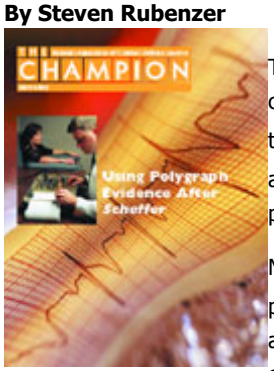
The Champion

May 2003 , Page 48

Search the Champion Looking for something specific?

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

By Steven Rubenzer



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 studies[1] 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 instructors[8] 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 test[14] 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 investigation[17] 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:

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.

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]

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]

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.

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 manual[26] 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-ophthalomologist) 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

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).
V. Tharp et al., supra. Theodore E. Anderson et al., Field Evaluation of a Behavioral Test Battery for DWI, DOT-HS-806-475 (1983).
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).
Marcelline Burns, First Annual DWI Training Seminar, Houston (2000).
V. Tharp et al., supra.
Jack Stuster & Marcelline Burns, Validation of the Standardized Field Sobriety Test Battery at BACs Below .10 Percent, DOT-HS-808-839 6 (1998).
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).
Marcelline Burns & Teresa Dioquino, A Florida Validation Study of the Standardized Field Sobriety Test (SFST) Battery, (1997).
Jack Stuster & Marcelline Burns, supra at 8.
Marcelline Burns & Ellen W. Anderson, supra. Marcelline Burns & Teresa Dioquino, supra. Jack Stuster & Marcelline Burns, supra.
National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000).
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).
See Louis M. Hsu, Diagnostic Validity Statistics and the MCMI-III, 14 Psych. Assess. 410, 410-411 (2002).
A. James McKnight et al, Development of a Standardized Boating Sobriety Test, 31 Accid. Anal. & Prev. 147 (1999).
A. James McKnight et al., Sobriety Tests for Low Alcohol Blood Concentrations, 34 Accid. Anal. & Prev. 305 (2002).
A. James McKnight et al., Development of a Standardized Boating Sobriety Test, 31 Accid. Anal. & Prev. 147, 152 (1999).
A. James McKnight et al., Sobriety Tests for Low Alcohol Blood Concentrations, 34 Accid. Anal. & Prev. 305 (2002)
Thomas D. Cook & Donald T. Campbell, Quasi-Experimentation: Design & Analysis Issues for Field Settings 2–9 (1979).
Id., Donald T. Campbell & Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research (1962).
See also Phillip E. Price & Spurgeon Cole, NHTSA Field Sobriety Tests: Validation and Invalidation, The Champion (Apr. 2001).
Marcelline Burns & Ellen W. Anderson, supra at 17.
See also J.L. Booker, End-Position Nystagmus as an Indictor of Ethanol Intoxication, 41 Sci. & Justice 113 (2001).
United States v. Horn, 185 F.Supp.2d 530, 558 (D.Md. 2002).
Marcelline Burns & Teresa Dioquino, supra, at 31.
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).
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.
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).
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).
V. Tharp et al., supra, at 30, 31.
Id. 30.
Id. 16.
Joseph Citron, MD, HGN: How to be Your Own Expert Witness (2002) (unpublished manuscript, http://www.ncdd.com/dsp_bookstore.cfm).
V. Tharp et al., supra at 7.
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).
J.L. Booker, supra.
Schultz v. State, 106 Md.App. 145, 664 A.2d 60 (1995).
Daubert v. Merrell Dow Pharma- ceuticals, 113 S.Ct. 2786 (1993).
United States v. Horn, 185 F.Supp.2d at 557.
New Hampshire v. Dahood, supra.
New Hampshire v. Dahood, (No. 99-510, December 20, 2002).
American Education Research Association, The American Psychological Association, and National Council on Measurement in Education, Standards for Education and Psychological Testing (1999).

1660 L St. NW • 12th Floor • Washington, DC 20036 • Phone: **(202) 872-8600** / Fax: **(202) 872-8690**