Field Sobriety Tests: Reliability and Validity

M.W. Perrine, R.D. Foss, A.R. Meyers, R.B. Voas, and C. Vélez
Vermont Alcohol Research Center
Colchester, VT 05446 USA

1. Introduction

Police officers and prosecutors use standardized field sobriety tests (SFSTs) to establish probable cause for the arrest and subsequent prosecution of motorists suspected of driving under the influence of alcohol (DUI). In our study of alcohol-tolerant drinking drivers, Vermont Alcohol Research Center (VARC) investigators (cf., Meyers, Perrine, & Foss, in this volume) planned to use drivers’ performance on the National Highway Traffic Safety Administration’s (NHTSA) standard FSTs as criteria of alcohol tolerance, i.e., we defined alcohol-tolerant drivers as those whose test performance was acceptable at blood alcohol concentrations (BACs) ≥ 0.10%. To this end, we incorporated into our research the three-part SFST battery (walk-and-turn [WAT], one-leg stand [OLS], and horizontal gaze nystagmus [HGN]) developed by Tharp, burns, and Moskowitz (1981), independently tested (Anderson et al., 1983; Compton, 1985), endorsed by NHTSA (NHTSA 1990), widely used by police officers in the United States, and supported by considerable case law.

First-year (1990) results were troubling. Although research assistants had been trained to NHTSA standards, there were only weak statistical associations between drivers’ BACs and their performance on all three SFSTs. Consequently, in Year 2 (1991), VARC employed five NHTSA-certified SFST experts to train staff according to NHTSA standards in the administration and scoring of SFSTs. This effort notwithstanding, early 1991 data continued to show only a weak relationship between SFST performance and BACs of motorists.

As the BAC test equipment (Intoximeter Alco-Sensor III) was both reliable and valid, there were three alternative explanations for the apparent discrepancies between BACs and test scores: Either (1) research staff had not had adequate training; or, (2) proper training notwithstanding, staff administered the tests poorly, or, (3) there were intrinsic inconsistencies between BAC and performance on SFSTs.

To examine these explanations, project Year 2 included a special study of (1) the validity and inter-rater reliability of the three NHTSA-sanctioned SFSTs as predictors of BAC and (2) of their ability to predict BACs in excess of different criterion levels (0.10%, 0.08%, and 0.05%, respectively). The special study was conducted in conjunction with the normal roadside surveys on 5 consecutive weekends (9 nights) in August and September, 1991. The study was based upon
the premise that experts' performances in field settings—in this case, roadsides—would approximate the tests' fundamental validity.

As part of standard interviewer training in Year 3 (1992), SFST experts provided VARC field staff with mandatory SFST training in May 1992, before the beginning of data collection, and again, in July 1992. This report includes a summary of both the 1991 special study and the 1992 field data, because the latter help to assess the longer-term effects of training upon the reliability and validity of SFSTs. This report concentrates upon test validity. Inter-rater reliability data appeared in an earlier report (Perrine, Foss, Vélez, Voas, & Meyers, 1992).

2. Method

A general description of the VARC roadside research procedures appears elsewhere in this volume (Meyers, Perrine, & Foss, 1992). The special study entailed modifications in subject recruitment and the methods and procedures used to administer and score SFSTs.

2.1 Subjects. Research subjects were all drivers with BACs >0 (measured by Alco-Sensor III), plus the 6% random sample of drivers with BACs=0 selected to take SFSTs (Meyers, Perrine, & Foss, 1992). As previous analyses of the Ohio roadside data indicated that only about 3% of Ohio's nighttime drivers would have BACs ≥ 0.10% (Foss, Perrine, & Wong, 1991), additional volunteer subjects were dosed with alcohol to BACs of approximately 0.12% and then driven by research staff to the survey sites. This procedure helped to increase the statistical power of the analysis. As roadside interviews were conducted out of sight of the SFST area (cf., Meyers, Perrine, & Foss, 1992), SFST administrators could not distinguish actual drivers from dosed volunteers.

2.2 Test Administrators. In contrast to standard VARC procedures, 3 kinds of assessors administered SFSTs: 10 VARC research assistants; 4 SFST experts (all of whom were police officers with substantial experience in DUI enforcement, experience as expert witnesses in DUI cases, and as SFST trainers), and 6 Ohio police officers who were, or who had been, active in DUI law enforcement and who had had NHTSA-specified training in SFST procedures.

2.3 Test Scoring. SFSTs were scored according to NHTSA specifications (1990) to assign one point for each of several possible "clues" (signs of intoxication). Maximum scores are 6 for the HGN, 8 for the WAT, and 4 for the OLS. Data were analyzed in two ways: total numbers of clues on each test and pass/fail scores, based on NHTSA guidelines (failure equals 4 or more clues on the HGN and two or more clues, each, for the WAT and the OLS).

3. Results

A total of 480 subjects participated in the study, 361 drivers (of 2,236 encountered) and 119 dosed subjects. The distribution of BACs was: 0.00%=14%; 0.001-0.049%=33%; 0.05%-0.099%=17%; 0.10%-0.149%=18%; and ≥ .15%=6%. SFST experts performed SFSTs on 185 of these subjects; Ohio officers tested 93, and VARC research assistants tested 202.
3.1 Validity of SFSTs. Performance on the HGN test, administered by the SFST experts, was strongly and monotonically related to respondent BAC (Somers' $d_{xy} = .47$, $p < .001$). As Figure 1 indicates, approximately 3% of individuals with zero BACs tested by SFST experts failed the HGN, whereas all those with BACs $\geq 0.15\%$ ($n = 10$) failed. The failure rate recorded by the experts for subjects with BACs 0.10-.149% was 81%.

In contrast, experts' assessments of performance on the WAT and OLS tests showed much weaker relationships to BAC. Although the relationship between BAC and WAT test failure was statistically significant ($d_{yx} = .13$, $p < .01$), it was not particularly strong. More than half of the individuals with zero BAC failed the WAT test, whereas fewer than half of those between 0.08% and 0.10% failed. Among persons with BACs $\geq 0.15\%$, 78% failed.

![Figure 1 Percent of persons who fail HGN test by subject BAC and administrator experience.](image)

OLS performance also was significantly, but weakly, related to BAC ($d_{yx} = .16$, $p < .01$). Thirty percent of subjects with zero BACs failed this test, whereas 49% of those with BACs between 0.10-0.149% failed, as did 88% (7/8) of those with BACs $\geq 0.15\%$. Thus, the HGN test appears to be a relatively effective measure for detecting drivers at various BAC levels, but neither the WAT nor the OLS appears to be sufficiently valid consistently to identify drivers with BACs $\geq 0.10\%$.

Combining the HGN and WAT test scores, as prescribed in the NHTSA SFST Manual (1990), does not significantly improve the predictive capability of the HGN test alone. Likewise, combining all 3 test scores fails to improve the predictive power that the experts can achieve with the HGN alone.
Figure 1 shows that neither VARC staff nor Ohio police officers were as proficient as the experts in identifying high BAC drivers using the HGN test. VARC staff appear to have failed fewer of the drivers with BAC ≥ 0.10%, but also fewer of those with BACs < 0.10%. Ohio police officers, who applied more liberal criteria, correctly identified 75% of subjects with BACs ≥ 0.10% on the basis of HGN performance. However, they also failed 41% of persons with BACs < 0.10%, a high false positive rate. By comparison, research staff members correctly identified only 57% of persons above 0.10%, but misidentified 13% of those < 0.10% as failing the test. The experience of VARC staff after their intensive pre-season (92-I) and mid-season (92-II) SFST training is especially revealing. They appear to increase true positive assessments, in some cases dramatically (at BACs ≥ 0.05%), but at a cost of increased false positives and only to a level well below that of experts.

Overall, the relationship of HGN scores to subjects’ BACs was significantly lower than the experts for both the police officers (d_{xy}=.47 vs .29, p< .10) and research staff (d_{xy}=.47 vs .24, p< .01), but the associations for officers and research staff did not differ significantly (.24 vs .29, p>.10).

3.2 Field Sobriety Tests at Lower BACs. When administered and scored by SFST experts, the HGN test appears to be approximately as effective for identifying persons above the lower criteria BACs (0.08% and 0.05%) as it is at 0.10% (Figure 2). The kappa values for the HGN test administered by experts generally reflect good performance in detecting persons above any of three BAC levels (.64, .61, and .53 for criterion BACs of 0.05%, 0.08%, and 0.10%
respectively; all $p < .01$). As failure rates on this test climbed sharply for persons with BACs $\geq 0.05\%$, the HGN was actually somewhat better at detecting persons above lower criterion BACs than for the $0.10\%$ BAC. However, these differences are not statistically significant.

Kappa values for the WAT and OLS tests, though statistically significant, are low at all BAC levels, reflecting poor agreement between individuals' test performance and their BACs at any of the criterion limits. For example, although the OLS was able to detect 56% of persons who had BACs in excess of 0.10%, it also resulted in failure by 27% of persons with BACs $< 0.10\%$, an unacceptably large number of false positive identifications. The WAT performs even worse. Although 73% of those persons with BACs above 0.10% failed the WAT, 49% of persons with BACs $< 0.10\%$ (and 44% of those $< 0.05\%$) also failed.

4. Discussion

Our results offer, at best, mixed reviews of effectiveness of the NHTSA battery of SFSTs. Figures 1 and 2 indicate that well-trained, experienced, and highly motivated test administrators can use the HGN test with some accuracy to identify drivers with BACs above not only the current legal limit of 0.10% (in most of the United States), but also potential lower legal limits of 0.08% and 0.05%. In contrast, even when administered and scored by highly-skilled administrators, the two psychomotor tests (WAT and OLS) appear to have relatively little utility in estimating BAC at any of the 3 criterion concentrations (0.05%, 0.08%, or 0.10%).

The data must be interpreted with caution because the study included relatively few police officers and because police officers who use SFSTs for law enforcement purposes have access to much more and a wider range of information than do VARC researchers. Nevertheless, the data do suggest the limited validity of the SFST battery, especially the WAT and OLS; more so, because these results are consistent with those of other, generally less rigorous, studies of the standard FSTs (Tharp et al., 1981; Anderson et al., 1983; Compton, 1985). In fact, the VARC data add credence to the earlier research because they were collected in a field setting, using drivers (predominately) as subjects and having SFSTs scored by a wide range of test administrators with a corresponding range of skill and experience.

Assuming that the data are accurate, they have significant implications for field sobriety testing and the use of SFST results by criminal justice authorities. It appears that there are several possibilities:

1. Abandon all such tests, because even the best of them (HGN) produces substantial numbers of false positive (and negative) results, and move in the direction of "chemical-based" enforcement (cf., Voas & Lacey, 1990).

2. Abandon WAT and OLS and concentrate efforts on better and more consistent use of HGN, the best predictor of BAC.
3. Continue to use all three tests, because OLS and WAT—their limitations notwithstanding—allow officers time to make additional observations of drivers suspected of DUI.

We are not yet prepared to make recommendations among these or other options. Before doing so, we need to know more about how police officers actually use these tests. We also need to know more about SFSTs in the context of other possible indicators of DUI (e.g., erratic driving and clinical signs of intoxication). We believe that, after blood and breath alcohol tests, SFSTs are the most valid and reliable indicators of BAC and impaired driving currently available.

Acknowledgment: This research was supported by PHS grant RO1-AA07876 from the National Institute on Alcohol Abuse and Alcoholism to M.W. Perrine, Principal Investigator.

5. References


