In Press: Journal of Forensic Sciences. May, 2005

Statistical Evaluation of Standardized Field Sobriety Tests

Michael P. Hlastala[1], Ph.D.; Nayak L. Polissar[2], Ph.D.; and Steven Oberman[3], J.D.

1. Division of Pulmonary and Critical Care Medicine, Department of Medicine, Department of Physiology and Biophysics, University of Washington, Seattle, WA 98195-6522
2. The Mountain-Whisper-Light Statistical Consulting, Seattle, WA 98112
3. Daniel and Oberman, an Association of Trial Lawyers, 550 W. Main St., Suite 950; Knoxville, TN 37902

Running Head: Field Sobriety Test Accuracy

ABSTRACT: Standardized Field Sobriety Tests (SFSTs) are used as qualitative indicators of impairment by alcohol in individuals suspected of DUI. Stuster and Burns authored a report on this testing and presented the SFSTs as being 91% accurate in predicting Blood Alcohol Concentration (BAC) as lying at or above 0.08%. Their conclusions regarding accuracy are heavily weighted by the large number of subjects with very high BAC levels. This present study re-analyzes the original data with a more complete statistical evaluation. Our evaluation indicates that the accuracy of the SFSTs depends on the BAC level and is much poorer than that indicated by Stuster and Burns. While the SFSTs may be usable for evaluating suspects for BAC, the means of evaluation must be significantly modified to represent the large degree of variability of BAC in relation to SFST test scores. The tests are likely to be mainly useful in identifying subjects with a BAC substantially greater than 0.08%. Given the moderate to high correlation of the tests with BAC, there is potential for improved application of the test after further development, including a more diverse sample of BAC levels, adjustment of the scoring system and a statistically-based method for using the SFST to predict a BAC greater than 0.08 %.

KEYWORDS: forensic science, alcohol, intoxication, horizontal gaze nystagmus, one leg stand, walk and turn.

In August of 1998, The National Highway Traffic Safety Administration published on their web page, a final report entitled "Validation of the Standardized Field Sobriety Test Battery at BACs Below 0.10%" (1) as a follow-up to the original work of Burns and Moskowitz (2) and of that of Tharp et al (3). This report has been used as a standard for Field Sobriety Testing (FST) by law enforcement agencies around the US. In the report, authors Stuster and Burns conclude that the use of SFSTs for "estimates of the 0.08% level were accurate in 91 percent of the cases, or as high as 94 percent "if explanations for some of the false positives are accepted". However the conclusion regarding accuracy is strongly influenced by the large number of subjects with BAC levels much greater than the 0.08% level The accuracy is substantially less for individuals with lower BAC levels, as will be shown below. Three additional papers have recently been published addressing accuracy of sobriety tests at lower alcohol levels. McKnight et al (4) evaluated BAC levels below 0.10 using Horizontal Gaze Nystagmus (HGN) and other modified tests. These authors used correlation analysis and concluded that HGN was the only valid indicator effective in identifying subjects between BAC levels of 0.04% and 0.08%. Another study by Heishman et al (5) focused on ethanol at low levels, cocaine and marijuana using correlation analysis with a variety of variables in addition to the SFSTs so it is difficult to correlate with the Stuster and Burns data. Cole and Nowaczyk (6) studied 21 sober (non-drinking) subjects using trained police officers to evaluate the SFSTs using videotapes of the individuals performing SFSTs. Forty-six percent of the officers' decisions were that the individual had "too much to drink".

SFSTs are usually used as tools by officers in the field to determine if an arrest followed by a breath test is justified. However, often breath test results are not available in court for a variety of reasons. Under these circumstances, the SFST's are frequently used as an indication of impairment and sometimes as an indicator that the subject has a BAC greater than 0.08 g/dl.

The purpose of this report is to outline the statistical strengths and weaknesses of the Stuster and Burns report (1) (SBR) and to suggest some improvements in the use of SFSTs. Our findings suggest that the SFSTs may be helpful in estimating blood alcohol concentration (BAC) or breath alcohol concentration (BrAC), but the results of the SBR must be interpreted more conservatively than suggested by the authors.

Methods

The original study was funded by the National Highway Traffic Safety Administration (NHTSA) and carried out in the San Diego area by seven police officers who administered the SFSTs on those stopped for suspicion of driving under the influence (DUI) of alcohol. The officers were instructed to carry out the SFSTs on the subjects, and then to note an estimated BAC based only on the SFST results: including the walk and turn (WAT), the one leg stand (OLS) and the horizontal gaze nystagmus (HGN) tests. Subjects driving appropriately were not stopped or tested. However, "poor drivers" were included because they attracted the attention of the officers. The data

collection did not include body weight, presence of prior injuries, and other factors that might influence either the SFSTs or the measured BAC (7, 8).

The officers were asked to estimate the BAC values[1] using SFSTs. Some of the subjects were arrested and given a breath test. The criteria used by the officers for estimation of BAC were not described in the report. There appears to be no specific quantitative combination of the FSTs, but rather there appears to be a subjective estimate of BAC. In other words, the decision to determine an estimated BAC was left to the subjective judgment of each officer. Each set of FSTs (for a given subject) was scored by only one officer. So it was not possible to assess inter-officer variability.

The data of Stuster and Burns were obtained via a request to the National Highway and Transportation Safety Administration (NHTSA) using the Freedom of Information Act (FOIA). Figure 1 shows the raw data {estimated BAC (EBAC) vs. measured BAC (MBAC)} for 297 subjects, who had a mean EBAC and MBAC of 0.117% and 0.122%, respectively. The figure shows the line of identity (EBAC = MBAC) and a least-squares regression line for EBAC vs. MBAC. In some cases the EBAC was greater than the MBAC resulting in a greater probability of arrest than if the MBAC had been used (points above the line of identity). In other cases EBAC was lower than MBAC resulting in a lower probability of arrest than if MBAC had been used

[1] The SFSTs are designed to estimate the blood alcohol concentration (BAC) in units of gm/dl. However, the SFSTs are evaluated with the breath alcohol concentration (BrAC) in units of gm/210L. We will use the term, BAC and express the values with units of % to be consistent with the original study.

(points below the line of identity). EBAC is plotted against MBAC for all observations. The MBAC of these points varies over a range of BAC = 0.00% to 0.33%.

Statistical Methods

The accuracy with which officers classified drivers as having a BAC above or below 0.08% is presented graphically by sorting the data on increasing MBAC and then using a moving window of 21 observations, shifting upward one observation at a time. The accuracy is calculated as the percentage of observations in the window that are correctly classified as $< 0.08\%$ or $\geq 0.08\%$ MBAC. The accuracy for the group of 21 observations in the window is plotted vs. the mean of the MBAC measurements in the window.

Four traditional test evaluation statistics were also calculated, namely, 1) sensitivity (percent of true positives who are correctly classified as such by the test), 2) specificity (percent of true negatives who are correctly classified as such by the test), 3) positive predictive value (percent of those with a positive test result who are true positives), and 4) negative predictive value (percent of those with a negative test result who are true negatives) (9). These test evaluation statistics are more commonly used than the accuracy measure defined by SBR. However, the term "accuracy" is used in related literature and in legal proceedings, and, therefore, we use it in this article along with the four more traditional test statistics. It is important to note that one may have very high accuracy yet have much weaker performance on one or more of the four traditional statistics, as happened with SBR.

The relationship of MBAC with the three sub-tests of the SFST, with the total SFST score, and with EBAC were analyzed using simple and multivariate linear regression and with Pearson correlation coefficients as a descriptive measure. (10)

Results

The accuracy of the SFST is not a single percentage, but depends very much on the level of MBAC. Using the 21-observation moving window, the accuracy of classifying individuals as above or below 0.08% MBAC can be pictured in relation to measured breath alcohol concentration (Figure 2). The data show that the officer's accuracy in estimating whether a person's BAC is over or under 0.08% depends on the MBAC. If MBAC is lower than 0.04, the officer is generally 80% or more accurate at predicting a subject's category (above or below 0.08% MBAC) in the sample studied. If the MBAC is greater than 0.09%, then the officer is about 90% or more accurate at predicting the subject's category. However, if the MBAC is around 0.08%, specifically, between 0.06 and 0.08, the SFSTs are only about 30-60% accurate in correctly predicting whether a subject's MBAC is $\geq$ 0.08% or $<$ 0.08%. The minimum accuracy in Figure 2 is 33%.

The data also provide evidence that the officers' estimates were not based only on the SFST. This is shown by an analysis where even very liberal use of **only** the SFST in a predictive model yields a BAC estimate with precision that is substantially inferior to the precision of the officers' estimates, even though the officers were instructed to base their estimates only on the SFST.

Specifically, regression models provide a method to estimate MBAC based only on the three tests in the SFST. A regression model was fitted to predict MBAC from

independent variables including linear and quadratic (squared) terms in the three tests: HGN, HGN^2, OLS, OLS^2, WAT, and WAT^2. The model is liberal in using the three tests, because not all of the variables add significantly or substantially to prediction of the MBAC. Nevertheless, all variables were retained (yielding an over-fitted model), in order to maximize use of the tests within this sample, attempting to mimic or even improve on how an officer might combine test results in practice. Interaction terms between tests were also tried (e.g., HGN*WAT), but they added so little to prediction of MBAC, with a negligible increase in R-squared, that they were not used in the liberal model. (A more appropriate regression model is presented later.)

The amount of variation in MBAC explained by the model based on the three tests alone (and their quadratic terms) is 56%, which increases to 76% when EBAC (the officer estimate of BAC) is added to the model, in addition to the tests. The gain in precision in predicting the quantitative value of MBAC from the model based only on tests to the model based on the tests plus the officer estimates is statistically very significant (20% increase in R-squared, $p < 0.001$). The mean absolute difference between the officer estimate, EBAC, and the measured value, MBAC, is 0.024% (in BAC units), versus a larger value of 0.031% indicating less precision, for the mean absolute difference between the model-based estimate and the MBAC.

The striking increase in precision when the officer estimates are added to a liberally-fitted model using only the tests suggests that the officers did not base their estimate solely on the test scores but most likely used other clues. This suggests that

it may be impractical to evaluate the three tests in isolation from other non-test clues used by the officers, such as slurred speech, odor of alcohol, appearance, admitted drinking or driving behavior. Another explanation may be the presence of other drugs in addition to alcohol. Or, as suggested by critics of the study, Price and Cole (9), it may be that the officers used portable breath testers (PBT) prior to recording their BrAC estimate and were then influenced by the known PBT values. The Stuster and Burns report (1, page 10) notes that "all police officers participating in the study were equipped with NHTSA-approved, portable breath testing devices to assess the BACs of all drivers who were administered the SFST...".

The utility of individual tests (HGN, OLS and WAT) and the combination of tests to predict MBAC can be evaluated by plotting MBAC against the total score from the individual tests. Figure 3 shows a plot of the measured breath alcohol concentration versus the total score from the three tests, with a reference line at MBAC = 0.08%. For Figure 2 only, a small amount of "jitter" (random noise) has been added to the score of each subject to avoid overlapping points. The jitter is less than ±0.25 points horizontally. The considerable variation in MBAC above each point score is apparent, and in addition, for total scores 4-18, there are MBAC values lying on both sides of the 0.08% cut-point. In order to be 95% confident that the subject has a MBAC greater than 0.08%, the total score (HGN + OLS + WAT) must exceed approximately 17 (based on the 95% lower confidence limit for predicted MBAC for an individual from the regression of MBAC on total score).

Figure 4 shows the percentage of measured breath alcohol concentration values that are above 0.08% in relation to each of the three individual test scores. For each score (horizontal axis), the percent of subjects with that score or higher who have an MBAC larger than 0.08% is plotted (Y-axis). In order to observe 95% of persons with MBAC > 0.08% in this sample, the score for WAT (circles in the plot) must be 5 or larger. None of the scores for HGN (crosses) reach the 95% point and the scores for OLS (triangles) reach over the 95% point only at 10 points and higher, where there are only two subjects. Note that the "failure" scores for these three tests, as specified by Stuster and Burns, are 4 for HGN, 2 for OLS, and 2 for WAT (12). Failure of an FST according to NHTSA standards simply estimate the 50% likelihood that a subject is > 0.08%. The data show that in order to be considerably more confident that the subject is above 0.08%, the scores should be much higher than the "failure" scores.

The correlation coefficients for individual tests vs. both MBAC and EBAC are shown in Table 1. The FST with the strongest correlation with MBAC is HGN followed by WAT and OLS. The strongest correlation is with the total test (determined by summing the scores for the three FSTs. However, total score and HGN have very similar correlations with MBAC and EBAC.

Discussion

Figure 5, redrawn from Figure 4 of SBR. illustrates the logic used by Stuster and Burns to describe the accuracy of SFST. A correct decision was registered if both MBAC and EBAC are ≥ 0.08% (upper, right quadrant; N=210) or both are ≤0.08% (lower, left quadrant; N=59). An incorrect decision occurred with a false positive (upper, left quadrant; N=24), when (EBAC ≥ 0.08% and MBAC < 0.08%) or a false negative (lower, right quadrant; N=4), when EBAC < 0.08% and MBAC ≥ 0.08%. Because such a large fraction of the points were between 0.08% and 0.33% (N = 214 of the 297 total points) and most of these had a MBAC > 0.12, Stuster and Burn's conclusion that the tests have 91% accuracy was strongly affected by the fact that a majority of points are in this high MBAC range, where correct classification as above 0.08 is more reliable. Of the correct results, 210 data points out of a study total of 297 were in the 0.08% to 0.33% range and 59 were in the 0.000% to 0.079% range. (The accuracy estimated by Stuster and Burns as 91% was calculated from the values in Figure 2 as (210 + 59)/297 = 0.91). The number of false positives (N=24) was much greater than the number of the false negatives (N=4). In the range of data near the 0.08% level, the estimated BAC by these experienced officers overestimates the measured BAC, introducing a bias against the subjects (see Figure 1). Using EBAC to determine whether the subject MBAC is greater than 0.08% is 100% accurate for all subjects with MBAC > 0.12%. In other words, if the subject is highly intoxicated, the SFST provide an accurate indication. It is not surprising that if the subject is clearly intoxicated, the officers can make this determination. If the MBAC is < 0.08%, there is a 24 / (24 +59) = 29% chance of a false arrest (determined from Figure 2).

To illustrate the problem with the SBR statistical strategy, let's apply the same logic to determine the level of accuracy at hypothetical cut-point ("legal limit") levels lower than 0.08%. For example, if Stuster and Burns were to use the same data set to examine the accuracy at lower threshold BAC (0.07% down to 0.01%) levels, they would determine an increasing accuracy level at lower threshold levels. The relative increase in apparent accuracy with decreasing BAC threshold is shown in Table 2, which indicates a hypothetical cut-point for designating a driver as "over the limit". For example, if the legal limit were 0.04%, the SBR method would conclude that SFST are 93.9% accurate. At a legal limit of 0.01%, the SBR conclusion would be that the SFST are 99.3% accurate. The method used by Stuster and Burns has determined a high degree of accuracy simply because most of the data points are at MBACs much greater than the cut-point of 0.08% used in their study. What underlies this problem is the weakness of "accuracy" as the sole performance statistic for this test, as well as the specific nature of this sample, weighted heavily toward individuals with high levels of MBAC.

An alternative way to explore the accuracy of SFST is to assess the accuracy over a range of points that is symmetric about the 0.08% cut-point (limit). In addition to accuracy, four traditional statistics of test performance also help in this exploration: sensitivity, specificity, positive predictive value and negative predictive value. Table 3 shows the accuracy of SFST when the range of interest extends above and below 0.08% by the same amount, along with the four traditional performance statistics. For

data with MBAC ranging between 0.07% and 0.09%, The SFST are 72.2% accurate. As the range broadens, the calculated apparent accuracy increases. At the broadest range of 0.04% – 0.12%, the calculated apparent accuracy is now 82.2%. Taken to the extreme, using all of the data points (MBAC = 0.00% to 0.033%), the apparent accuracy is 91% as calculated by Stuster and Burns. The accuracy of SFST in the vicinity of 0.08% is poorer than estimated in the SBR for the whole data set.

Parallel with the reduced level of accuracy in the range 0.07-0.09% MBAC, the four traditional test performance statistics in Table 3 also show varying performance in this range. Specificity is low (36%), indicating that a large fraction of subjects (64%) would be falsely declared over the limit. The sensitivity is excellent in this range, 96%, due to the tendency of EBAC to overestimate alcohol level compared to MBAC. Positive predictive value (PPV) is fair, 70%, indicating that 30% of the subjects declared over-limit would not be so. Negative predictive value (NPV) is good, 83%, indicating that most of those declared under-limit would really be so, but this, again, due to the over-estimation by EBAC. As the range of MBAC in Table 3 steadily widens to finally include all cases, specificity increases to a maximum of only 71%, while sensitivity, PPV and NPV all reach at least 90%, due to predominance in this sample of high levels of measured alcohol.

A closer examination of the data between 0.04% and 0.12% is shown on Figure 6 (by expanding a section of Figure 1). Another way of determining the officer's accuracy in estimating the BAC is to compare the fraction of observations (EBAC)

overestimating and underestimating the MBAC. If we consider three ranges of MBAC, 0.00% ≤ MBAC < 0.04%, 0.04% ≤ MBAC < 0.08%, and 0.08% ≤ MBAC < 0.12%, the officers' EBAC overestimated the MBAC 76%, 67% and 48% of the time, respectively, estimated it exactly 10%, 7%, and 24% of the time, and underestimated it 14%, 26% and 28% of the time. Overestimation occurs more frequently than underestimation for MBAC < 0.12%. Further evidence of officer overestimation at lower MBAC levels can be taken from the regression line and the line of identity in figures 1 and 6. At MBAC values **at or** below approximately 0.10%, the EBAC tends to overestimate the MBAC. Out of 123 points in this range, 80 overestimate and 26 underestimate the MBAC. For MBAC above 0.10%, the EBAC values tend to underestimate the MBAC. Out of 174 data points with MBAC > 0.10%, 50 are overestimates and 108 are underestimates of MBAC. Thus, the experienced officers used in this study tended to overestimate the BAC at low levels (< 0.10) and underestimate the BAC at higher levels (> 0.10).

The optimal predictive capability of the SFST depends on the scaling for the particular test and the predictive capacity of the test. The maximum scores **permitted** for HGN, OLS and WAT are 6, 4, and 8, respectively. However, some officers assigned scores that were greater than the maximum score allowable for a given FST. The highest scores assigned in this study were 6, 12, and 9 for the HGN, OLS and WAT, respectively.

By adjusting the weight given to each test and taking account of the precision of the test in predicting MBAC, we find the following linear regression model (equation 1)

maximizes the precision of the SFST for estimating MBAC, using only linear versions of the three test variables. The quadratic terms (squared values of the three test variables), while statistically significant as a group ($p = 0.004$) increase R-squared by only 2%, from 54% to 56%, and have been omitted for parsimony. The model is based on the 261 cases without any missing values for the three tests. Note in the equation below that the increase in BAC per point increase in the score is largest for HGN, with a 0.017 increase in BAC, on the average, for each point increase in the HGN score.

MBAC = -0.007 + 0.017 x (HGN Score) + 0.0012 x (OLS Score) + 0.011 x (WAT Score) (Eq. 1)

The equation does quite well in predicting the mean MBAC, but there is still a large spread of individuals around the predicted value. The standard deviation of individual MBAC values around the predicted regression value is 0.044%. A 95% confidence interval for the true MBAC of an individual, predicted from this regression model, would have a minimum width of ± 0.09%, certainly a wide range.

Using the predictors (HGN, OLS, WAT), the additive model from equation 1 accounted for 54% of the variability in MBAC (corresponding to a correlation of 0.73). Including EBAC as an additional predictor in the model resulted in a substantial and significant increase ($p < 0.001$) in the variance of MBAC explained, increasing it to 75%. As noted earlier, this marked increase in predictability of MBAC by adding in the

officer's EBAC indicates that the officers' estimates were probably influenced by factors other than the three FSTs

We believe that the accuracy of the SFST can be improved if a weighted sum of scores from the three standard tests is combined as described in Equation 1. However, this relationship should be tested in a variety of populations, and, in a larger sample, it is possible that non-linear and other functions of the test scores may help in prediction. The evaluation should include an assessment of accuracy and bias in estimating the numerical BAC and, as well, the accuracy in classifying individuals above or below specified limits (such as 0.08%) for various low, medium and high levels of measured BAC. In follow-up trials of the FST, the instructions given to officers for converting test scores into estimates of BAC should be stated more explicitly (such as using equation 1 above, or another algorithm). Further, some attempt should be made to identify and incorporate (or control) other factors, aside from the SFST scores, that influence BAC estimates. It may be difficult or impossible to "turn off" other cues that officers use in estimating BAC or in making a decision about an arrest.

The magnitude of the correlations between the tests and MBAC suggests that this type of testing could be developed further, either through re-formulation of the tests, or through different scoring systems, **or** by other means. In the current framework, the test scores have to be quite high to provide confidence that the subject is above 0.08%, but further development could potentially improve confidence in the **three** test results, both singly and in combination. And, anticipating the possibility that

some jurisdictions may now or in the future have lower (or higher) legal limits than 0.08%, testing could include more representation from lower levels of BrAC.

The SFST total score and sub-test scores are undoubtedly correlated with breath alcohol level (Table 1). However, predicting a numeric blood alcohol concentration from the SFST scores, as the SFST methodology is defined in the Stuster and Burns report, has limited accuracy and precision. The evidence for this is a) considerable over- and under-estimation of MBAC (see Results section); b) a large range of observed MBAC values corresponding to any given total SFST score (Figure 3); and, c) a large spread of observed MBAC values around predicted MBAC values from a liberal regression model that attempts to optimize the use of the SFST, yet has a minimum prediction uncertainty of ±0.09%.

If our interest is not in quantitative prediction, but in classifying individuals, such as below vs. equal to or above a limit of 0.08%, the utility of the SFST depends very much on how intoxicated an individual is. Accuracy (and specificity) are low when individuals are close to 0.08% MBAC (Figure 2 and Table 3), but if the individuals are quite intoxicated, such as above 0.12%, then accuracy is high (Figure 2).

The use of a single test performance statistic, accuracy, and the calculation of this one statistic for the entire study sample is an over-simplification of the more complex relationship between the SFST score and the MBAC level.

SFSTs could become more useful if much more data are accumulated and analyzed using statistical methods such as those presented in this paper, including some of the traditional test evaluation statistics. It is likely that the usefulness of SFSTs will be greatest for drivers who have high test scores. The moderate to strong correlations between the tests and MBAC suggest a potential for further test development. Enhanced understanding would come from tests applied to a more diverse population sample as well as from the development of a statistical approach to predicting the probability of a subject having a BAC greater than 0.08 % from a particular set of SFST scores.

References:

1. Stuster J, Burns M. Validation of the standardized field sobriety test battery at BACs below 0.10 percent. August, 1998. National Highway Traffic Safety Administration.

2. Burns M, Moskowitz H. Psychophysical tests for DWI arrest. Technical Report DOT-HS-5-01242. National Highway Traffic Safety Administration. Washington, DC.

3. Tharp V, Burns M and Moskowitz H. Development and field test of psychophysical tests for DWI arrest. US Department of Transportation, National Highway Traffic Safety Administration Final Report DOT-HS-805-864, Washington, DC.

4. McKnight, AJ, Langston, EA, McKnight AS, Lange, JE. Sobriety tests for low blood alcohol concentrations. Acc Anal & Prevent 2002;34: 305-311.

5. Heishman, SJ, Singleton, EG, Crouch, DJ. Laboratory validation study of drug evaluation and classification program: ethanol, cocaine, and marijuana. J Anal Toxicol 1996; 20: 468-481.

6. Cole S and Nowaczyk, RH. Field sobriety tests: Are they designed for failure? Perceptual and Motor Skills 1994; 79: 99-104.

7. Hlastala M. The alcohol breath test - A brief review. J Appl Physiol 1998; 84: 401-408.

8. Hlastala M. Invited editorial on "The alcohol breath test". J Appl Physiol 2002; 93: 405-406.

9. Price P, Cole S. NHTSA field sobriety tests validation v. invalidation, 25 The Champion. 2001; 25: 37-42.

10. Fisher LD, van Belle G. Biostatistics. Wiley, 1993.

11. Weisberg S. Applied linear regression, 2nd edition. Wiley, 1985.

12. NHTSA DWI Detection and Standardized Field Sobriety Testing Student Manual, DOT-HS-178-R1/02.

Additional information and reprint requests:

Michael P. Hlastala, Ph.D.
Division of Pulmonary and Critical Care Medicine, Department of Medicine
Department of Physiology and Biophysics
Box 356522
University of Washington
Seattle, WA 98195-6522
Email: hlastala@u.washington.edu

TABLES

Table 1. Pearson correlation of three Field Sobriety Tests with measured breath alcohol (MBAC) and officer-estimated breath alcohol (EBAC).

TEST	MBAC	EBAC
TOTAL score (3 tests)	0.69	0.74
HGN Horizontal Gaze Nystagmus	0.65	0.71
WAT Walk and turn	0.61	0.64
OLS One leg stand	0.45	0.51

Table 2. Accuracy of "over-limit" designation based on estimated breath alcohol concentration for defined cut-points (hypothetical legal "limit") of measured breath alcohol concentration (MBAC)

Legal "limit" (%)	N: All	N: MBAC < cut-point	N: MBAC ≥ cut-point	Accuracy*
0.10	297	107	190	90.6%
0.09	297	97	200	89.2%
0.08	297	83	214	90.6%
0.07	297	69	228	89.6%
0.06	297	58	239	90.6%
0.05	297	43	254	92.3%
0.04	297	29	268	93.9%
0.03	297	19	278	93.9%
0.02	297	9	288	97.6%
0.01	297	4	293	99.3%

Accuracy = 100%(# correctly classified as ≥ limit or < limit)/total

Table 3. Accuracy and other statistics related to "over-limit" designation based on estimated breath alcohol concentration for defined ranges of MBAC.

Range of MBAC	Total in Range	Accuracy	Sensitivity	Specificity	PPV	NPV
0.07 – 0.09	36	72%	96%	36%	70%	83%
0.06 – 0.10	65	75%	95%	44%	73%	85%
0.05 – 0.11	97	79%	97%	55%	75%	92%
0.04 – 0.12	135	82%	95%	63%	79%	90%
All cases	297	91%	98%	71%	90%	94%

Accuracy = (# correctly classified as ≥ 0.08 or < 0.08)/total
PPV = positive predictive value
NPV = negative predictive value

Figure Legends:

Figure 1. Estimated BAC vs. Measured BAC for all subjects in the Stuster and Burns study. The line of identity (Estimated BAC = Measured BAC; thin line) and linear regression line (heavy solid line) are shown.

Figure 2. Accuracy of classification of individuals as ≥ 0.08% or < 0.08% MBAC using the officer estimate. Accuracy is plotted vs. measured breath alcohol concentration (horizontal axis).

Figure 3. Measured breath alcohol concentration versus total of three test scores.

Figure 4. Percent of subjects with MBAC greater than 0.08% vs. the individual test score, with the percentage calculated for all individuals at or above the designated score.

Figure 5. Decision matrix at 0.08% BAC (modified from figure 4 in Stuster and Burns).

Figure 6. Data from Figure 1 expanded to show points between 0.04% and 0.12%. The line of identity (EBAC = MBAC), dashed line and linear regression line (heavy solid line) are shown.

Figure 1

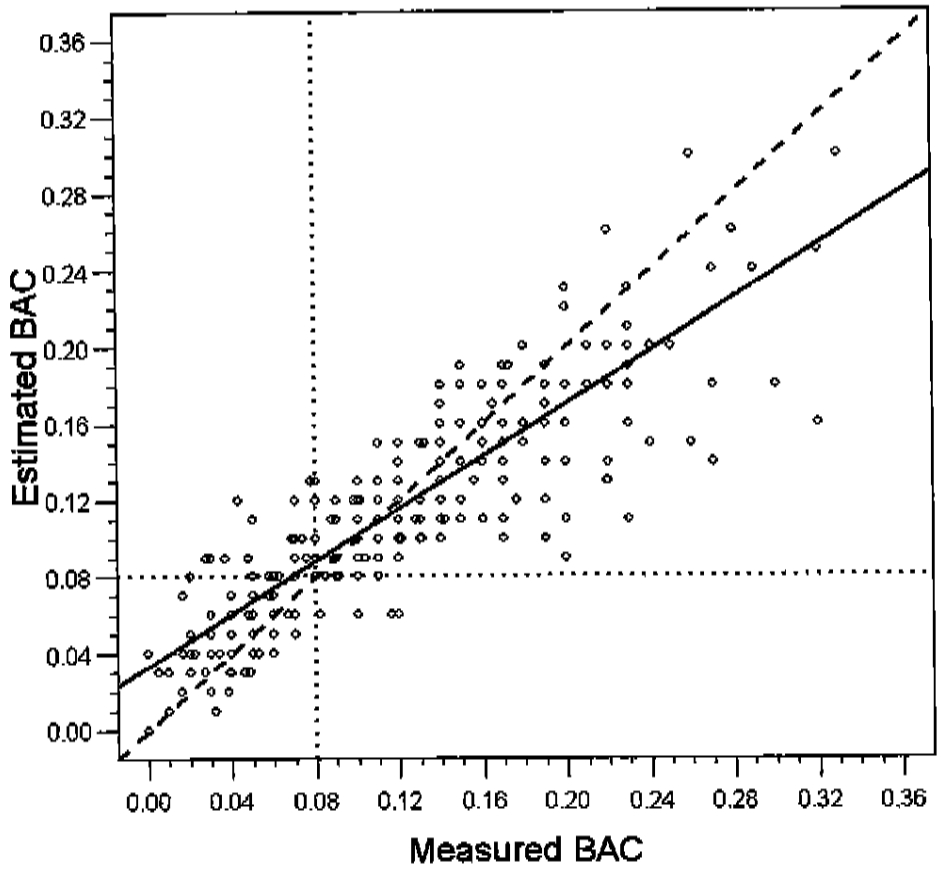

Estimated BAC
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
0.36
Measured BAC

Figure 2

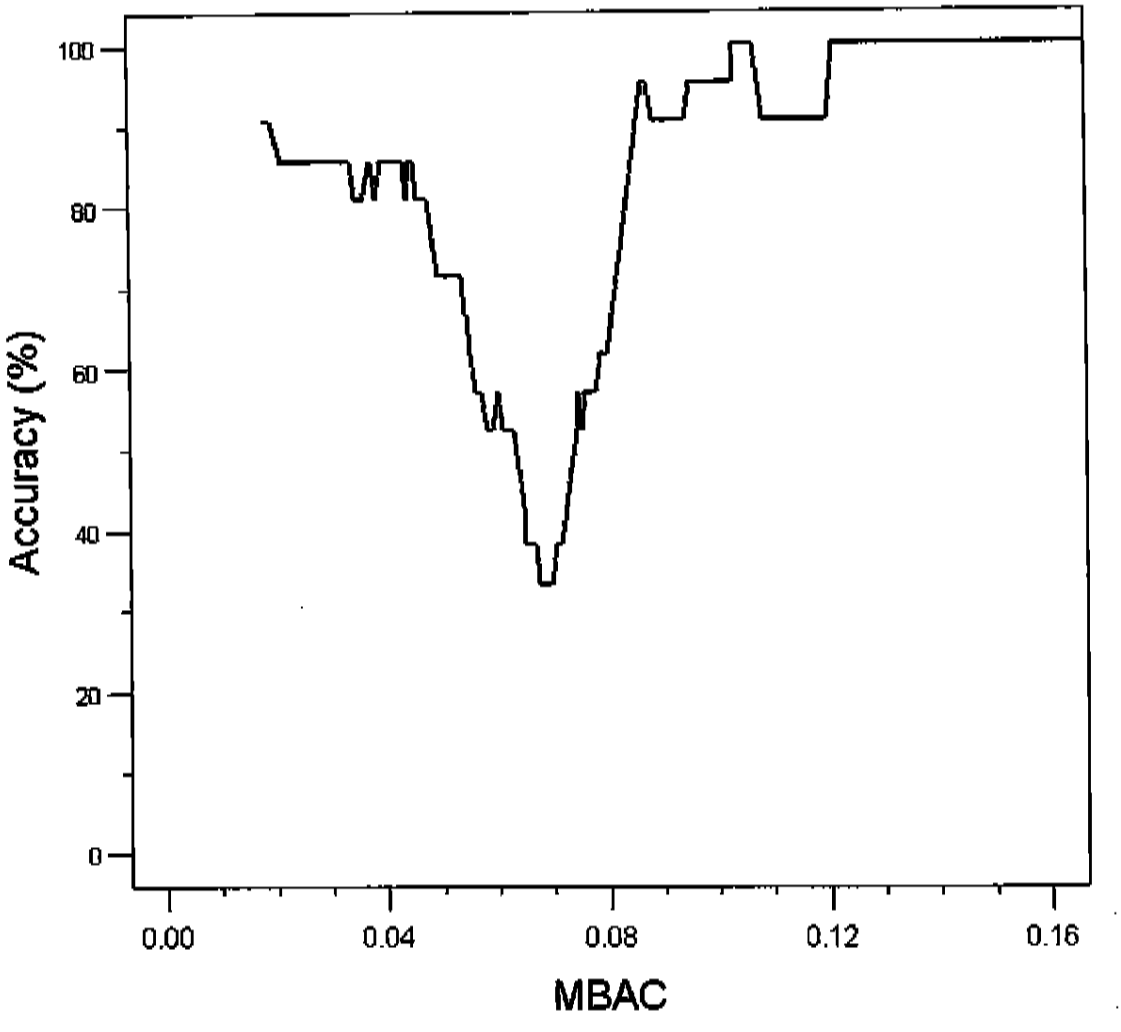


Figure 3.

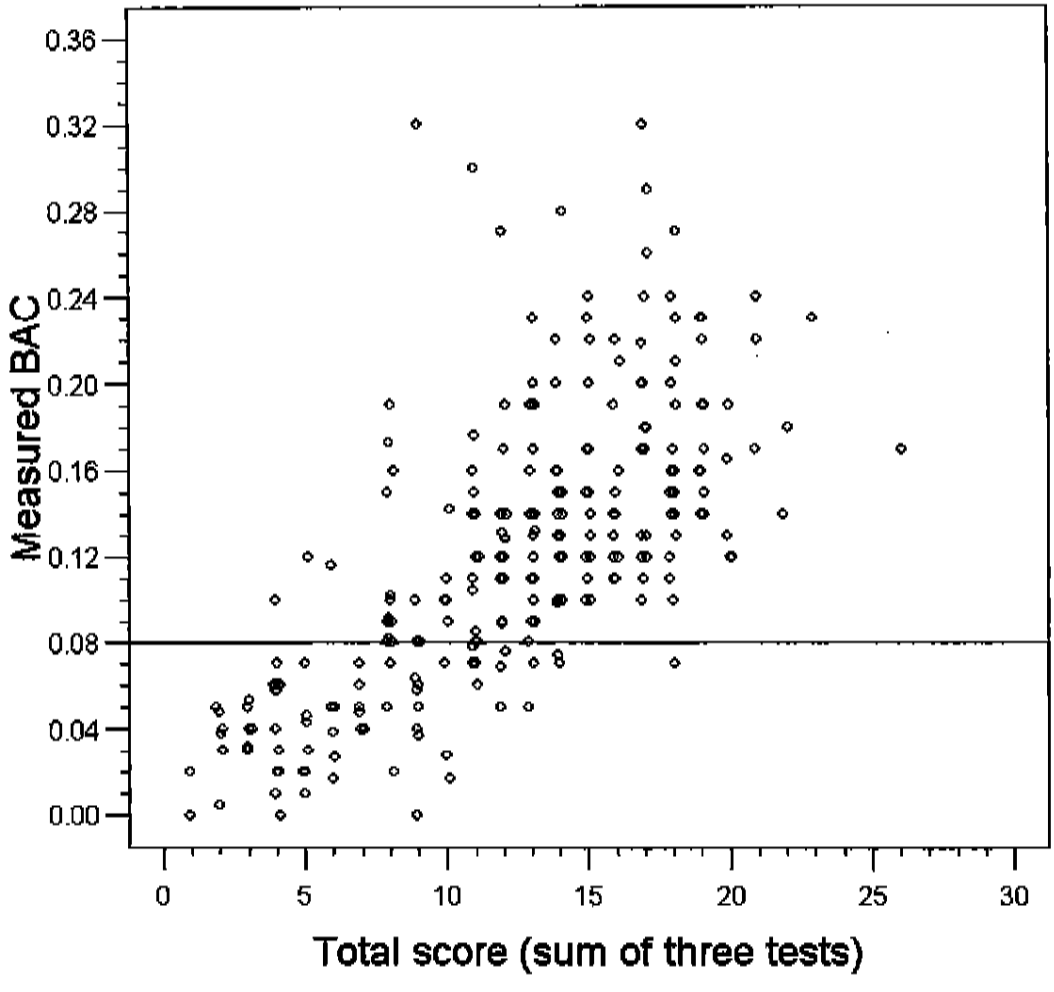

Measured BAC
0.36
0.32
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00
0
5
10
15
20
25
30
Total score (sum of three tests)

Figure 4.

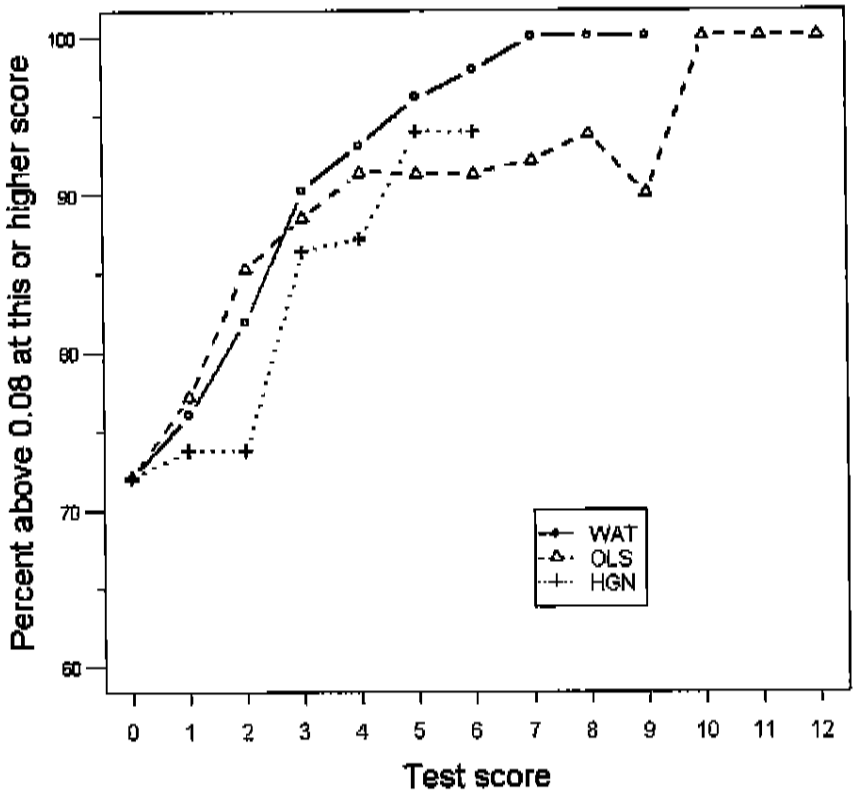

Percent above 0.08 at this or higher score
100
90
80
70
60
0
1
2
3
4
5
6
7
8
9
10
11
12
Test score
WAT
OLS
HGN

Figure 5.

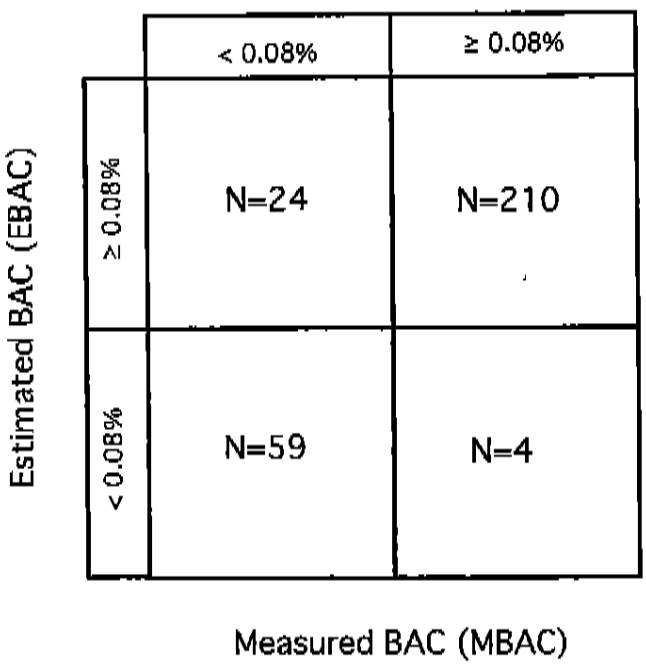
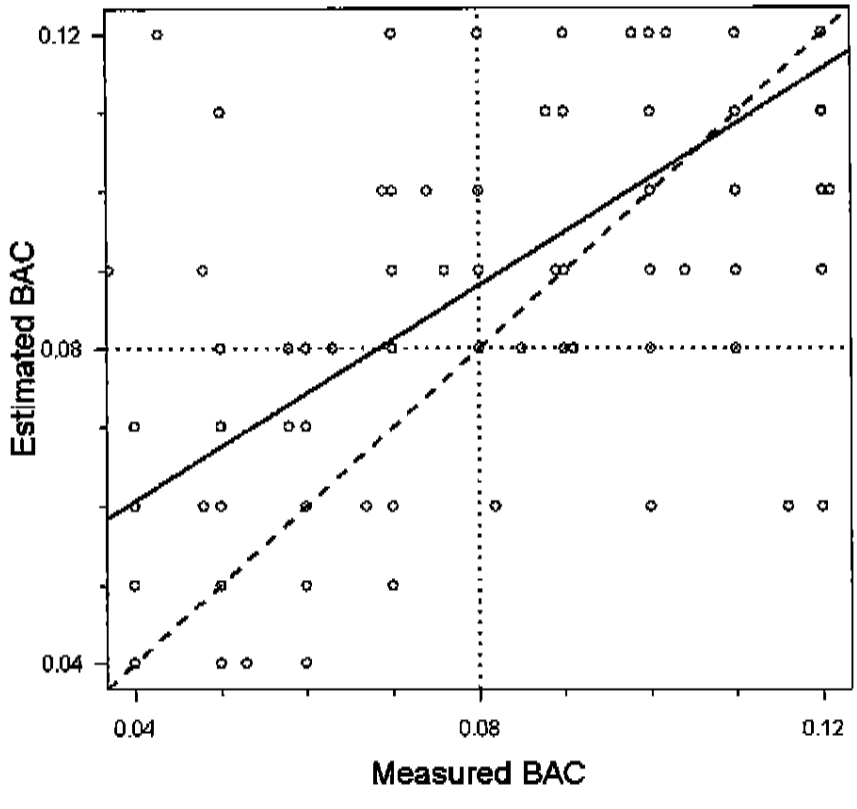

< 0.08%
≥ 0.08%
≥ 0.08%
< 0.08%
N=24
N=210
N=59
N=4
Estimated BAC (EBAC)
Measured BAC (MBAC)

Figure 6.



J Forensic Sci, May 2005, Vol. 50, No. 3
Paper ID JFS2003386
Available online at: www.astm.org

Michael P. Hlastala,[1] Ph.D.; Nayak L. Polissar,[2] Ph.D.; and Steven Oberman,[3] J.D.

Statistical Evaluation of Standardized Field Sobriety Tests

ABSTRACT: Standardized Field Sobriety Tests (SFSTs) are used as qualitative indicators of impairment by alcohol in individuals suspected of DUI. Stuster and Burns authored a report on this testing and presented the SFSTs as being 91% accurate in predicting Blood Alcohol Concentration (BAC) as lying at or above 0.08%. Their conclusions regarding accuracy are heavily weighted by the large number of subjects with very high BAC levels. This present study re-analyzes the original data with a more complete statistical evaluation. Our evaluation indicates that the accuracy of the SFST's depends on the BAC level and is much poorer than that indicated by Stuster and Burns. While the SFSTs may be usable for evaluating suspects for BAC, the means of evaluation must be significantly modified to represent the large degree of variability of BAC in relation to SFST test scores. The tests are likely to be mainly useful in identifying subjects with a BAC substantially greater than 0.08%. Given the moderate to high correlation of the tests with BAC, there is potential for improved application of the test after further development, including a more diverse sample of BAC levels, adjustment of the scoring system and a statistically-based method for using the SFST to predict a BAC greater than 0.08%.

KEYWORDS: forensic science, alcohol, intoxication, horizontal gaze nystagmus, one leg stand, walk and turn

In August of 1998, The National Highway Traffic Safety Administration published on their web page, a final report entitled "Validation of the Standardized Field Sobriety Test Battery at BACs Below 0.10%" (1) as a follow-up to the original work of Burns and Moskowitz (2) and of that of Tharp et al. (3). This report has been used as a standard for Field Sobriety Testing (FST) by law enforcement agencies around the U.S. In the report, authors Stuster and Burns conclude that the use of SFSTs for "estimates of the 0.08% level were accurate in 91 percent of the cases, or as high as 94 percent "if explanations for some of the false positives are accepted." However, the conclusion regarding accuracy is strongly influenced by the large number of subjects with BAC levels much greater than the 0.08% level The accuracy is substantially less for individuals with lower BAC levels, as will be shown below. Three additional papers have recently been published addressing accuracy of sobriety tests at lower alcohol levels. McKnight et al. (4) evaluated BAC levels below 0.10 using Horizontal Gaze Nystagmus (HGN) and other modified tests. These authors used correlation analysis and concluded that HGN was the only valid indicator effective in identifying subjects between BAC levels of 0.04% and 0.08%. Another study by Heishman et al. (5) focused on ethanol at low levels, cocaine and marijuana using correlation analysis with a variety of variables in addition to the SFSTs so it is difficult to correlate with the Stuster and Burns data. Cole and Nowaczyk (6) studied 21 sober (non-drinking) subjects using trained police officers to evaluate the SFSTs using videotapes of the individuals performing SFSTs. Forty-six percent of the officers' decisions were that the individual had "too much to drink."

SFSTs are usually used as tools by officers in the field to determine if an arrest followed by a breath test is justified. However, often breath test results are not available in court for a variety of reasons. Under these circumstances, the SFST's are frequently used as an indication of impairment and sometimes as an indicator that the subject has a BAC greater than 0.08 g/dL.

The purpose of this report is to outline the statistical strengths and weaknesses of the Stuster and Burns report (1) (SBR) and to suggest some improvements in the use of SFSTs. Our findings suggest that the SFSTs may be helpful in estimating blood alcohol concentration (BAC) or breath alcohol concentration (BrAC), but the results of the SBR must be interpreted more conservatively than suggested by the authors.

Methods

The original study was funded by the National Highway Traffic Safety Administration (NHTSA) and carried out in the San Diego area by seven police officers who administered the SFSTs on those stopped for suspicion of driving under the influence (DUI) of alcohol. The officers were instructed to carry out the SFSTs on the subjects, and then to note an estimated BAC based *only* on the SFST results: including the walk and turn (WAT), the one leg stand (OLS) and the horizontal gaze nystagmus (HGN) tests. Subjects driving appropriately were not stopped or tested. However, "poor drivers" were included because they attracted the attention of the officers. The data collection did not include body weight, presence of prior injuries, and other factors that might influence either the SFSTs or the measured BAC (7,8).

The officers were asked to estimate the BAC values[4] using SFSTs. Some of the subjects were arrested and given a breath test. The criteria used by the officers for estimation of BAC were

[1] Division of Pulmonary and Critical Care Medicine, Department of Medicine, Department of Physiology and Biophysics, University of Washington, Seattle, WA 98195-6522.
[2] The Mountain-Whisper-Light Statistical Consulting, Seattle, WA 98112.
[3] Daniel and Oberman, an Association of Trial Lawyers, 550 W. Main St., Suite 950; Knoxville, TN 37902.
Received 16 Nov. 2003; and in revised form 31 July 2004; accepted 12 Nov. 2004; published 6 April 2005.

[4] The SFSTs are designed to estimate the blood alcohol concentration (BAC) in units of g/dL. However, the SFSTs are evaluated with the breath alcohol concentration (BrAC) in units of g/210L. We will use the term, BAC and express the values with units of % to be consistent with the original study.

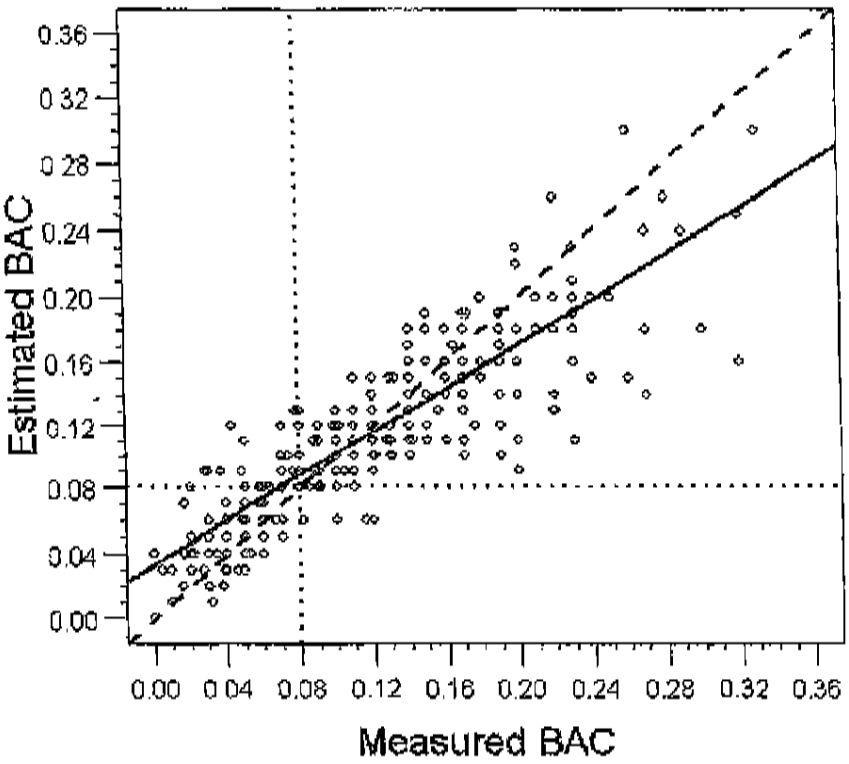


FIG. 1—*Estimated BAC vs. Measured BAC for all subjects in the Stuster and Burns study. The line of identity (Estimated BAC = Measured BAC; thin line) and linear regression line (heavy solid line) are shown.*

not described in the report. There appears to be no specific quantitative combination of the FSTs, but rather there appears to be a subjective estimate of BAC. In other words, the decision to determine an estimated BAC was left to the subjective judgment of each officer. Each set of FSTs (for a given subject) was scored by only one officer. So it was not possible to assess inter-officer variability.

The data of Stuster and Burns were obtained via a request to the National Highway and Transportation Safety Administration (NHTSA) using the Freedom of Information Act (FOIA). Fig. 1 shows the raw data {estimated BAC (EBAC) vs. measured BAC (MBAC)} for 297 subjects, who had a mean EBAC and MBAC of 0.117% and 0.122%, respectively. The figure shows the line of identity (EBAC = MBAC) and a least-squares regression line for EBAC vs. MBAC. In some cases the EBAC was greater than the MBAC resulting in a greater probability of arrest than if the MBAC had been used (points above the line of identity). In other cases EBAC was lower than MBAC resulting in a lower probability of arrest than if MBAC had been used (points below the line of identity). EBAC is plotted against MBAC for all observations. The MBAC of these points varies over a range of BAC = 0.00% to 0.33%.

Statistical Methods

The accuracy with which officers classified drivers as having a BAC above or below 0.08% is presented graphically by sorting the data on increasing MBAC and then using a moving window of 21 observations, shifting upward one observation at a time. The accuracy is calculated as the percentage of observations in the window that are correctly classified as <0.08% or ≥0.08% MBAC. The accuracy for the group of 21 observations in the window is plotted vs. the mean of the MBAC measurements in the window.

Four traditional test evaluation statistics were also calculated, namely: 1) sensitivity (percent of true positives who are correctly classified as such by the test), 2) specificity (percent of true negatives who are correctly classified as such by the test), 3) positive predictive value (percent of those with a positive test result who are true positives), and 4) negative predictive value (percent of those with a negative test result who are true negatives) (9). These test evaluation statistics are more commonly used than the accuracy measure defined by SBR. However, the term "accuracy" is used in related literature and in legal proceedings, and, therefore, we use it in this article along with the four more traditional test statistics. It is important to note that one may have very high accuracy yet have much weaker performance on one or more of the four traditional statistics, as happened with SBR.

The relationship of MBAC with the three sub-tests of the SFST, with the total SFST score, and with EBAC were analyzed using simple and multivariate linear regression and with Pearson correlation coefficients as a descriptive measure (10).

Results

The accuracy of the SFST is not a single percentage, but depends very much on the level of MBAC. Using the 21-observation moving window, the accuracy of classifying individuals as above or below 0.08% MBAC can be pictured in relation to measured breath alcohol concentration (Fig. 2). The data show that the officer's accuracy in estimating whether a person's BAC is over or under 0.08% depends on the MBAC. If MBAC is lower than 0.04, the officer is generally 80% or more accurate at predicting a subject's category (above or below 0.08% MBAC) in the sample studied. If the MBAC is greater than 0.09%, then the officer is about 90% or more accurate at predicting the subject's category. However, if the MBAC is around 0.08%, specifically, between 0.06 and 0.08, the SFSTs are only about 30–60% accurate in correctly predicting whether a subject's MBAC is ≥0.08% or <0.08%. The minimum accuracy in Fig. 2 is 33%.

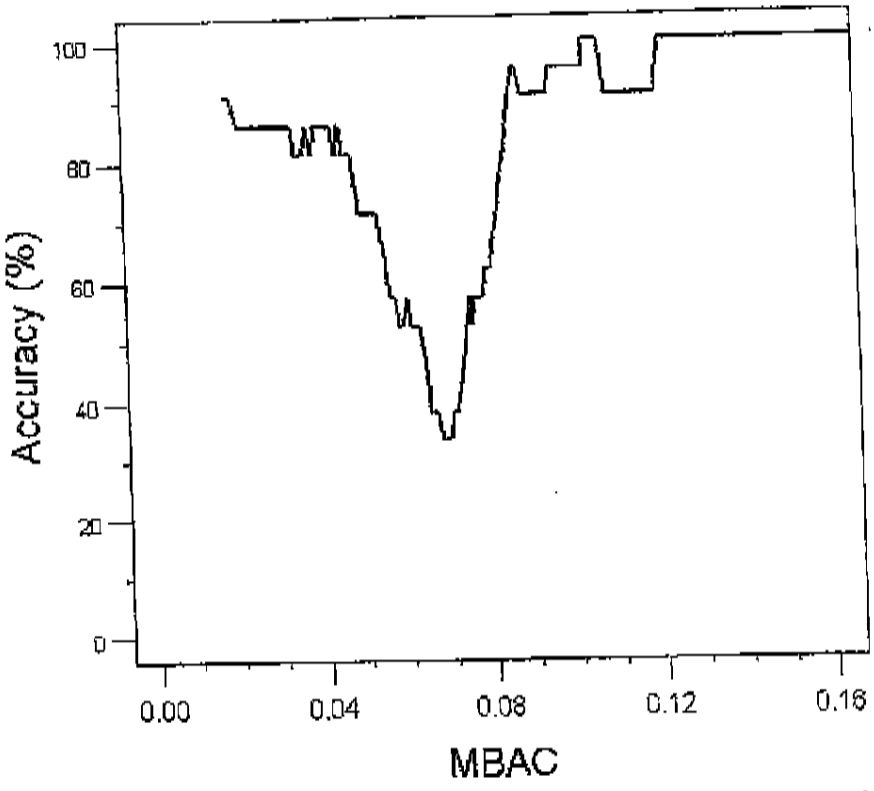


FIG. 2—*Accuracy of classification of individuals as ≥0.08% or <0.08% MBAC using the officer estimate. Accuracy is plotted vs. measured breath alcohol concentration (horizontal axis).*

The data also provide evidence that the officers' estimates were not based only on the SFST. This is shown by an analysis where even very liberal use of only the SFST in a predictive model yields a BAC estimate with precision that is substantially inferior to the precision of the officers' estimates, even though the officers were instructed to base their estimates *only* on the SFST.

Specifically, regression models provide a method to estimate MBAC based only on the three tests in the SFST. A regression model was fitted to predict MBAC from independent variables including linear and quadratic (squared) terms in the three tests: HGN, HGN^2, OLS, OLS^2, WAT, and WAT^2. The model is liberal in using the three tests, because not all of the variables add significantly or substantially to prediction of the MBAC. Nevertheless, all variables were retained (yielding an over-fitted model), in order to maximize use of the tests within this sample, attempting to mimic or even improve on how an officer might combine test results in practice. Interaction terms between tests were also tried (e.g., HGN * WAT), but they added so little to prediction of MBAC, with a negligible increase in R-squared, that they were not used in the liberal model. (A more appropriate regression model is presented later.)

The amount of variation in MBAC explained by the model based on the three tests alone (and their quadratic terms) is 56%, which increases to 76% when EBAC (the officer estimate of BAC) is added to the model, in addition to the tests. The gain in precision in predicting the quantitative value of MBAC from the model based *only* on tests to the model based on the tests plus the officer estimates is statistically very significant (20% increase in R-squared, $p < 0.001$). The mean absolute difference between the officer estimate, EBAC, and the measured value, MBAC, is 0.024% (in BAC units), versus a larger value of 0.031% indicating less precision, for the mean absolute difference between the model-based estimate and the MBAC.

The striking increase in precision when the officer estimates are added to a liberally-fitted model using only the tests suggests that the officers did not base their estimate solely on the test scores but most likely used other clues. This suggests that it may be impractical to evaluate the three tests in isolation from other non-test clues used by the officers, such as slurred speech, odor of alcohol, appearance, admitted drinking or driving behavior. Another explanation may be the presence of other drugs in addition to alcohol. Or, as suggested by critics of the study, Price and Cole (9), it may be that the officers used portable breath testers (PBT) prior to recording their BrAC estimate and were then influenced by the known PBT values. The Stuster and Burns report (1, page 10) notes that "all police officers participating in the study were equipped with NHTSA-approved, portable breath testing devices to assess the BACs of all drivers who were administered the SFST..."

The utility of individual tests (HGN, OLS and WAT) and the combination of tests to predict MBAC can be evaluated by plotting MBAC against the total score from the individual tests. Figure 3 shows a plot of the measured breath alcohol concentration versus the total score from the three tests, with a reference line at MBAC = 0.08%. For Fig. 3 only, a small amount of "jitter" (random noise) has been added to the score of each subject to avoid overlapping points. The jitter is less than ±0.25 points horizontally. The considerable variation in MBAC above each point score is apparent, and in addition, for total scores 4–18, there are MBAC values lying on both sides of the 0.08% cut-point. In order to be 95% confident that the subject has a MBAC greater than 0.08%, the total score (HGN + OLS + WAT) must exceed approximately 17 (based on the 95% lower confidence limit for predicted MBAC for an individual from the regression of MBAC on total score).

Figure 4 shows the percentage of measured breath alcohol concentration values that are above 0.08% in relation to each of the

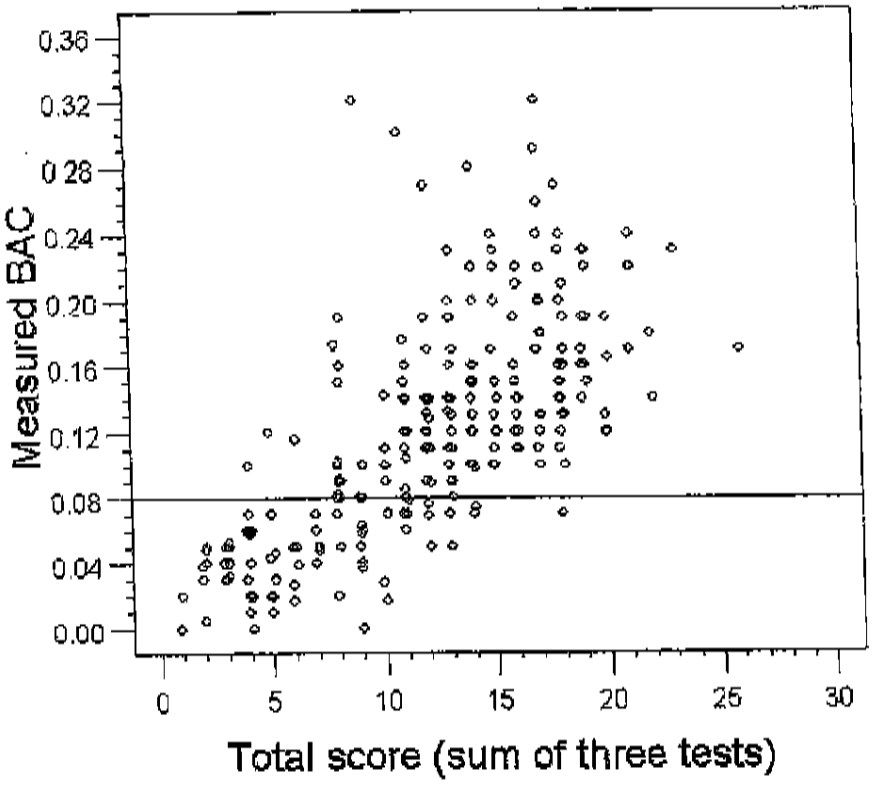


FIG. 3—*Measured breath alcohol concentration versus total of three test scores.*

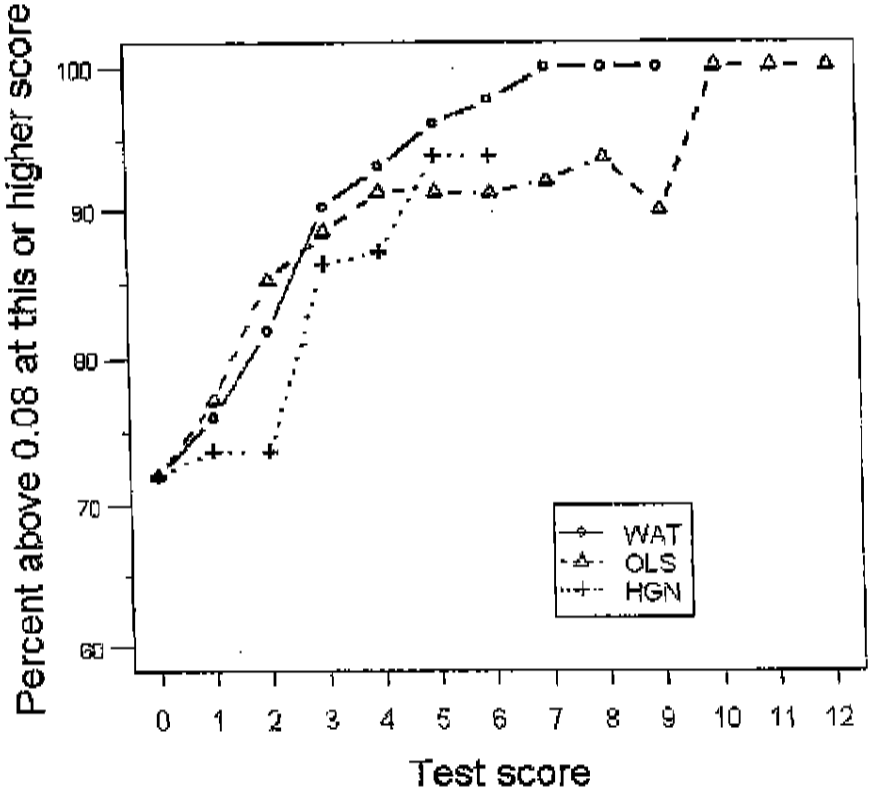


FIG. 4—*Percent of subjects with MBAC greater than 0.08% vs. the individual test score, with the percentage calculated for all individuals at or above the designated score.*

TABLE 1—*Pearson correlation of three Field Sobriety Tests with measured breath alcohol (MBAC) and officer-estimated breath alcohol (EBAC).*

Test	MBAC	EBAC
TOTAL score (3 tests)	0.69	0.74
HGN Horizontal Gaze Nystagmus	0.65	0.71
WAT Walk and turn	0.61	0.64
OLS One leg stand	0.45	0.51

three individual test scores. For each score (horizontal axis), the percent of subjects with that score or higher who have an MBAC larger than 0.08% is plotted (Y-axis). In order to observe 95% of persons with MBAC > 0.08% in this sample, the score for WAT (circles in the plot) must be 5 or larger. None of the scores for HGN (crosses) reach the 95% point and the scores for OLS (triangles) reach over the 95% point only at 10 points and higher, where there are only two subjects. Note that the "failure" scores for these three tests, as specified by Stuster and Burns, are 4 for HGN, 2 for OLS, and 2 for WAT (12). Failure of an FST according to NHTSA standards simply estimate the 50% likelihood that a subject is >0.08%. The data show that in order to be considerably more confident that the subject is above 0.08%, the scores should be much higher than the "failure" scores.

The correlation coefficients for individual tests vs. both MBAC and EBAC are shown in Table 1. The FST with the strongest correlation with MBAC is HGN followed by WAT and OLS. The strongest correlation is with the total test (determined by summing the scores for the three FST's. However, total score and HGN have very similar correlations with MBAC and EBAC.

Discussion

Figure 5, redrawn from Fig. 4 of SBR, illustrates the logic used by Stuster and Burns to describe the accuracy of SFST. A correct decision was registered if both MBAC and EBAC are ≥0.08% (upper, right quadrant; N + 210) or both are ≤0.08% (lower, left quadrant; N = 59). An incorrect decision occurred with a false positive (upper, left quadrant; N = 24), when (EBAC ≥ 0.08% and MBAC < 0.08%) or a false negative (lower, right quadrant; N = 4), when EBAC < 0.08% and MBAC ≥ 0.08%. Because such a large fraction of the points were between 0.08% and 0.33% (N = 214 of the 297 total points) and most of these had a MBAC > 0.12, Stuster and Burn's conclusion that the tests have 91% accuracy was strongly affected by the fact that a majority of points are in this high MBAC range, where correct classification as above 0.08 is more reliable. Of the correct results, 210 data points out of a study total of 297 were in the 0.08% to 0.33% range and 59 were in the 0.000% to 0.079% range. (The accuracy estimated by Stuster and Burns as 91% was calculated from the values in Fig. 2 as (210 + 59)/297 = 0.91). The number of false positives (N = 24) was much greater than the number of the false negatives (N = 4). In the range of data near the 0.08% level, the estimated BAC by these experienced officers overestimates the measured BAC, introducing a bias against the subjects (see Fig. 1). Using EBAC to determine whether the subject MBAC is greater than 0.08% is 100% accurate for all subjects with MBAC > 0.12%. In other words, if the subject is highly intoxicated, the SFST provide an accurate indication. It is not surprising that if the subject is clearly intoxicated, the officers can make this determination. If the MBAC is <0.08%, there is a 24/(24 + 59) = 29% chance of a false arrest (determined from Fig. 2).

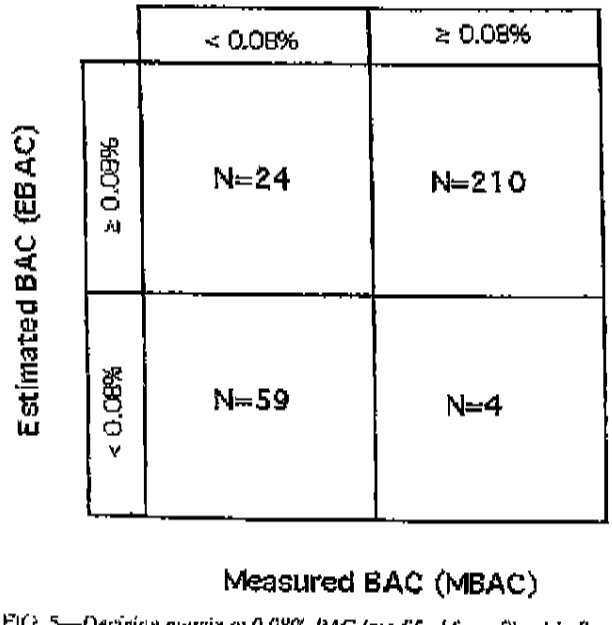


FIG. 5—*Decision matrix at 0.08% BAC (modified from Fig. 4 in Stuster and Burns).*

To illustrate the problem with the SBR statistical strategy, let's apply the same logic to determine the level of accuracy at hypothetical cut-point ("legal limit") levels lower than 0.08%. For example, if Stuster and Burns were to use the same data set to examine the accuracy at lower threshold BAC (0.07% down to 0.01%) levels, they would determine an increasing accuracy level at lower threshold levels. The relative increase in apparent accuracy with decreasing BAC threshold is shown in Table 2, which indicates a hypothetical cut-point for designating a driver as "over the limit." For example, if the legal limit were 0.04%, the SBR method would conclude that SFST are 93.9% accurate. At a legal limit of 0.01%, the SBR conclusion would be that the SFST are 99.3% accurate. The method used by Stuster and Burns has determined a high degree of accuracy simply because most of the data points are at MBACs much greater than the cut-point of 0.08% used in their study. What underlies this problem is the weakness of "accuracy" as the sole performance statistic for this test, as well as the specific nature of this sample, weighted heavily toward individuals with high levels of MBAC.

TABLE 2—*Accuracy of "over-limit" designation based on estimated breath alcohol concentration for defined cut-points (hypothetical legal "limit") of measured breath alcohol concentration (MBAC).*

Legal "limit" (%)	N: All	N: MBAC < cut-point	N: MBAC ≥ cut-point	Accuracy*
0.10	297	107	190	90.6%
0.09	297	97	200	89.2%
0.08	297	83	214	90.6%
0.07	297	69	228	89.6%
0.06	297	58	239	90.6%
0.05	297	43	254	92.3%
0.04	297	29	268	93.9%
0.03	297	19	278	93.9%
0.02	297	9	288	97.6%
0.01	297	4	293	99.3%

Accuracy = 100% (# correctly classified as ≥limit or <limit)/total.

An alternative way to explore the accuracy of SFST is to assess the accuracy over a range of points that is symmetric about the 0.08% cut-point (limit). In addition to accuracy, four traditional statistics of test performance also help in this exploration: sensitivity, specificity, positive predictive value and negative predictive value. Table 3 shows the accuracy of SFST when the range of interest extends above and below 0.08% by the same amount, along with the four traditional performance statistics. For data with MBAC ranging between 0.07% and 0.09%, The SFST are 72.2% accurate. As the range broadens, the calculated apparent accuracy increases. At the broadest range of 0.04%–0.12%, the calculated apparent accuracy is now 82.2%. Taken to the extreme, using all of the data points (MBAC = 0.00% to 0.033%), the apparent accuracy is 91% as calculated by Stuster and Burns. The accuracy of SFST in the vicinity of 0.08% is poorer than estimated in the SBR for the whole data set.

TABLE 3—*Accuracy and other statistics related to "over-limit" designation based on estimated breath alcohol concentration for defined ranges of MBAC.*

Range of MBAC	Total in Range	Accuracy	Sensitivity	Specificity	PPV	NPV
0.07–0.09	36	72%	96%	36%	70%	83%
0.06–0.10	65	75%	95%	44%	73%	85%
0.05–0.11	97	79%	97%	55%	75%	92%
0.04–0.12	135	82%	95%	63%	79%	90%
All cases	297	91%	98%	71%	90%	94%

Accuracy = (# correctly classified as ≥0.08 or <0.08)/total.
PPV = positive predictive value.
NPV = negative predictive value.

Parallel with the reduced level of accuracy in the range 0.07–0.09% MBAC, the four traditional test performance statistics in Table 3 also show varying performance in this range. Specificity is low (36%), indicating that a large fraction of subjects (64%) would be falsely declared over the limit. The sensitivity is excellent in this range, 96%, due to the tendency of EBAC to overestimate alcohol level compared to MBAC. Positive predictive value (PPV) is fair, 70%, indicating that 30% of the subjects declared over-limit would not be so. Negative predictive value (NPV) is good, 83%, indicating that most of those declared under-limit would really be so, but this, again, due to the over-estimation by EBAC. As the range of MBAC in Table 3 steadily widens to finally include all cases, specificity increases to a maximum of only 71%, while sensitivity, PPV and NPV all reach at least 90%, due to predominance in this sample of high levels of measured alcohol.

A closer examination of the data between 0.04% and 0.12% is shown on Fig. 6 (by expanding a section of Fig. 1). Another way of determining the officer's accuracy in estimating the BAC is to compare the fraction of observations (EBAC) overestimating and underestimating the MBAC. If we consider three ranges of MBAC, 0.00% ≤ MBAC < 0.04%, 0.04% ≤ MBAC < 0.08%, and 0.08% ≤ MBAC < 0.12%, the officers' EBAC overestimated the MBAC 76%, 67% and 48% of the time, respectively, estimated it exactly 10%, 7%, and 24% of the time, and underestimated it 14%, 26% and 28% of the time. Overestimation occurs more frequently than underestimation for MBAC < 0.12%. Further evidence of officer overestimation at lower MBAC levels can be taken from the regression line and the line of identity in Figs. 1 and 6. At MBAC values at or below approximately 0.10%, the EBAC tends to overestimate the MBAC. Out of 123 points in this range, 80 overestimate and 26 underestimate the MBAC. For MBAC above

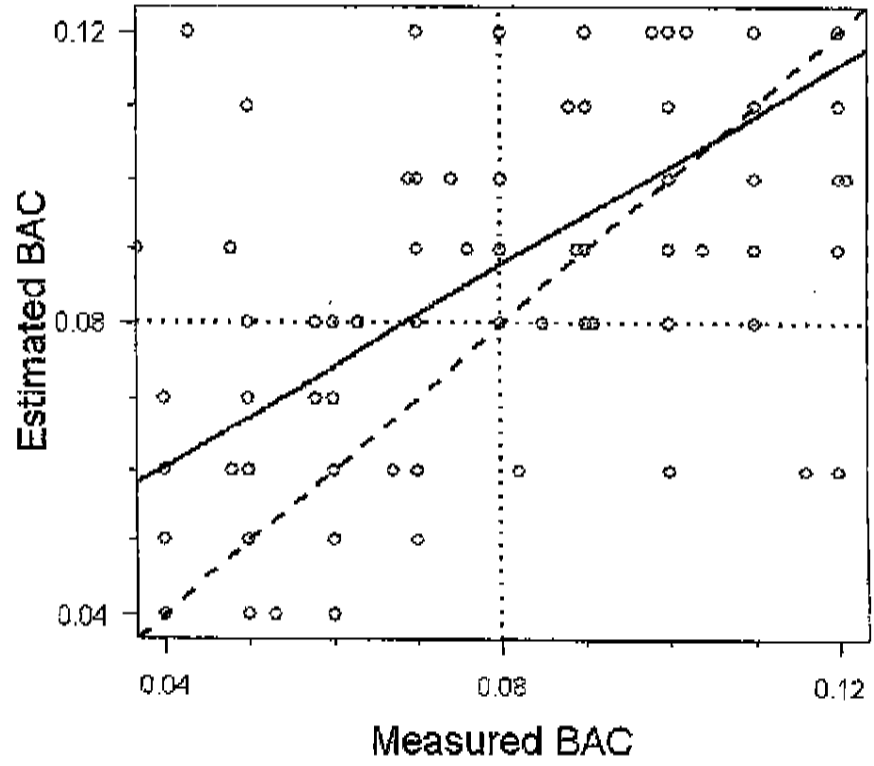


FIG. 6—*Data from Figure 1 expanded to show points between 0.04% and 0.12%. The line of identity (EBAC = MBAC), dashed line and linear regression line (heavy solid line) are shown.*

0.10%, the EBAC values tend to underestimate the MBAC. Out of 174 data points with MBAC >0.10%, 50 are overestimates and 108 are underestimates of MBAC. Thus, the experienced officers used in this study tended to overestimate the BAC at low levels (<0.10) and underestimate the BAC at higher levels (>0.10).

The optimal predictive capability of the SFST depends on the scaling for the particular test and the predictive capacity of the test. The maximum scores *permitted* for HGN, OLS and WAT are 6, 4, and 8, respectively. However, some officers assigned scores that were greater than the maximum score allowable for a given FST. The highest scores assigned in this study were 6, 12, and 9 for the HGN, OLS and WAT, respectively.

By adjusting the weight given to each test and taking account of the precision of the test in predicting MBAC, we find the following linear regression model (Eq 1) maximizes the precision of the SFST for estimating MBAC, using only linear versions of the three test variables. The quadratic terms (squared values of the three test variables), while statistically significant as a group ($p = 0.004$) increase R-squared by only 2%, from 54% to 56%, and have been omitted for parsimony. The model is based on the 261 cases without any missing values for the three tests. Note in the equation below that the increase in BAC per point increase in the score is largest for HGN, with a 0.017 increase in BAC, on the average, for each point increase in the HGN score.

$$\text{MBAC} = -0.007 + 0.017 \times (\text{HGN Score}) + 0.0012 \times (\text{OLS Score}) + 0.011 \times (\text{WAT Score}) \quad (1)$$

The equation does quite well in predicting the mean MBAC, but there is still a large spread of individuals around the predicted value. The standard deviation of individual MBAC values around the predicted regression value is 0.044%. A 95% confidence interval for the true MBAC of an individual, predicted from this regression model, would have a minimum width of ±0.09%, certainly a wide range.

Using the predictors (HGN, OLS, WAT), the additive model from equation 1 accounted for *54%* of the variability in MBAC (corresponding to a correlation of 0.73). Including EBAC as an additional predictor in the model resulted in a *substantial and* significant increase ($p < 0.001$) in the variance of MBAC explained, increasing it to *75%*. As noted earlier, this marked increase in predictability of MBAC by adding in the officer's EBAC indicates that the officers' estimates were probably influenced by factors other than the three FSTs.

We believe that the accuracy of the SFST can be improved if a weighted sum of scores from the three standard tests is combined as described in Eq 1. However, this relationship should be tested in a variety of populations, and, in a larger sample, it is possible that non-linear and other functions of the test scores may help in prediction. The evaluation should include an assessment of accuracy and bias in estimating the numerical BAC and, as well, the accuracy in classifying individuals above or below specified limits (such as 0.08%) for various low, medium and high levels of measured BAC. In follow-up trials of the FST, the instructions given to officers for converting test scores into estimates of BAC should be stated more explicitly (such as using Eq 1 above, or another algorithm). Further, some attempt should be made to identify and incorporate (or control) other factors, aside from the SFST scores, that influence BAC estimates. It may be difficult or impossible to "turn off" other cues that officers use in estimating BAC or in making a decision about an arrest.

The magnitude of the correlations between the tests and MBAC suggests that this type of testing could be developed further, either through re-formulation of the tests, or through different scoring systems, *or* by other means. In the current framework, the test scores have to be quite high to provide confidence that the subject is above 0.08%, but further development could potentially improve confidence in the *three* test results, both singly and in combination. And, anticipating the possibility that some jurisdictions may now or in the future have lower (or higher) legal limits than 0.08%, testing could include more representation from lower levels of BrAC.

The SFST total score and sub-test scores are undoubtedly correlated with breath alcohol level (Table 1). However, predicting a numeric blood alcohol concentration from the SFST scores, as the SFST methodology is defined in the Stuster and Burns report, has limited accuracy and precision. The evidence for this is: a) considerable over- and under-estimation of MBAC (see Results section); b) a large range of observed MBAC values corresponding to any given total SFST score (Fig. 3); and, c) a large spread of *observed* MBAC values around *predicted* MBAC values from a liberal regression model that attempts to optimize the use of the SFST, yet has a minimum prediction uncertainty of ±0.09%.

If our interest is not in quantitative prediction, but in classifying individuals, such as below vs. equal to or above a limit of 0.08%, the utility of the SFST depends very much on how intoxicated an individual is. Accuracy (and specificity) are low when individuals are close to 0.08% MBAC (Fig. 2 and Table 3), but if the individuals are quite intoxicated, such as above 0.12%, then accuracy is high (Fig. 2).

The use of a single test performance statistic, accuracy, and the calculation of this one statistic for the entire study sample is an over-simplification of the more complex relationship between the SFST score and the MBAC level.

SFSTs could become more useful if much more data are accumulated and analyzed using statistical methods such as those presented in this paper, including some of the traditional test evaluation statistics. It is likely that the usefulness of SFSTs will be greatest for drivers who have high-test scores. The moderate to strong correlations between the tests and MBAC suggest a potential for further test development. Enhanced understanding would come from tests applied to a more diverse population sample as well as from the development of a statistical approach to predicting the probability of a subject having a BAC greater than 0.08 % from a particular set of SFST scores.

References

1. Stuster J, Burns M. Validation of the standardized field sobriety test battery at BACs below 0.10 percent. August, 1998. National Highway Traffic Safety Administration.
2. Burns M, Moskowitz H. Psychophysical tests for DWI arrest. Technical Report DOT-HS-5-01242. National Highway Traffic Safety Administration. Washington, DC.
3. Tharp V, Burns M, Moskowitz H. Development and field test of psychophysical tests for DWI arrest. US Department of Transportation, National Highway Traffic Safety Administration Final Report DOT-HS-805-864. Washington, DC.
4. McKnight, AJ, Langston, EA, McKnight AS, Lange, JE. Sobriety tests for low blood alcohol concentrations. Acc Anal & Prevent 2002;34: 305–11.
5. Heishman, SJ, Singleton, EG, Crouch, DJ. Laboratory validation study of drug evaluation and classification program: ethanol, cocaine, and marijuana. J Anal Toxicol 1996;20: 468–81.
6. Cole S, Nowaczyk, RH. Field sobriety tests: Are they designed for failure? Perceptual and Motor Skills 1994;79:99–104.
7. Hlastala M. The alcohol breath test—A brief review. J Appl Physiol 1998;84:401–8.
8. Hlastala M. Invited editorial on "The alcohol breath test." J Appl Physiol 2002;93:405–6.

9. Brice D, Cole C. NHTSA field sobriety tests validation v. invalidation. 25 The Champion. 2001;25:37–42.
10. Fisher LD, van Belle G. Biostatistics. Wiley, 1993.
11. Weisberg S. Applied linear regression, 2nd edition. Wiley, 1985.
12. NHTSA DWI Detection and Standardized Field Sobriety Testing Student Manual, DOT-HS-178-R1/02.

Additional information and reprint requests:
Michael P. Hlastala, Ph.D.
Division of Pulmonary and Critical Care Medicine, Department of Medicine
Department of Physiology and Biophysics
Box 356522
University of Washington
Seattle, WA 98195-6522
E-mail: hlastala@u.washington.edu