<td>9.01</td>
<td>0.0033</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Tracking error</td>
<td>18.63</td>
<td>&lt;0.0001</td>
<td>18.05</td>
<td>&lt;0.0001</td>
<td>0.05</td>
<td>0.8152</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Post-stimulus tracking error</td>
<td>2.05</td>
<td>0.1551</td>
<td>1.69</td>
<td>0.1956</td>
<td>1.68</td>
<td>0.1972</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Reaction time</td>
<td>21.59</td>
<td>&lt;0.0001</td>
<td>3.76</td>
<td>0.0550</td>
<td>1.02</td>
<td>0.3145</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Interaction between Treatment and Day for paired associate learning number of errors. Data from Run 1 (30–75 min post-drink) are shown as mean ± standard deviation. Light bars: placebo; dark bars: alcohol.

Fig. 3. Period effect for critical tracking task tracking error C (last third of test). Data from Run 1 (30–75 min post-drink) are shown as mean ± standard deviation. Light bars: placebo; dark bars: alcohol.

cross-validation demonstrates that the correct classification results are robust.

It was found that the SART + CRT + PAL test combination was the most efficient in classifying subjects as impaired/unimpaired with an overall accuracy of 74.6% (73.7%, cross-validated). However, when the same analytical approach was taken using Day 2 data, the mean cross-validated classification accuracy for the SART + CRT + PAL combination was found to be 61.8%. This result reflects the reduction in discriminatory capability of the PAL test due to the asymmetry problems discussed in Section 3.5.

An LDFA model was constructed for the SART + CRT combination alone. The cross-validated classification accuracy was found to be 67.4% for Day 1 and 66% for Day 2. In Table 5 the key results from all the linear discriminant function analyses are compared

Table 4
Summary of LDFA classifications for Day 1.

<table>
<thead>
<tr>
<th>Test combination</th>
<th>Mean correct classification %</th>
<th>Cross-validated %</th>
</tr>
</thead>
<tbody>
<tr>
<td>SART + PAL + LE</td>
<td>72.1</td>
<td>69.7</td>
</tr>
<tr>
<td>SART + CRT + PAL</td>
<td>74.6</td>
<td>73.7</td>
</tr>
<tr>
<td>SART + CTT + PAL</td>
<td>73.8</td>
<td>71.3</td>
</tr>
<tr>
<td>SART + AFT + PAL</td>
<td>69.7</td>
<td>68.9</td>
</tr>
<tr>
<td>SART + CTT</td>
<td>67.2</td>
<td>66.4</td>
</tr>
<tr>
<td>SART + CRT + AFT</td>
<td>71.3</td>
<td>67.7</td>
</tr>
</tbody>
</table>
with the accuracy found for FIT. The best overall accuracy for a RITA score combination that is consistent over the 2 days is about 66%, compared to the overall accuracy of FIT of about 63%.

### 3.7. Comparison of RITA and FIT results

Fig. 4 compares the failure rates recorded for the best cross-validated RITA model (SART + CRT + PAL) and FIT by treatment and age on Day 1. This combination shows the best discrimination for RITA, with an accuracy of 73.7% compared to 59.8% for FIT (Table 5). From Fig. 4 it can be seen that FIT had the higher false positive rate, and that this difference was most pronounced in subjects aged over 40 years. When the analysis was restricted to between-subjects under the age of 40, the difference in accuracies for FIT and RITA was marginal; 69.3% compared with 71%.

### 3.8. Time after ingestion of alcohol

The main analysis described above has used data only from the first post-drink time point (starting 30 min after the drink) when the BACs were highest and the subjects were expected to be most impaired. Data from the second time-point (starting 90 min after the drink) were compared with data from the first time-point. This comparison showed that the proportion of volunteers failing the FIT tests decreased for both alcohol and placebo treatments at the second time-point. In contrast, the failure rates recorded for RITA remained almost constant at both time points. Overall, the diagnostic accuracies calculated for FIT and RITA at the second assessment were not significantly different from those presented earlier for the first assessment.

4. Discussion

The sample we studied was intended to cover the full age range from 18 to 70 with an even spread. This was achieved, with at least 9 volunteers in each of the 12 age/gender groups, no obvious “clumping”, and a minimum age of 18 and a maximum of 68 in each of the treatment groups. Although the variability of blood alcohol concentration was greater than we had hoped for, the mean was close to the target, and a BAC two standard deviations below the sample mean was still above 50 mg/100 ml, a level at which measurable impairment would be expected. Thus the study groups were appropriate for comparing impairments assessed by different candidate methods for field assessment.

The maximum percentage of correct classifications given by RITA on Day 1 was 76%. This used all the tests in the battery, and is thus not a realistic figure in practise, but gives an upper bound to the available discriminatory power. A more realistic figure for a useful subset of tests is 66–70%. This is somewhat better than the 63% accuracy obtained for FIT in the present study, and the 66% recorded from the Glasgow study (Oliver et al., 2006), but does not represent a major advance in accuracy of assessment.

The essential problem of accurately identifying drug-impaired drivers is the wide range of performance found in those who are not affected by drugs. This is illustrated in Fig. 5, which shows the distributions of composite scores under the two treatment conditions. The range of scores in the placebo condition is widely spread, and the overlap between alcohol and placebo very considerable.

Thus while all tests in the battery were effective at discriminating alcohol from placebo, combination of measures from the various tests did not give a clear demarcation between impaired and non-impaired performance, even though the composite measures included a variety of distinct elements including measures of speed and accuracy of performance, and also different types of error score, including false positives and false negatives. Having both speed and accuracy represented in the summary measure is important because of the possibility of trade-offs. Some people may be impaired mainly in speed of performance, others in accuracy. Trade-offs may be voluntary or involuntary. A measure that assessed only one aspect would be vulnerable to deliberate compensatory efforts by suspected drivers.

It is important to consider the way the tests perform with repeated usage. While the majority of drivers monitored will be experiencing these procedures for the first time, a significant and important minority will be assessed on more than one occasion. Thus tests should continue to detect impairment on the second and subsequent administrations. Two tests showed problems in this respect. Both PAL and LE showed a reduced ability to discriminate between placebo and alcohol on the second test day. Tracking showed a tendency to improve performance overall on the second day. Other tests, including FIT, were robust in this respect.
Although RITA did not show a substantial improvement in ability to discriminate impairment, it has some other advantages. The first is to provide an objective assessment. While FIT depends on the judgement of the administering police officer to assess impairment, a system such as RITA gives a result that does not depend on judgement. Thus a more standardised result may be obtained than with FIT with less requirement for training.

The second advantage to RITA is that the tests included in the battery appear to be less dependent on age than for FIT. While it was expected that some age effect would be found for both methods, the size of the effect with FIT was not anticipated. While the proportion of failures for FIT in volunteers on placebo (the false positive rate) was 27% for those under 40, it was 67% for those over 40, raising concerns about the use of FIT to assess older drivers. The age effects for the RITA results are much less marked, even though older users are often expected to have more difficulty with electronic systems than the young.

The age effect with FIT was not seen in the Glasgow study (Oliver et al., 2006), presumably because their group was predominantly a young one. Their age-range was wide (15–74), but the mean was 27–28, suggesting that the great majority of their subjects were under 40. Thus their placebo failure rate of about 25% is consistent with our findings. Their group presumably reflects the overall patterns of stopping suspected divers, but neither alcohol nor drug driving is confined to a particular age group, and methods should be valid for older as well as younger drivers.

Nystagmus was included in our study as it is part of the SFST used in the USA, but not in FIT. Nystagmus showed good results on its own, giving 74% accuracy. This is comparable to RITA, and takes much less time to carry out. Given that nystagmus is capable of detecting the effect of a number of drugs as well as alcohol (e.g. see Logan et al., 2000; Cochems et al., 2007), this method warrants further study. As used in the SFST, nystagmus requires judgement in a similar way to the other FIT tests, but it should be possible to develop methods to standardise administration and record resulting eye movements in a field setting. The combination of nystagmus with objective tests could be a promising approach.

In conclusion, the automated approach to testing offers advantages in terms of providing an objective measure of assessment and being relatively less affected by age than are the field impairment tests, but does not offer a great benefit over FIT in terms of accuracy of discriminating impairment. Other methods such as nystagmus may warrant further investigation.

Acknowledgements

This research was funded by the UK Government’s Home Office. The study would not have been possible without the software skills of HOSDB’s Chris Treglohan and Tom Smith and the organisational skills of Leah George. Thanks are due to the Supervising Consultant, Dr. Gordon Drummond, to the Research Nurses, Stephanie Dodds and Caroline McAlver, and to the staff of the Clinical Research Facility, Royal Infirmary of Edinburgh. The authors are also grateful to Chief Inspector Mick Trosh and national FIT instructor, Steve Collier for their advice on the operational policing aspects of this study.

References


