

The psychometrics and science of standardized field sobriety tests, Part 1

By Steven Rubenzer, (www.SteveRubenzerPhD.com)

The National Highway Transportation Safety Administration (NHTSA) standardized field sobriety tests (SFSTs) came under intense scrutiny by the defense community when they went into widespread use in the 1980s. At that time, the scientific literature to support their use was limited to two NHTSA-sponsored laboratory studies¹ and two very modest field studies.² Both the NHTSA researchers and critics pointed out that the tests had not proven themselves in the field and that studies done under roadside conditions were badly needed.

Many critics trenchantly derided the SFSTs and their supporting empirical base and detailed other significant problems.³ In the past seven years, three large-scale field studies have been conducted that potentially address some of the problems noted earlier. Indeed, Marcelline Burns, a primary researcher in the development of the SFSTs, has stated the initial laboratory studies have limited relevance to understanding the use and accuracy of the SFSTs 25 years later in field settings.⁴

Have the subsequent Colorado, Florida, and San Diego SFST field studies rectified the earlier problems? What about research by other researchers or agencies? This column reviews the NHTSA SFST field studies and related works, appraises their impact on the research base for the SFSTs, and reviews the SFSTs' standing as psychological tests in light of current standards.

NHTSA SFST field studies

The original NHTSA laboratory studies examined field sobriety tests as applied to volunteers in indoor, well-lighted conditions. For horizontal gaze nystagmus (HGN), examiners had the benefit of equipment to stabilize the subject's head and a protractor for measuring the angle of onset of nystagmus. Could officers obtain usable or valid results under traffic stop conditions? The field studies were designed to address this question. The first such study was completed in 1981, but encountered such poor cooperation from participating officers that the data were deemed unsuitable for analysis.⁵

Presumably because of this initial negative experience, subsequent field testing locations were chosen largely based on the cooperation and support of the administration and officers that would carry out the testing (".....only agencies that could assume an extremely high level of cooperation and commitment would be recommended for participation."⁶). The officers that would perform SFSTs in the new generation of studies were not reluctant draftees, but volunteers,⁷ SFST instructors⁸ or trained officers who exhibited "genuine interest in the study and eagerness to be selected."⁹

The three major NHTSA field studies consist of investigations carried out in Colorado, Florida and San

Diego in 1995, 1997 and 1998, respectively.¹⁰ The designs of the studies were highly similar, so they will be discussed together. In each, actual traffic stops using the SFSTs were investigated. Police officers were recruited to participate in the study from agencies that supported the research efforts. Officers had previous training and experience in the SFSTs (in the Florida study, all 16 were SFST instructors) and received ““refresher”” training before beginning data collection.

In the Colorado and Florida studies, observers from the study (either researchers or participating police officers) monitored about half of the stops to ensure they observed the study protocols (no use of portable breath tests [PBTs] until after the SFSTs were given and scored) and that the SFSTs were administered correctly. In the Colorado and Florida studies, researchers obtained PBTs on the majority of drivers who were tested but released. This allowed an estimate of false negatives — failures to make an arrest when warranted. The different studies investigated the SFSTs performance at blood alcohol concentration (BAC) levels of .05 percent and .08 percent. All three studies reported that correct arrest decisions based on the SFSTs exceeded 90 percent, with two of the three reporting higher levels of false negatives (erroneous releases).

In all three studies, the proportion of drivers arrested to those tested was quite high — well over 50 percent. Mean BAC level of those arrested were .138 percent (San Diego), .150 percent (Florida), and .152 percent (Colorado). In the Colorado study, HGN was scored differently than in all other studies, as scores for left and right eyes were not distinguished, and the scores ranged only from 0-3. There is no indication of what instructions were given for the ““walk-and-turn”” (WAT) and the ““one-leg stand”” (OLS) in the Colorado and San Diego studies, while the instructions used in the Florida study differ substantially from the 2000 NHTSA Student Manual. The Colorado study reported that only 13 errors of administration and 6 errors in instructions were observed in 305 SFST administrations (only 41 percent were observed). No errors were observed in the 313 SFST batteries given in the Florida study, although only two-thirds of the administrations were monitored.

The NHTSA *Student Manual*,¹¹ the official SFST training guide for police officers, provides cutoff scores for each test to optimally classify a person as above or below .10 percent. It appears that the NHTSA-suggested decision rules for the SFSTs were *not* used in the Colorado and Florida studies — officers had access to test scores but used their own best judgment as the final criterion for arrest. Officers’ failures to follow the recommended SFST-decision rules were cited as a significant problem in the San Diego study. In the Colorado study, incorrect arrest decisions were attributed to officers focusing on poor WAT and OLS performance when the suspects’ HGN performance was normal.

Other studies

Several investigations besides the three NHTSA field studies examined the performance of SFSTs in detecting BAC levels. Two optometrists analyzed the results from 2429 administrations of the HGN test conducted during normal traffic stops in Ohio.¹² They reported results, in the form of a table, that suggest high levels of accuracy (92 percent) for HGN — the other SFSTs were not examined. However, all of the suspects were arrested (even those that passed the HGN), and 92 percent of them had a BAC of

above the .10 percent standard used in Ohio. In other words, the officers would be right 92 percent of the time by arresting everybody (which they did) or by randomly arresting suspected drunk drivers: the test added nothing.¹³ The authors report very few details of the data collection, there were no observers present, and there is no indication whether PBTs were used.

The only NHTSA-sponsored sobriety test studies that have been published in *peer-reviewed* journals detail the development of a standardized *boating* sobriety test¹⁴ and an investigation of various sobriety tests at detecting BAC at .04 percent.¹⁵ The marine environment is unique because the motion of a watercraft makes the WAT and OLS unsuitable for on-the-spot testing. Like the 1981 SFST study, both laboratory and field observations were made. HGN and three other tests were identified as most promising based on their correlation with BAC.

In the field portion of the boating study, the Maryland Department of Natural Resources Police administered the four SBST candidate measures. Officers all had been certified on the SFSTs, were described as ““highly experienced”” regarding DUI/BUI, and were given an additional day and a half of training before beginning the study. Officers were instructed not to obtain PBT readings until after recording the SBST results, but no observers monitored this or administration and scoring procedures. HGN was found the best individual test, correlating .77 with BAC in the field stops. Using HGN scores alone resulted in 100 percent classification of BAC >.10 percent and 90 percent correct classification below .10 percent. Two tests used in the field battery, ““saying the alphabet”” and ““hand-pat,”” showed respectable correlations with BAC but did not improve upon decisions based on HGN alone. The authors nonetheless recommended the full battery because the latter tests provide some measure of performance impairment (vs. BAC level), whereas HGN does not.¹⁶

A very recent investigation¹⁷ found that only HGN was effective at distinguishing persons above or below a BAC of .04 percent, a standard sometimes applied to drivers of commercial vehicles and, in some states, to drivers younger than 21. Both laboratory and simulated field conditions were investigated, and several variations of HGN and the OLS were tried. The variations did not matter much, but the optimum cut-score for HGN was two clues rather than four. Even so, the observed accuracy level obtained was lower than for higher BAC levels: 79 percent of those above .04 percent were correctly identified, while 38 percent of those below .04 percent were wrongly classified.

Critique of the SFST field studies

A scientific study should evaluate the effect of a variable, or a test, controlling for the effects of extraneous variables as much as possible.¹⁸ In the case of the SFSTs, a rigorous test of their validity would be to examine the correct classification rate (*i.e.*, BAC > .08 percent) using only information from the test(s) — not from the suspect’s driving performance, demeanor, smell, previous arrest record, etc. Accomplishing this level of control would probably require video taping only the relevant (officially scored) aspects of SFST performance. The test performance would be scored by officers who had no other information regarding the suspects and no opportunity to observe, smell, or talk to them. A rigorous study

of HGN, probably only feasible in a laboratory study, would involve partial masking of the eyes, so eye redness, glassiness, and eyelid droop could not be observed.

Ideally, subjects in an experiment are randomly assigned to a control or experimental group. In this way, differences between the groups are minimized. The original NHTSA laboratory studies assigned subjects to a target BAC group based on their drinking history. In the field studies, there were no experimentally created groups — just drivers stopped for one reason or another. Therefore, the NHTSA field studies are quasi-experiments, not experiments.¹⁹

All the officers employed the SFSTs and no control group was used. A control group is considered a near-essential feature of a rigorous study because it duplicates all the relevant factors that might account for the results in the experimental group except for the variable under study. In the case of the SFSTs, adjacent jurisdictions might be compared — one department using the SFSTs and another not. Or some members of the department might be trained in the SFSTs, others given other DWI-detection training. Without *some* control group, the results observed are ambiguous. Is 90-95 percent a better accuracy rate than without the SFSTs?²⁰ Was the high accuracy rate due to the quality of the officers? Their sensitization to DWI detection because of their recent training? The fact that they were observed by researchers and supervisors?

Significant defects of the SFST field studies as rigorous scientific studies can be summarized in the following five points:

1. *The field studies validated the arrest decisions of the officers in the studies, not the SFSTs.* Because officers had access to driver behavior and demeanor, the field studies did not specifically test the accuracy of the SFSTs as stand-alone tests. They were not conducted ““blind,”” much less double-blind. As stated in the Colorado study, ““Some of the information underlying an officer’s decision is not documented and cannot be examined.””²¹ In the San Diego and the boating studies, officers may have also had use of PBTs, which would contaminate the test with the criterion — a fatal flaw. Even in the other two studies, large proportions of the stops were unobserved, so officers could have used PBTs before scoring the SFSTs. In sum, the officers’ judgments of intoxication and arrest decisions were not solely due to the SFSTs, and cannot provide solid evidence for SFST validity.

2. *The police officers and the degree of supervision in the field studies were not typical of typical DWI stops.* In each study, participating officers were highly motivated, highly experienced volunteers. In two studies, they were monitored by either civilian research observers or their colleagues. It is well known that people who are watched tend to perform better — in social psychology this is known as the *Hawthorne Effect*. Supervision likely made officers more attuned to accurate administration and recording than an officer working on his own would be. The very low rate of administration errors reported for the Colorado and Florida studies attest to this, and contrasts greatly with the experience of many DUI attorneys.²²

3. *The studies are insufficiently documented for scientific papers, a point made in *United States v. Horn*.*²³ For example, two of the SFST studies do not specify the instructions used to administer the tests

(the instructions have changed considerably since the initial 1977 study). None of the studies examined the combination of HGN and WAT that is referenced in the NHTSA manuals, or examined interrater reliability (how well different observers agreed on scoring or arrest decisions) or internal reliability (how well the different scoring clues agreed). There is no discussion of the weaknesses or limitations of the studies, as is customary in the discussion section of a published paper. Instead, the Florida study ends with an astonishingly strong conclusion: ““There appears to be little basis for continuing legal challenge [to the SFSTs].””²⁴

4. The authors did not report the accuracy of arrest decisions for stops that were observed vs. those that were not, or for SFSTs performed under adverse climate conditions versus those that were not. This is surprising, since this latter issue was a one of the primary goals of the Colorado study.

5. None of the SFST field studies have been published in peer-reviewed scientific journals. The reports were submitted to state DOT agencies or simply ““written up.”” Peer review exposes the work to the criticism of other researchers and authors who may not share the same beliefs and purposes, and who have training and experience in valid experimental design. The scrutiny that this process brings is crucial to detecting error and bias.

Because of the limitations of the field studies cited above, it could be argued that the 1981 laboratory study, and a similar work by non-NHTSA authors,²⁵ remain the primary evidence of SFST reliability and validity. Supporting this claim, NHTSA continues to cite the accuracy figures from the 1981 study in the student manual²⁶ rather than much higher figures obtained in the field studies.

Although the laboratory studies were rigorous in some respects, they have several significant limitations: 1) subjects had no reason to fear detection/arrest; 2) testing was conducted during the day rather than night, when most DWIs occur; 3) officers were able to observe, talk to, and smell the subjects; 4) for the NHTSA study, subjects were recruited from the state employment office and are not representative of the general population, and no attempt was made to justify this source as representative of DWI stoppees; and 5) the same subjects were used to create the cutoff scores for the test and to evaluate the accuracy of these cutoff scores. This procedure will lead to inflated estimates of accuracy, because the test decision rules are tailored to the subjects on which it was calibrated.²⁷ The cutoff rules from the first group should be cross-validated on a new group of subjects. The accuracy level achieved in the second group will be an unbiased estimate of the accuracy when applied to a new group of similar subjects, such as DWI suspects, assuming the base rates (frequency) of intoxicated persons are similar in both groups.

A comment on HGN

Horizontal gaze nystagmus has repeatedly been found in NHTSA-sponsored studies to be the best psychophysiological test to estimate BAC.²⁸ Conducted by medical or optometry personnel in laboratory conditions with healthy, rested subjects, there is little doubt that HGN can be a good indicator of BAC. However, most police officers lack in-depth training, and estimating a 45-degree angle is a poor substitute

for laboratory apparatus that can measure angles to a tenth of degree. Data from the 1981 study indicate that most officers had difficulty accurately estimating 45 degrees,²⁹ which the authors stated ““is a critical factor in making accurate decisions from sobriety test battery performance.””³⁰ Officers were deemed proficient if they could estimate an angle within 3 degrees *with use of a protractor*.³¹ Thus, even when officers are freshly trained and use an apparatus to assist in their observations, a 6-degree range of error is expected. One of the clues for HGN is onset of nystagmus before 45 degrees of lateral deviation. If a 6-point spread is acceptable, one officer may estimate 45 degrees at 42 degrees, another at 48. If the officers are consistent in their scoring, the first officer will score this clue much less often than the second will.

Difficulties can arise in several other ways when interpreting HGN. Are a subject’s eye movements smooth pursuit movements with nystagmus or natural saccadic movements? At least one board-certified ophthalmologist wrote that NHTSA’s recommended ““smooth pursuit”” administration (two seconds across each eye) invites saccadic movements because it requires the eye to move too fast.³² The 1981 study authors acknowledged that as many as 50 percent of people show some nystagmus at maximum deviation in at least one eye.””³³ In *New Hampshire v. Dahood*, the court reported ““Drs. Citron (an ophthalmologist) and Rizzo (a neuro-ophthalmologist) were adamant in their opinion that the distinct nystagmus at maximum deviation clue should be eliminated from the HGN test.””³⁴ Recently, it has been reported that fatigue can induce nystagmus at maximum deviation in 50 percent of people, and that nystagmus persists after BAC levels have fallen to zero.³⁵ Lastly, the Maryland Court of Appeals in *Shultz v. State* recognized 35 causes of nystagmus in addition to alcohol.³⁶

Two recent court opinions have held that HGN does not meet *Daubert*³⁷ standards to be admissible as direct evidence of intoxication or impairment. In *United States v. Horn*, the court held HGN is not generally accepted among psychologists.³⁸ In *New Hampshire v. Dahood*,³⁹ the trial court, on remand from the New Hampshire Supreme Court on the issue of admissibility, cited an inability to determine error rates and concluded HGN is *not* generally accepted among ophthalmologists. On appeal, however, the New Hampshire Supreme Court held that HGN does meet the four *Daubert* criteria, and reaffirmed other state court opinions that the relevant professional communities for HGN include behavioral psychology, highway safety, neurology, and criminalistics in addition to optometry and ophthalmology.⁴⁰

However, it maintained that HGN is only circumstantial evidence of impairment and cannot be introduced at trial to estimate BAC.

SFSTs as standardized tests

SFSTs are quite similar to the neuropsychological tests, which detect brain damage and assess sensory, motor, and cognitive impairment. To the extent that the SFSTs are standardized tests, they should meet the relevant professional standards. *Standards for Educational and Psychological Testing*⁴¹ is an authoritative guide that enumerates many criteria for test construction, reliability, validity, documentation, and implementation, and provides a useful introduction to these issues. Some of these are directly relevant for the SFSTs. For example, Standard 1.10 states, ““When interpretation of performance on

specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretations should be provided.” This is not addressed in the SFST literature. Table 2 lists the standards that are most relevant for examination of the SFSTs. Part 2 in the next issue addresses problem areas regarding standardization, reliability, and validation.

Notes:

1. Marcelline Burns & Herbert Moskowitz, *Psychophysiological Tests for DWI Arrest*, Final Report, DOT-HS-802-424 (1977). V. Tharp *et al.*, *Development and Field Test of Psychophysiological Tests for DWI Arrest*, Final Report, DOT-HS-805-864 (1981).
2. V. Tharp *et al.*, *supra*. Theodore E. Anderson *et al.*, *Field Evaluation of a Behavioral Test Battery for DWI*, DOT-HS-806-475 (1983).
3. William A. Pangman, *Horizontal Gaze Nystagmus: Voodoo Science*, 2 *DWI J. Law & Sci.* 1 (1987). G. Simpson, *Attacking NHTSA’s Three-Test Field Sobriety Assessment*, 3 *DWI J. Law & Sci.*, 97. Jonathon D. Cowen & Susannah G. Jaffee, *Field Sobriety Tests: The Flimsy Scientific Underpinnings*, 5 *DWI J. Law & Sci.* 121 (1990). Mark Rouleau, *Unreliability of the Horizontal Gaze Nystagmus Test*, 4 *Am. Jur. POF* 3d 439 (1990). Jonathon D. Cowan & Susannah G. Jaffe, *Proof and Disproof of Alcohol-Induced Impairment Though Evidence of Observable Intoxication and Coordination Testing*, 9 *Am. Jur.*, *POF* 3d, 459 (1990). Ronnie M. Cole & Spurgeon N. Cole, *New Proof That Field Sobriety Tests Are “Failure-Designed,”* 6 *DWI J.: Law & Sci.* 1 (1991). Spurgeon Cole & Ronald H. Nowaczyk, *Field Sobriety Tests: Are They Designed for Failure?* 79 *Percep. & Motor Skills*, 99 (1994). Randy T. Leavitt, *Horizontal Gaze Nystagmus*, 22 *Voice for the Defense* 17 (1994). Ronald H. Nowaczyk & Spurgeon Cole, *Separating Myth From Fact: A Review of Research on the Field Sobriety Tests*, *The Champion* (Aug. 1995) 40. Charles R. Honts & Susan L. Amato-Henderson, *Horizontal Gaze Nystagmus Test: The State of the Science in 1995*, 71 *N. Dak. L. Rev.* 671 (1995). Joseph R. Meaney, *Horizontal Gaze Nystagmus: A Closer Look*, 36 *Jurimetrics J.* 383 (1996).
4. Marcelline Burns, First Annual DWI Training Seminar, Houston (2000).
5. V. Tharp *et al.*, *supra*.
6. Jack Stuster & Marcelline Burns, *Validation of the Standardized Field Sobriety Test Battery at BACs Below .10 Percent*, DOT-HS-808-839 6 (1998).
7. Marcelline Burns & Ellen W. Anderson, *Field Evaluation Study of the Standardized Field Sobriety Test (SFST) Battery* (Final Report Submitted to the Colorado DOT, November, 1995).
8. Marcelline Burns & Teresa Dioquino, *A Florida Validation Study of the Standardized Field Sobriety Test (SFST) Battery*, (1997).
9. Jack Stuster & Marcelline Burns, *supra* at 8.
10. Marcelline Burns & Ellen W. Anderson, *supra*. Marcelline Burns & Teresa Dioquino, *supra*. Jack Stuster & Marcelline Burns, *supra*.
11. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, *DWI Detection and Standardized Field Sobriety Testing*, Student Manual (2000).
12. Gregory W. Good & Carol R. Augsburger, *Use of Horizontal Gaze Nystagmus as a Part of Roadside Field Sobriety Testing*, 63 *Amer. J. Optometry & Physiological Optics*, 467 (1986).

13. See Louis M. Hsu, *Diagnostic Validity*

Statistics and the MCMI-III, 14 *Psych. Assess.* 410, 410-411 (2002).

14. A. James McKnight et al, *Development of a Standardized Boating Sobriety Test*, 31 *Accid. Anal. & Prev.* 147 (1999).

15. A. James McKnight et al., *Sobriety Tests for Low Alcohol Blood Concentrations*, 34 *Accid. Anal. & Prev.* 305 (2002).

16. A. James McKnight et al., *Development of a Standardized Boating Sobriety Test*, 31 *Accid. Anal. & Prev.* 147, 152 (1999).

17. A. James McKnight et al., *Sobriety Tests for Low Alcohol Blood Concentrations*, 34 *Accid. Anal. & Prev.* 305 (2002)

18. Thomas D. Cook & Donald T. Campbell, *Quasi-Experimentation: Design & Analysis Issues for Field Settings* 2—9 (1979).

19. *Id.*, Donald T. Campbell & Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (1962).

20. See also Phillip E. Price & Spurgeon Cole, *NHTSA Field Sobriety Tests: Validation and Invalidation*, *The Champion* (Apr. 2001).

21. Marcelline Burns & Ellen W. Anderson, *supra* at 17.

22. See also J.L. Booker, *End-Position Nystagmus as an Indicator of Ethanol Intoxication*, 41 *Sci. & Justice* 113 (2001).

23. *United States v. Horn*, 185 F.Supp.2d 530, 558 (D.Md. 2002).

24. Marcelline Burns & Teresa Dioquino, *supra*, at 31.

25. Jack E. Richman & John Jakobowski, *The Competency and Accuracy of Policy Academy Recruits in the Use of the Horizontal Gaze Nystagmus Test for Detecting Alcohol Impairment*, 47 *New Eng. J Optom.* 5 (1994).

26. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, *DWI Detection and Standardized Field Sobriety Testing*, Student Manual (2000) at VIII-8, VIII-12, VIII-14.

27. Elazar J. Pedhazur, *Multiple Regression in Behavioral Research* 147-150 (2nd ed. 1982). Jum C. Nunnally & Ira H. Bernstein, *Psychometric Theory* 333 (3rd ed.1994).

28. Marcelline Burns & Herbert Moskowitz, *Psychophysiological Tests for DWI Arrest*, Final Report, DOT-HS-802-424 (1977). V. Tharp et al., *supra*. A. James McKnight et al., *Development of a Standardized Boating Sobriety Test*, 31 *Accident Analysis and Prevention* 147 (1999). A. James McKnight et al., *Sobriety Tests for Low Alcohol Blood Concentrations*, 34 *Accid. Anal. & Prev.* 305 (2002). See also Antti Penttila & Martti Tenju, *Clinical Examination as Medicolegal Proof of Alcohol Intoxication*, 16 *Med. Sci. law* (1976) 95. Robert S. Kennedy et al., *Indexing Cognitive Tests to Alcohol Dosage and Comparison to Standardized Field Sobriety Tests*, 55 *J. of*

Studies on Alcohol 615 (1994).

29. V. Tharp et al., *supra*, at 30, 31.

30. *Id.* 30.

31. *Id.* 16.

32. Joseph Citron, MD, *HGN: How to be Your Own Expert Witness* (2002) (unpublished manuscript, <http://www.ncdd.com/>

dsp_bookstore.cfm).

33. V. Tharp *et al.*, *supra* at 7.

34. New Hampshire v. Dahood, #96-JT-707 (Concorde District Court, April 2002) at 11. (New Hampshire Supreme Court remanded the case to the Concord District Court to hold an evidentiary hearing and rule whether the HGN test incorporates scientific principles. If so, the court was to make findings as to the reliability of the test under New Hampshire Rule of Evidence 702).

35. J.L. Booker, *supra*.

36. Schultz v. State, 106 Md.App. 145, 664 A.2d 60 (1995).

37. Daubert v. Merrell Dow Pharmaceuticals, 113 S.Ct. 2786 (1993).

38. *United States v. Horn*, 185 F.Supp.2d at 557.

39. New Hampshire v. Dahood, *supra*.

40. New Hampshire v. Dahood, (No. 99-510, December 20, 2002).

41. American Education Research Association, The American Psychological Association, and National Council on Measurement in Education, Standards for Education and Psychological Testing (1999).ⁿ

From *The Champion*

June 2003, Page 40

The psychometrics and science of the standardized field sobriety tests (Part 2)

By *Steve Rubenzer* (www.SteveRubenzerPhD.com)

The psychometrics and science of the standardized field sobriety tests (Part 2)

Standardization problems—As the name implies, the SFSTs gain their special status because they have been *standardized*, meaning specific rules for administering, scoring, and interpretation have been specified and researched.

Standardization is crucial if research findings are used to support the validity of the tests, since a test that is modified is no longer the same test. As the National Highway Transportation Safety Administration (NHTSA) states, ““If any one of the standardized field sobriety test elements is changed, the validity is compromised.””¹ A number of courts have held that if not properly administered, the SFSTs are not admissible.²

The following problem areas are organized in the chronological order that the SFSTs are administered and scored.

1. *Screening questions for possible medical problems and conditions should be standardized and validated.* The NHTSA Student Manual states the officer should ask about certain topics, but does not specify the form of the questions. The wording of a question, and how it is asked, are crucial to obtaining valid data. Screening questionnaires are used in a variety of medical fields. A good screening test should identify virtually everyone who has the condition being queried about — and should be demonstrated to do so. In the case of the SFSTs, the questions should uncover relevant conditions that could invalidate or affect SFST performance. No research has been conducted on this issue.

2. *The SFST instructions have changed repeatedly from the initial laboratory studies to the field studies to the current NHTSA Student Manual used to train police officers.*

3. *SFST training does not emphasize rigorous adherence to the standardized instructions.* Psychologists routinely administer standardized tests. Many, like the Wechsler intelligence tests, come with materials that direct the examiner to read the instructions verbatim. This was my expectation when I learned the SFSTs. Although the NHTSA instructions are given in quotation marks, suggesting they should be delivered verbatim, this level of proficiency is not specifically endorsed. Consequently, students and instructors do not seem to aspire to it. Some training films actually demonstrate inaccurate delivery.³

4. *SFST training materials do not address how instructions are to be delivered (attitude, speed, and tone).* Should the officer be polite? Authoritative? Commanding? Is it all right to be impatient, surly, and condescending? How does this affect performance? What about speed of delivery? Should the officer's demeanor facilitate maximum performance? That is the usual standard for neuropsychological tests. 4 In contrast, some officers appear to make the tests harder by delivering instructions in a rapid, bored, monotone voice.

It is unlikely that the officers in the laboratory studies, using volunteers and monitored by the researchers, adopted the hostile, impatient demeanor sometimes displayed by officers during SFST administrations. To the extent that arresting officers behave differently than the officers in the NHTSA studies (which was not recorded), the validation evidence is diminished.

5. *For the Walk-and-Turn, a variety of line situations are permitted.* There is no research on the effect of using an imaginary line, a crooked line, an offset line, or one that the line creates an uneven surface.

6. *What constitutes "demonstrates understanding"?* For the WAT and One-Leg-Stand (OLS), officers are directed to determine that the suspect understands the instructions. A "yes" or "no" question often suffices. If a suspect equivocates, the officer may become impatient and demand an answer. Clearly, this is not an adequate assessment. The tests are designed to test ability to follow directions and perform after the instructions are understood. (Standard 9.3)

7. *Scoring rules are often inadequately specified.* What constitutes an "inappropriate turn?" In horizontal gaze nystagmus (HGN), the examiner must make two passes for each eye to assess each of the three signs. Does the clue have to occur on both passes, or just one? If it occurs on just one, should the examiner administer another pass and make a decision based on two out of three?

8. It is unclear, both in the studies and the Student Manual, what the criteria are for failing the SFST battery. The Student Manual provides cutoff scores for each test, plus a decision grid for the combination of the HGN and WAT. What it does not say is what criterion is primary. Thus, a suspect apparently can fail at least four ways (from each of the three tests and from the combination of the HGN and WAT). If the defendant is given multiple chances of failing, the risk of a false positive finding will accumulate with each additional test unless credit is given for those tests passed.

9. *Officers are not specifically directed to record their observations immediately.* Failure to do so encourages a tendency to assign scores consistent with the officer's arrest decision and, for example, to remember seeing a particular clue in both eyes rather than one. As the authors of the 1981 laboratory study stated, ".....many of the advantages of standardized scoring are lost when the scoring is left to memory."⁵

Reliability and validity problems

1. The SFSTs have not been subjected to a rigorous "blind" assessment of their validity. As discussed above, none of the studies of the SFSTs have been truly double blind, as expected in medical research. The laboratory studies came close; the field studies do not. (Standard 1.17)

2. The effects of fatigue, drowsiness, circadian rhythm, driver stiffness or roadside conditions on SFST performance have not been adequately investigated. (Standard 10.1) The angle of onset of nystagmus was found to advance five degrees in the hours after midnight, while the other laboratory studies were conducted during daytime hours.⁶ In the 1981 study, the authors stated that exercise, sleep loss, elevated temperatures, and antihistamines are associated with increased body sway.⁷ Strobe and emergency lights, gusts of wind from passing traffic—all have unknown effects on SFST performance and validity given the limitations of the field studies.

3. Drivers suspected of DWI and subjected to the SFSTs may be highly anxious, which alone or in combination with small amounts of alcohol, may influence their performance. In the laboratory studies, subjects were volunteers who had no reason to be anxious, aside from possible self-consciousness. There are theoretical reasons to believe that fear, anxiety, or stress may affect performance on the WAT and OLS,⁸ and no study has demonstrated these factors are not relevant.

4. *The clues for the WAT and OLS lack documentation of their individual validity and reliability.* The validation and reliability data focus solely on the total scores, not the individual clues. For the WAT, it is possible that all eight clues are valid — or that half of them are not. Since there is no published data on this issue, it cannot be assumed that the clues your client failed are valid ones. (Standard 1.10)

5. *Reliability data are lacking or below accepted standards for psychological tests used for making decision about individuals.* Reliability refers to the consistency with which a test produces results across conditions that can change, such as testing at different times or by different evaluators. Authorities recommend such tests show “a bare minimum” reliability of .90, with .95 “considered the desirable standard.”⁹ None of the reliability figures for the SFSTs are this high, and most are much lower. Different raters scoring the same subject at the same time show reliability coefficients between .62 and .74 on the SFSTs, and lower figures (.58-.59) for their decisions about whether the person is impaired and should be arrested. Other NHTSA researchers assessed the SFSTs to be quite low on “Ease of Scoring,” providing ratings on a 1-100 scale of 5 for HGN, 25 for WAT, and 30 for OLS.¹⁰ No figures have been reported to assess the internal reliability (coherence) of the SFST items. This is a standard, expected piece of information for a psychological test.

The reliability coefficients are estimates of how much of the test score is reliable — a reliability coefficient of .70 indicates 70 percent of the score is reliable and 30 percent is error. However, each reliability coefficient reflects only some of the potential sources of error: The observed score is a function of the quality that is being measured (intoxication) plus numerous sources of error, including who administered the test, the particular occasion and conditions it was administered under, and the quality of the items composing the test. Unfortunately, you cannot simply add up the errors from the different reliability estimates. However, one dramatic illustration of the role of multiple sources of error comes from the 1981 study: The test-retest coefficient for the WAT scored by a different rater is .34, as opposed to .61 when scored by the same rater. The moderate reliability figures cast doubt on the high accuracy rates reported in the field studies, since high reliability is a prerequisite for high validity.¹¹

6. *Standard errors of measurement (SEM) are not provided.* (Standard 6.5) The standard error of measurement is the average amount of error in the typical measurement for that test. The SEM is used to create confidence intervals around an observed score to show how precise the estimate (observed score) is. For example, a 95 percent confidence interval around a score of 4 on the HGN might be 2 to 6. But NHSTA studies do not include basic descriptive statistics of the data (means and standard deviations) that would allow calculation of these values.

7. *SFSTs have not been normed on sober people.* As acknowledged in the 1981 study, “Balance tests of various sorts show large individual differences in the performance of sober individuals.....”¹² When

most psychological tests are developed, they are tested on a large sample to determine what is “normal.” The Personality Assessment Inventory is a self-report test designed to assess psychopathology. Before it was published, the author administered it to some twelve hundred psychiatric patients — the intended population for the test. But he also administered it to over twelve hundred volunteers from around the country. Then, volunteers were dropped in order to obtain a census-projected nationally representative sample in terms of age, race, and education.¹³ The SFSTs have never been administered to a large, representative group of sober people. There is no “normal” score.

8. *There is very limited data on the SFSTs for people under 21 or over 50-55. (Standard 3.6)* Only 3.1 percent of the NHTSA 1981 study sample used to standardize, calibrate, and validate the SFSTs were older than 55. Reporting of age groups is inconsistent across the field studies, but in all three, people above 50-60 made up a very small portion of the sample. There have been no comparisons made of the validity of the SFSTs for younger versus older groups. (Standards 7.2, 7.3, 10.1)

9. *SFSTs have questionable validity for those who are elderly, in poor physical condition, or overweight.* If the SFSTs are of questionable validity for people more than 50 pounds overweight,¹⁴ what about short people who are 45 or 40 pounds over the ideal? Proportionately, a person who is 4’8” and 40 pounds overweight is likely to be more physically impaired than someone 6’3” and 51 pounds overweight. Why does the test suddenly become invalid when one goes from 50 to 51 pounds over the ideal? Obviously, the impediment due to weight is likely to be gradual. The same issue applies to people in their late 50’s versus the arbitrary cutoff of 6015 or 65.¹⁶ Physical health and condition are likely to be more important than age. (Standards 7.2, 7.3, 10.1)

10. Even NHTSA claims the SFSTs, when optimally used, are only 80 percent accurate.¹⁷ This is perhaps the most direct and compelling evidence of the SFST validity problems. Although a 20 percent error rate may be acceptable in a test used for evidence of probable cause of a BAC of .08 percent or more, it seems insufficient when the SFSTs are used as to establish, beyond a reasonable doubt, intoxication or impairment. Further, consider that the SFSTs were (1) evaluated by the tests’ developers, (2) under laboratory conditions, (3) only a fraction of subjects were in the critical .05-.15 percent BAC range, and (4) the same subjects used to calibrate the tests were used to assess their accuracy. Given all of these potential biases in their favor, a hit rate of 80 percent is unimpressive.

Another perspective on SFST accuracy is provided by using a bathroom scale as an analogy. Even a cheap scale might be expected accurate within a few pounds. Yet, the NHTSA authors state, “[I]t is unrealistic to attempt to use behavioral tests to discriminate BACs in a $\pm 0.02\%$ margin around a given level.”¹⁸ This is equivalent to a 100-pound woman stepping on a scale, seeing a reading of 120, and being told the scale is functioning within its design limits. And this is under ideal conditions. But how well can police officers actually estimate individuals’ BACs? In the 1981 laboratory study, police officers’ estimates of BAC (measured by Intoximeters) were incorrect by an average of .03 percent ¹⁹ — meaning approximately half the errors were larger than this.

Psychologists often calculate confidence intervals to communicate that a given score, like an IQ, is an imprecise measurement. For example, an IQ of 100 may have a confidence interval of 94 to 106. If someone obtained an IQ of 100 on one occasion, it is likely that he or she would obtain a score within the confidence interval if tested again. Confidence intervals are not absolute, but based on probability. The most common probability used is 95 percent, meaning that on 95 of 100 retests, the new score would fall within the confidence interval created from the first score.

Let’s return to the analogy of a 100-pound woman stepping on a bathroom scale using the SFST BAC estimation errors. Using the most conservative average error reported (.03 percent), and using standard tools to create a confidence interval,²⁰ we find that a 100-pound woman would observe a scale reading of between 25 and 175 pounds on 95 of 100 trials. The other five percent of readings would be more

inaccurate. In the 1981 field study, officers' average BAC estimates were off by an incredible .077 percent before training and a whopping .0537 percent after training.²¹ Creating a 95 percent confidence interval from the "before training" figure (.077 percent) means our 100-pound woman will weigh anywhere from -93 to 293 pounds on our SFST bathroom scale — 95 percent of the time.

Miscellaneous issues

1. The SFSTs have been evaluated primarily by NHTSA-supported researchers, with no rigorous evaluation by disinterested researchers in a field settings. Replication by impartial researchers is the *sine qua non* of reliable scientific knowledge.

2. *SFSTs have usually been evaluated in high base rate settings where up to 92 percent of the persons tested were legally intoxicated.* Base-rates have a major effect on the confidence that can be given to a test result.²² In both the laboratory and field studies, the majority of subjects or drivers tested were intoxicated so generalization to settings (sobriety checkpoints or daytime stops) where the incidence of DUI is much lower is not warranted. An earlier NHTSA study²³ showed high rates of false positives when the frequency of intoxicated (BAC > .10 percent) drivers was experimentally set to 48 percent. HGN, either alone or in combination with observations of driver behavior and appearance, showed false positive rates of up to 75 percent for those with a BAC between .05 percent and .09 percent. Officers who received only three hours of training in administration of HGN²⁴ assessed 24 percent of those in the .00-.04 percent BAC range as impaired — and the majority of these were probably completely sober.

3. *SFST scoring is potentially biased by the officer's suspicion of intoxication.* The SFSTs require subjective judgment to score, as acknowledged by Marcelline Burns,²⁵ NHTSA reviewers,²⁶ and as indicated by their moderate inter-rater reliability coefficients. An officer could easily decide a WAT turn is improper based, in part, of how the driver smelled and his clarity of speech. When these biases seep in, the test has been contaminated.

4. *The SFSTs may be harder than driving.* The WAT and OLS are unfamiliar and probably strain many sober peoples' abilities, especially those that are not in good physical condition. To quote the NHTSA student manual, "Tests that are difficult for a sober person to perform have little or no evidentiary value."²⁷ A recent survey of British police surgeons found about half expressed concern about the SFSTs being too difficult or the grading too harsh. Amongst those with advanced credentials (a Diploma of Medical Jurisprudence or Diploma of Forensic Medicine) over 60 percent of respondents expressed reservations for the Walk and Turn and One Legged Stand.²⁸

5. *Although the SFSTs were not designed as indications of driving impairment and have undergone little validation for this purpose, they are still frequently admitted as evidence for establishing the driver was impaired.* The SFSTs were expressly developed and validated to distinguish between BACs of above and below .10 percent — not driving impairment. Marcellin Burns has emphasized this distinction,²⁹ but NHTSA materials³⁰ and court decisions³¹ wrongly equate the two terms. While the SFSTs attempt to gauge BAC, NHTSA plainly states "Impairment varies widely among individuals with the same BAC level."³²

Only a couple of studies have attempted to correlate SFST scores with driving impairment. In one of the original NHTSA laboratory studies, subjects were tested on both the SFSTs and a divided attention performance test designed to simulate the demands of driving. Each SFST correlated about .30 with the different performance measures. When the results of the tests were combined statistically, the two psychomotor tests (WAT and OLS) carried all the weight — HGN added nothing.³³ Other NHTSA-associated researchers stated "there is no evidence that the eye movements that constitute Nystagmus seriously impair the visual processes involved in driving or operating a boat."

34 A third group of NHTSA researchers evaluated dozens of behavioral tests to determine their potential to assess driver impairment and other desired qualities. These authors recommended a completely different battery than the SFSTs,³⁵ and the SFSTs received low ratings of relevance to driving skills: HGN received a rating of 0 (zero on a scale of 0-100) for its value in assessing driver impairment, while WAT and OLS received ratings of 40 and 20, respectively.³⁶

6. *SFSTs, particularly HGN, arguably are more prejudicial than probative on the issue of impairment.* When a person suspected of DWI/DUI has difficulty performing a field sobriety test, the jury viewing the performance may logically assume the suspect is drunk. Given the circumstances, this is the natural interpretation. A study published by Cole and Nowacyk³⁷ had 21 completely sober people perform the sobriety tests (not including HGN) and other tasks. Police officers perceived 46 percent of the subjects performing sobriety tests as drunk and worthy of arrest. If prejudicial value = high and probative value = low or medium, then high > low or medium = nonadmissible. Meaney argued that a negative HGN is more probative than a positive finding because ““no study suggests the possibility of intoxication without nystagmus.””³⁸

Conclusion

The SFSTs claim to be standardized and validated psychological tests. The first claim is justified if they are administered, scored, and interpreted in line with NHTSA guidelines. Much more serious questions arise regarding their validation and other psychometric properties. The SFSTs have been evaluated primarily by their proponents and there have been no studies of the SFSTs as a group in either laboratory or field studies by disinterested researchers. Just as important, NHTSA training has not encouraged officers to consider other plausible causes of poor performance, such as anxiety or sleepiness.

The SFSTs have significant limitations as tests that should be understood by those who encounter them in the legal arena. Like most tests, they can be useful, but are also easily abused and misunderstood. Defense attorneys must challenge their empirical bases where they can and expose failures to follow the standardized instructions. More importantly, prosecutors and judges need to critically examine the SFST evidence offered in DUI cases so that innocent people are not wrongly convicted.

Notes

1. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000) at VIII-3.
2. *Emerson v. State*, 880 S.W.2d 759 (Tx.Cr.-App., 1994), *Homan v. State*, 732 N.E.2d 952 (Ohio 2000).
3. *Crime to Court: Rappin' Up the DUI* (instructional video), South Carolina ETV, in cooperation with the South Carolina Criminal Justice Academy, SC Law Enforcement Division (1995). *The Truth Is in the Eyes* (instructional video) cited in *New Hampshire v. Dahood*, *supra*.
4. Muriel D. Lezak, *Neuropsychological Assessment* 139-140 (3rd ed., 1995).
5. V. Tharp *et al.*, *Development and Field Test of Psychophysiological Tests for DWI Arrest*, Final Report, DOT-HS-805-864 (1981) at 70.
6. *Id.* at 17.
7. *Id.* at 83.
8. Phillip B. Price, Sr., *Fear and Sobriety Testing* (2000) (unpublished manuscript, available from Mr. Price's office, 217 Randolph Avenue, Huntsville, AL 35801, 256-536-6000, dwilawyer@aol.com).
9. Jum C. Nunnally & Ira H. Bernstein, *Psychometric Theory* 265 (3rd ed. 1994).
10. K.J. Snapper *et al.*, *An Assessment of Behavioral Tests to Detect Impaired Drivers*, Final Report, DOT-HS-806-211, (1981) at 3-34 to 3-37.
11. Jum C. Nunnally & Ira H. Bernstein, *Psychometric Theory* 214 (3rd ed.1994). Hoi K. Suen, *Principles of Test Theories* 141 (1990).
12. Tharp *et al.*, *supra* at 83.
13. Leslie Morey, *Personality Assessment Inventory Professional Manual* (1989).

14. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000)
15. *Improved Sobriety Testing*, DOT-HS-806-512 (1984) at 7.
16. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000).
17. *Id.* at VIII-12.
18. Marcelline Burns & Herbert Moskowitz, *Psychophysiological Tests for DWI Arrest*, Final Report, DOT-HS-802-424 (1977) at 41.
19. V. Tharp *et al.*, *supra*, at 72.
20. A confidence interval is set using the *standard deviation* rather than the *average deviation*. For a normal distribution, the standard deviation equals 1.25332 * average deviation. (see R. J. Senter, *Analysis of Data: Introductory Statistics for the Behavioral Sciences* 92 (1969). A 95 percent confidence interval is set by taking the mean +/- twice the standard deviation.
21. V. Tharp *et al.*, *supra*, at 63.
22. See Louis M. Hsu, *Diagnostic Validity Statistics and the MCMI-III*, 14 *Psych. Assess.* 410, 410-411 (2002).
23. Richard P. Compton, *Pilot Test of Selected DWI Detection Procedures for Use at Sobriety Checkpoints*, National Highway Traffic Safety Administration, DOT-HS-806-724 (1985).
24. Three hours of training in administration of HGN may not be atypical — most of the 24 hour NHTSA student course is devoted to topics other than administration of the SFSTs. The Texas A&M University System TEEEX — Law Enforcement Training Division, Texas Standardized Field Sobriety Testing Program Instructor Manual (2002).
25. Marcelline Burns, *supra*.
26. K.J. Snapper *et al.*, *supra*.
27. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000) at VII-3.
28. Michael O'Keefe, *Drug Driving — Standardized Field Sobriety Tests: A Survey of Police Surgeons in Strathclyde*. 8 *J. Forensic Med.* 57, 60-61 (2001).
29. Marcelline Burns, *supra*.
30. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000). *Horizontal Gaze Nystagmus: The Science & The Law* (A Resource Guide for Judges, Prosecutors, and Law Enforcement), National Highway Transportation Safety Administration, <http://www.nts.dot.gov/peopole/injury/nystagmus/hgntxt.html>
31. *State v. Baue*, 258 Neb. 968 (2000), *U.S. v. Horn*, 185 F.Supp.2d 530, 561 (D.Md. 2002).
32. National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000) at VII-6.
33. Marcelline Burns & Herbert Moskowitz, *Psychophysiological Tests for DWI Arrest*, Final Report, DOT-HS-802-424 (1977) *supra*, at 54.
34. A. James McKnight *et al.*, *Development of a Standardized Boating Sobriety Test*, 31 *Accid. Anal. & Prev.* 147 (1999).
35. K.J. Snapper *et al.*, *supra*, at 4-2.
36. *Id.* 3-34 to 3-37.
37. Spurgeon Cole & Ronald H. Nowaczyk, *Field Sobriety Tests: Are They Designed for Failure?* 79 *Percep. & Motor Skills*, 99 (1994).
38. Joseph R. Meaney, *Horizontal Gaze Nystagmus: A Closer Look*, 36 *Jurimetrics J.* 383, 406 (1996).