Law Hum Behav (2008) 32:293–313
DOI 10.1007/s10979-007-9111-y

ORIGINAL ARTICLE

The Standardized Field Sobriety Tests: A Review of Scientific and Legal Issues

Steven J. Rubenzer

Published online: 6 October 2007

Abstract This article details the history and development of the National Highway and Safety Administration's Standardized Field Sobriety Tests. They are reviewed in terms of relevant scientific, psychometric, and legal issues. It is concluded that the research that supports their use is limited, important confounding variables have not been thoroughly studied, reliability is mediocre, and that their developers and prosecution-oriented publications have oversold the tests. Further, case law since their development has severed the tests from their validation data, so that they are not admissible on the criterion for which they were validated (blood alcohol concentration), and admissible for a criterion for which they were not (mental, physical, or driving impairment). Directions for further research are presented.

Keywords Driving while intoxicated · Sobriety test · Horizontal Gaze Nystagmus · HGN · Walk and Turn · One Leg Stand · Driving safety

Field sobriety tests are the only behavioral tests that, if failed, are likely to lead to arrest. Conversely, they may have contributed to saving many lives by helping take drunk drivers off the streets. Developed by psychologists, the Standardized Field Sobriety Tests (SFSTs) have become one of the major types of evidence used in DWI cases. Failure on the tests will usually lead to arrest and a demand that the suspect take a roadside breath test to seek evidence that alcohol is present and the likely cause of any impairment observed. If the defendant refuses the breath test, his or her license will be suspended in many states and the SFSTs may become primary evidence of intoxication.

Most US jurisdictions now have two or more types of DWI statutes (National Highway and Traffic Safety Administration [NHTSA], 2004). The first requires evidence of mental, physical, or driving impairment due to alcohol use. Under this type of statute, observation of driving errors and the driver's demeanor, performance on simple motor tasks ("show me your license") and the SFSTs are important pieces of evidence. The second type of DWI statute provides that a breath or blood test indicating an alcohol concentration (BAC) over the legal limit creates a rebuttable presumption of intoxication. The third type is a *per se violation*, in which an alcohol concentrations above a specific value (.08% BAC in most jurisdictions) are illegal. No demonstration of impairment is necessary for a per se charge. A defendant may be charged under one or more of the various statutes (NHTSA, 2004). In both law enforcement and SFST research, BAC is typically estimated from breath alcohol concentration readings, and this should be kept in mind when reading this paper. Although such readings tend to correlate very highly with analyses from blood in laboratory settings (i.e., $r = .98$; Jones et al. 1992), there are many factors that can cause inaccuracies in practice (Hlastala 1998; Hlastala and Anderson 2007; Taylor and Oberman 2005).

The law has traditionally held that judgments of intoxication are within the powers of laypersons and that no special expertise is required. However, a substantial literature has found physicians, bartenders, and police officers to have very limited accuracy in making such judgments. Some studies have found police officers to perform better than other groups (Langenbucher and Nathan 1983; Pisoni and Martin 1989), whereas others (Pagano and Taylor 1979; Taubenslag and Taubenslag 1975; Vingilis 1983)

S. J. Rubenzer (✉)
11475 Sagecreek, Houston, TX 77089, USA
e-mail: srubenzer@comcast.net

found police performed poorly for BACs between .10 and .15%. Brick and Carpenter (2001) found officers had poor accuracy in the .08–.16% BAC range, but nonetheless had high confidence in their judgments. The difficulty in judging intoxication was one impetus for development of the SFSTs (Burns and Moskowitz 1977; NHTSA 2004).

Because of their importance in DWI prosecutions, the SFSTs have been the subject of intense, sometimes impassioned criticism since their introduction in the 1980s. In fact, there are far more published critiques by defense lawyers (Coffey 2004; Essen and Levenstein 1989; Nichols 1998; Pangman 1987; Price 1996; Rouleau 1990), legal scholars (Honts and Amato-Henderson 1995; Meaney 1996), professionals (Booker 2001, 2004; Citron n.d.; Cohen 2002; Cowan and Jaffee 1989, 1990; Glass 2003; Nowaczyk and Cole 1995; Rubenzer 2003a, b, 2006; Simpson 1988), and combinations of the above (Cole and Cole 1991; Price and Cole 2001) than empirical studies. This list is by no means exhaustive.

This article will review the development, validation, and legal and scientific status of NHTSA's three-test SFST battery. Coverage of case law is limited to the most prominent cases. I made extensive attempts to obtain every empirical article that evaluated the SFSTs, one of its components, or a close variant as a sobriety test. Nonetheless, a few foreign or unpublished works (Burns, in press; McKnight and Langston 1993; Pentillä et al. 1971; Pierce 1984) could not be obtained. Although Burns (in press) is cited by NHTSA (2004) as "Burns (2004)," a Department of Transportation spokesperson recently stated that it has not yet been released (J. Korkor, personal communication). I will focus on the SFST's relation to intoxication by alcohol, which has been the basis of most research. Recent studies have found the SFSTs somewhat effective at detecting marijuana (Leggett et al. 2007; Papafotiou et al. 2005) and benzodiazepines (Bramness et al. 2003; Kelly et al. 1997), but perhaps only high doses of amphetamines (McCloskey et al. 2007; Silber et al. 2005).

At the onset, and to avoid repetition, it is useful to note common deficits of nearly all SFST studies. In the great majority, there is scant reporting of subject variables; some do not even disclose age or gender and only a few report race. In most laboratory and analog field studies, subjects were relatively young and healthy. Examiners administering sobriety tests spoke with the subjects about their condition and observed general demeanor, and few studies utilized control groups. The reader can assume testing was not truly double blind unless explicitly stated. The reporting of testing procedures and statistical analyses is often very sparse, with many papers failing to report basic descriptive data such as mean BAC values or test scores or statistical tests of significance, and almost none report standard deviations. All SFST studies have utilized BAC (as estimated by breath alcohol content) as the criterion, using .10% in most early studies, and .08% or less in more recent investigations. No study to date has used mental, behavioral, or driving impairment as the SFST's criterion.

Statistics from various studies are cited throughout the article. The reader can assume that individual correlations and diagnostic statistics cited are statistically significant unless otherwise noted.

Review of the Empirical Literature

Early Developments and Research Through 1994

Several groups of medical researchers (Penner and Coldwell 1958; Pentillä et al. 1971 (cited in Burns 2003); Pentillä et al. 1974; Tenhu and Pentillä 1976; Widmark 1932/1981) have examined subjects' demeanor, vital signs, and performance on informal sobriety tests (walking a straight line, turning while walking, finger to finger, picking up small objects, the Romberg test) as indicators of BAC. While a number of tests correlated substantially with BAC, only one test did so at BAC levels below .10%: breakdown of "smooth pursuit" eye movements. Alcohol decreases the capacity to accurately track faster moving objects, leading the eyes to fall behind the target, then make a saccade to catch up. The two movements (drift off the visual target and saccade to re-fixate) together are referred to as *nystagmus* (Leigh and Zeigh 2006). Failure rates on most other sobriety tests did not approach 100% until BAC levels exceeded .32% BAC—four times the current legal limit in most jurisdictions. Pettilla et al. (1974) reported poor correlations between BAC and physicians' judgments of intoxication, especially for lower BACs, and about 20% of sober subjects (BAC $\leq$ .016%) were judged to be under the influence of alcohol.

The SFSTs were developed in the mid 1970s under contract from the Department of Transportation and NHTSA (Burns and Moskowitz 1977; NHTSA 2004). Burns and Moskowitz evaluated fifteen candidate sobriety tests in a laboratory study. A primary goal was to reliably detect lower levels of BAC, as previous studies showed that police officers frequently did not arrest drivers with BAC levels in the .10–.20% range. One of the tests evaluated, *Gaze Nystagmus*, had not been used in previous sobriety test evaluations but had been examined in laboratory studies (Aschan 1958; Lehti 1976). Subjects were seated with their chin stabilized (facing straight forward) and asked to fix their gaze on an object either 30 or 40 degrees to the side. The examiner then looked for inability to maintain gaze, with a drift off target (toward the midline) followed by a saccade to bring it back on target. One eye was tested at a time, the other occluded, and angles were mechanically measured. The other tests evaluated by Burns and Moskowitz were face valid tasks that have been used by police for decades.

Six tests were selected for further evaluation, with the provision that "it was also considered essential for the battery to represent a variety of skills ... (and) include measures of balance, large muscle coordination, cognitive skills, and oculomotor control" (p. 10). Each test had a behaviorally anchored scoring guide with a range of possible scores from one to ten. Two hundred and thirty eight subjects were recruited from the California State Employment Office and paid for their participation. The study report did not include a *subjects* section and provided little information about participants other than gender and drinking history. Each subject was classified regarding their drinking history (light, medium, heavy) and assigned to groups targeted for .00%, .05%, .10%, and .15% BAC, with only those classified as heavy drinkers assigned to the .15% BAC group. Subjects arrived at 8:00 in the morning and began drinking. Subjects assigned to the .00% BAC group were served orange juice with a small amount of vodka floated on the top to simulate the smell and initial taste of an alcoholic drink. Ten police officers and two observers were paired to evaluate each subject (one officer and observer per subject). They did not know which subjects were assigned to which BAC group, making the study ostensibly double blind, although some subjects were reported to volunteer their feelings of intoxication and likely showed other differences in behavior or demeanor. Both officers and observers scored the tests, estimated BAC, and made analog arrest decisions. All tests showed decrements in performance with increasing BAC levels. As in the Finnish studies, nystagmus was by far the best predictor, correlating .67 with BAC. At .10% officers achieved an overall accuracy of .76, a sensitivity (percentage of intoxicated subjects identified) of .84, and a false positive rate (percentage of sober subjects misclassified) of .27. In part because the base rate of subjects over .10% was low (27%), nearly 47% of arrest decisions were erroneous.

Stepwise discriminant analysis was used to reduce the six test battery to three tests that, with further refinement, were to become the SFST battery: Gaze Nystagmus, Walk and Turn, and One Leg Stand. These three tests were as accurate as the full six test battery when the discriminant analysis was used to classify subjects.

The authors also conducted a driving simulation task to ascertain whether SFST scores predicted driving errors. However, the simulation was highly abstract and cannot be described as a close facsimile of actual driving. SFST tests correlated with tracking errors (*r*s = .31–.44) but had lower correlations with reaction time and errors in signal detection. Walk and Turn and One Leg Stand generated much higher canonical weights (.52, .62) than Gaze Nystagmus (.04), indicating they contributed to the prediction of simulated driving errors despite being less effective indicators of BAC.

The second major US study (Tharp et al. 1981a) modified the standardized administration and scoring rules for the three tests selected in the 1977 study (see Table 1) and evaluated the tests' interrater and test–retest reliability. Gaze Nystagmus was replaced with a new eye test called Horizontal Gaze Nystagmus (HGN) consisting of three clues: (a) Breakdown of the smooth pursuit when visually tracking a slowly moving object, (b) distinct nystagmus when the eyes are maximally deviated to the side and held for four seconds (distinct nystagmus at maximum deviation), and (c) onset of nystagmus at less than 45 degrees of lateral gaze deviation. Each eye was scored separately, leading to a possible score of six clues. Angle of onset of nystagmus was no longer mechanically measured and subjects no longer covered one eye or used a chin rest. No rationale for these changes is provided, but practicality is the likely reason. The authors referred to pilot studies that examined the effect of variables that might influence performance on the SFSTs, but little information is provided about subjects, procedures, examiners, or results.

The main portion of the study was very similar in design to that of Burns and Moskowitz (1977). About half of the 297 subjects were invited back for re-testing one week later. Officers and observers scored the three SFSTs and also made summary judgments such as whether the person was too impaired to drive or would warrant arrest in an actual traffic stop. Each subject's BAC was also estimated. No cut-off scores were specified for any of the tests.

Table 1 Description of the NHTSA SFSTs and scoring criteria

SFST	Clues
HGN	(1) Lack of smooth pursuit, (2) distinct nystagmus at maximum deviation, (3) onset of nystagmus before 45 degrees. Scored for each eye, cutoff is four "clues"
Walk and Turn	(1) Inability to maintain balance in starting position, (2) starting too soon, (3) stepping off line, (4) not touching toe to heel, (5) raising arms, (6) improper turn, (7) stopping, (8) wrong number of steps. Any two "clues" is a failure
One Leg Stand	(1) Putting foot down, (2) hopping, (3) swaying, (4) raising arms. Any two "clues" is a failure

Table 2 Reliability coefficients for the SFSTs reported in Tharp et al. (1981a)

Test or judgment	Interrater Reliability	Test–retest reliability	
		Same rater	Different rater
HGN	.62	.66	.59
Walk and Turn	.74	.72	.34
One Leg Stand	.70	.61	.60

Note. Interrater reliability figures taken from Session 1 (n = 297). Figures from Session 2 were higher but based on half the number of subjects and may reflect practice effects with a given subject. Test–retest coefficients for 77 cases were based on the *same officer* condition and 64 cases for the *different officer* condition

Interrater and test–retest correlations were reported and are summarized in Table 2. The interrater reliability coefficient for HGN (r = .62) is far below that reported for Gaze Nystagmus (r = .90) in Burns and Moskowitz, possibly because of the increased complexity of the test and the fact that a chin rest and protractor were no longer used. The very low coefficient (r = .34) for Walk and Turn test–retest by different raters is corroborated by a similarly low figure (r = .39) when a research assistant (rather than a police officer) retested the same subject on the same occasion (not shown in table).

Officers obtained a correct classification rate of 81%, which was not significantly improved by use of a discriminant function. Sensitivity was .64 whereas the false positive rate was .12. Test–retest reliability was assessed by dosing subjects a week later to the same BAC level and retesting, both with the same and different examiners. Interrater reliability improved on the second testing session, and reliabilities were somewhat higher for test scores than for judgments regarding impairment or whether to arrest. Officers' estimates of actual BAC had an average error of .03% BAC, which yields a 95% confidence interval around .08% BAC of .007–.153%.[1] Clearly, officers were unable to estimate BAC levels with precision.

Tharp et al. (1981a) also attempted a field study to examine the performance of the SFSTs but experienced poor cooperation and deemed their data insufficient for statistical analysis. Despite this, the authors concluded, "This project has confirmed ... that gaze nystagmus is an outstanding useful tool for the officer at roadside" (p. 72).

Only two other studies were conducted before the SFSTs were introduced to law enforcement. Giguire (1983) reported an unusually high false positive rate (.77) for HGN and poor sensitivities for Walk and Turn and One Leg Stand (both .50). Anderson et al. (1983) utilized data from the 1981 NHTSA laboratory study to establish cut-off scores for each test and created a decision rubric for combinations of HGN and Walk and Turn scores. Using this template and the Tharp et al. data, the authors reported 80% correct classification of subjects of above or below .10% BAC. Respective figures for the individual tests were HGN, 77%; Walk and Turn, 68%; and One Leg Stand, 65%. No description of the analysis or principles that guided determination of the cut-off scores is presented. The authors then conducted a field study involving police from Virginia, Maryland, North Carolina, and Washington DC. Officers were trained in the SFSTs and asked to perform the tests before administering portable breath tests. Although results were encouraging, the authors noted that police officers were not randomly assigned to groups and no observers were present to confirm police adhered to instructions.

In another NHTSA-sponsored study, Compton (1985) found HGN was more accurate than standard "sniff and chat" assessments used at simulated sobriety checkpoints. Compton focused on sensitivity as the primary index of accuracy, citing an impressive figure of .95, but high false positive rates (.15–.75) were also observed, particularly for less experienced officers.

Good and Augsburger (1986), two optometrists, analyzed 2429 administrations of HGN conducted during traffic stops in Ohio. The authors stated that "over 92% of subjects scoring 4 points or higher on this test (HGN) registered BACs above .10%" (p. 471), but 92% of all subjects had a BAC of above the .10% BAC standard then used in Ohio. Thus, officers would be right 92% of the time by arresting everybody (which they did)—HGN added nothing to the overall accuracy rate. Goding and Dobie (1986) examined the relationship between BAC and an HGN-related clue, AON (angle of onset of nystagmus), in 46 emergency room patients and, separately, 159 persons suspected of DWI. In both portions of the study, a handheld cardboard protractor aided accurate assessment of AON. Major strengths of the study were the ability to draw blood and test for presence of drugs other than alcohol in the ER portion of the study, and that the tests were conducted late at night or early morning, as is typical of most DWI stops. In the ER portion of the study, 25 of 26 legally intoxicated persons had an AON of less than or equal to 40 degrees, for a sensitivity of .96. However, the BAC levels of subjects extended beyond .40%, which is an astoundingly high value and probably the reason for admission to the ER for many subjects. The mean BAC value was not reported. Four of 11 subjects with a BAC < .10% had AON at less than 40 degrees, although blood work revealed three of these had consumed drugs that exacerbate the CNS effects of alcohol. In the DWI stop study, two police officers

[1] A confidence interval is set using the SD rather than the average deviation. For a normal distribution, the SD equals 1.25332 times the average deviation (see Senter 1969).

independently administered and scored the AON (and lack of smooth pursuit). AON correlated .88 with a breathalyzer test, and BAC estimates were accurate within .02% for nearly 90% of suspects.

Boating while intoxicated is also a crime. Sussman et al. (1990) examined the use of several sobriety tests in a marine environment. HGN showed the highest correlation with BAC, was as valid when administered on a boat as on land, and provided incremental validity over other tests. McKnight et al. (1999) also found HGN to be the most accurate boating sobriety test, obtaining 100% sensitivity and 90% specificity at .10% BAC.

Several studies suggest that cut-off scores are set too low on the psychomotor SFSTs. Perrine et al. (1993) found that over 50% of drivers at .00% BAC failed Walk and Turn, and that correlations of Walk and Turn and One Leg Stand with BAC failed to reach significance. Cole and Nowaczyk (1994) reported that experienced police officers judged 46.5% of completely sober subjects videotaped performing psychomotor field sobriety tests to be impaired. O'Keefe (2001) reported that 60% of forensic physicians in Scotland with advanced credentials expressed concern that Walk and Turn and One Leg Stand were too difficult and the grading too stringent. On the other hand, Mundt et al. (1997) reported mean scores (# clues) for Walk and Turn and One Leg Stand of 0.3 and 0.5, respectively, in a group of 94 men (mean age 30.5) when sober, but no standard deviations or false positive rates are reported.

Kennedy et al. (1994) examined the ability of the SFST tests and a computer administered battery of cognitive tests to predict BAC levels. The cognitive tests included finger tapping speed, mental arithmetic, grammatical reasoning, pattern discrimination, code substitution, simple reaction time, and mental rotation tasks. Three studies examined prediction of BAC when alcohol levels were increasing, falling, and across both phrases. In all three studies, HGN correlated highest with BAC, averaging .77. One Leg Stand and Walk and Turn averaged .57 and .54, respectively, but two of the computerized cognitive tests (mental rotation, code substitution) showed higher correlations. The authors concluded that both SFST and cognitive test batteries contributed significantly to prediction of BAC and that Walk and Turn is redundant with One Leg Stand. Lastly, the authors presented inter-correlations of the various tests (typically .30–.60), providing some evidence that the SFSTs tap cognitive impairment associated with alcohol use. HGN's correlation was significantly higher with BAC than with any test, suggesting its relationship with performance impairment is indirect.

McKnight and associates (McKnight et al. 1995, 2002) found that only HGN was useful at a BAC criterion of .04%, used in some states for commercial drivers. A reduced cut-off score of two clues was used. Many variations of HGN were examined, including faster or diagonal administration of smooth pursuit, but correlations with BAC were not significantly affected. Effects on sensitivity and specificity were not addressed.

The NHTSA SFST Field Studies

The original NHTSA laboratory studies examined field sobriety tests as applied to volunteers in controlled, indoor conditions during daytime hours. Could officers obtain usable or valid results under traffic stop conditions, which might include inclement weather? Three large field studies conducted in the 1990s addressed this question (Burns and Anderson 1995; Burns and Dioquino 1997; Burns and Stuster 1998). Because initial field studies encountered poor cooperation among police officers, "Only agencies that could assume an extremely high level of cooperation and commitment would be recommended for participation" (Stuster and Burns 1998, p. 6). In each, actual traffic stops using the SFSTs were investigated. Officers had previous training and experience in the SFSTs and received refresher training before beginning data collection. In the first two studies, researchers obtained portable breath tests on the majority of drivers who were tested but released. This allowed an estimate of false negative rates—failures to make an arrest when warranted. Research observers were present in about half of these stops to assess accuracy of test administration for these studies. For all studies, the mean BAC of those arrested was over .138%, the base rate of legally intoxicated drivers exceeded 72%, and few people over 50 years of age were examined. Many stops were occasioned by observing driving errors, such as weaving, and occurred late at night in close proximity to bars. In the introduction to one of the studies (Burns and Stuster 1998), the authors stated, "A firm study requirement was that ... all officer-estimates of BAC must be based exclusively on the results of the NHTSA SFST battery" (p. 8). This assertion overlooks obvious situational cues and driver statements and demeanor.

Since the studies were quite similar in many respects and produced comparable results, they are discussed here together. Overall accuracy rates equaled or exceeded 86% in all three studies, although no data was reported for the .10% BAC criteria in Burns and Anderson (1995). Where SFST administrations were observed, very few errors of administration or scoring were reported. Varying weather and field conditions were reported not to impact SFST performance, but no data were presented in support of this claim. The authors reported that both prosecutors and officers regarded the SFSTs as practical and valuable, but again presented no data. False positive rates were significant in all studies and quite high for Walk and Turn (at least .72)[2] in Burns and Dioquino (Medley 2005), but were not openly reported. Diagnostic statistics for these and other SFST studies are presented in the Appendix. The reports for all three studies issued by NHTSA are lacking much of the material and analysis expected in a scientific paper, and none have been published in peer-reviewed journals. Nonetheless, Burns and Dioquino (1997) declared, "There appears to be little basis for continuing legal challenge (to the SFSTs)" (p. 31).

Recent Research

Mundt et al. (1997) examined the retest reliability of the SFSTs and of instrument-measured smooth pursuit and gaze fixation in a group of 94 male heavy drinkers. Test–retest reliability figures (median retest period 21 days) were presented for subjects when sober, when dosed to .08% (descending limb), and across conditions. The results were sobering. Baseline, sober performance was a strong predictor of dosed performance for the psychomotor tests, while alcohol-dosed test–retest correlations are very low (both .29). Figures were somewhat better for HGN, and considerably better for eye variables measured by instrument, a finding corroborated by Ettinger et al. (2003). Mundt et al. did not examine interrater reliability, a factor that may have contributed to the discouraging results. The fact that all subjects were alcohol tolerant drinkers may have done so as well. The authors noted that alcohol's effect varies with different settings and that behavioral measures, including eye movement variables, typically show only modest test–retest reliabilities.

Booker (2001) reported that 19% of subjects exhibited distinct nystagmus at maximum deviation in one or both eyes with no exposure to alcohol and 62% of subjects continued to do so immediately after all alcohol had cleared from the blood. Fifty-five percent of subjects showed this HGN clue in one or both eyes after sleep deprivation of 24.5 and 13–14 h of continuous mental and physical activity. Booker also examined the video tapes of 52 DWI arrests involving HGN. Using a very conservative scoring system, only one HGN administration was reported to be performed correctly—in great contrast to the NHTSA field studies. There is no description of how Booker chose the HGN videos, so non-random selection cannot be ruled out, and interrater scoring agreement was not examined.

Citek et al. (2003) evaluated the ability of 40 SFST Instructors and Drug Recognition Experts to use HGN to classify subjects above or below .08% and .10% BAC. Ninety six volunteer subjects consumed alcohol and underwent repeated testing in different physical positions. Testing position reportedly did not affect accuracy, but more HGN signs were observed at a given BAC level for supine positions and standing positions than in sitting subjects. Correlations of total HGN scores with BAC were consistent with previous findings (*r*s = .59–.63), but testing was not blind and it is not reported if scores from the two officers at each position were averaged, how raters' differences were resolved, or if they were able to confer with each other. Immediate retesting in the same position by different officers resulted in correlations of .59–.71. Interrater agreement for above/below the BAC criteria was 73.0–77.3% at .08% and 61.3–68.9% at .10% BAC. False positive rates were substantial, ranging from .29 to .56, with higher figures being observed for the .10% BAC criterion. Sensitivities ranged from .80 to .97 (higher for .10% BAC), while Cohen's *d*s ranged from 1.3 to 1.7. Lastly, the authors reported observations for a substantial number (n = 164) of subjects at .00% BAC. They reported, somewhat ambiguously, that "fewer than 10% at any posture demonstrated (at most) three HGN signs" (p. 703). Only one subject obtained a failing score, showing four clues in the standing position, but not others.

Citek et al. seem to rationalize data not favorable to the use of HGN in DWI investigation. The modest interrater agreement rates are cited to deem HGN "highly reliable" (pp. 706, 708), while false positive rates as high as .56 "are acceptable, given the fact that the result of the HGN test provides only one of many possible pieces of evidence of impairment" (p. 705).

Hlastala et al. (2005) published evidence that accuracy rates on NHTSA field study depended on more than SFST results. They found that the SFST *scores* accounted for 54% of the variance in BAC (estimated by evidentiary breath tests). However, adding the officers' subjective estimate of BAC, based on all his/her observations, increased the explained variance to 75%. In the absence of such observations, the standard error of predicted BAC values was .044% BAC, resulting in a 95% confidence interval of ±.09% around the predicted BAC value. Sensitivity was excellent (.95–.97), but accuracy varied greatly with BAC: Accuracy rates for those below .04% BAC and above .09% BAC exceeded 80%, but fell to a low of about 37% at .07% BAC. False positive rates reached .56 for those in the .06% to .10% range and .64 for the .07% to .09% range.

Entry into the Legal System

In 1984, NHTSA released the forerunner to the Student Practitioner Workbook (NHTSA 2004), the companion to the basic police training course for DWI detection. The

[2] A full explanation of this figure and analysis is available from the author.

SFSTs were given special attention and the importance of standardized administration stressed. However, some important details are not covered: Although cut-off scores are provided for individual tests, there is little guidance about decision rules for passing or failing the SFST battery. How is the officer supposed to relate to the defendant—firmly, politely, or in an authoritarian manner? *Improved Sobriety Testing* (NHTSA, 1984), advised that people who are over 60 years old, 50 pounds overweight, or physically impaired should not take the Walk and Turn or One Leg Stand tests. It noted that both tests required a level, hard, dry, non-slippery surface, and that the Walk and Turn required a visible line. If such conditions could not be met at roadside or if the suspect had a medical condition that might impair performance, the officer was advised to rely solely on HGN.

By the mid-1980s, it was apparent that in order to introduce HGN in court, the prosecution might have to provide a scientific foundation through expert testimony. Tiffany (1986), a prosecutor, introduced readers of the *Journal of Optometry and the Law* to the NHTSA research, the need for expert testimony, and its desired form and content. Two more balanced articles appeared in the following year, both by optometrists. Halperin and Yolton (1986) noted several potential problems with HGN:

> A small but significant number of suspects have chronic nystamoid eye movements. Second, and not mentioned in the Training Manual, suspects who have high refractive errors could have trouble seeing the test target with their glasses removed and may therefore have problems with the test. Third, as most optometrists know, many subjects will have jerky eye movements even with .00% BAC. (p. 657)

Forkiotis (1987) conceded optometrists have little training or experience in eye testing for determining sobriety. All three articles rely almost entirely on NHTSA studies. Nonetheless, in 1993, the American Optometry Association passed a resolution declaring HGN "to be a scientifically valid and reliable tool for trained police officers to use in field sobriety testing" (p. 653) and encouraging members to become involved as expert witnesses.

Defense attorneys responded in force to the new challenge of HGN. In one of the first critiques by defense attorneys, Pangman (1987) cited a number causes of nystagmus other than alcohol such as poor lighting conditions, distractions from traffic and wind, and medical conditions. He noted that a police officer's scoring cannot be verified, the difficulty of getting a reliable reading, that established sources have reported that HGN clues can appear at .06% BAC (the legal limit was then .10%), and that cautionary information from NHTSA studies is omitted from law enforcement materials. He claimed that all the NHTSA studies used protractors and chin rests to measure angle of onset of nystagmus, but this assertion (and several others) is not accurate: Only the original laboratory study did so. Both Pangman and Rouleau (1990) suggested that an authoritative medical source (Toglia 1976) declared normal, transient end-point nystagmus "indistinguishable" from the NHTSA's distinct nystagmus at maximum deviation clue. Aside from the implausibility of this statement, the author could locate no material in Toglia on this topic.

Professionals from outside the law also weighed in. Essen and Levinstein (1989) declared that the psychomotor tests are clearly scientific tests and that the prosecution should be required to prove the reliability and validity of the tests, that the test administrator is properly qualified, and that the tests were administered properly. A criminologist (Busloff 1993) argued the "extremely subjective" (p. 235) nature of HGN contrasted with its scientific status and supposed aura of infallibility. She recommended HGN be held to emerging evidentiary standards and that juries be expressly cautioned about the limits of the test and the officer's expert testimony.

Three detailed and sophisticated critiques of the SFSTs appeared in the mid-1990s, one by a pair of psychologists and two by legal scholars. Nowaczyk and Cole (1995) cited high false positive rates and noted the SFSTs have not been validated in the field or been demonstrated to correlate with driving impairment. Citing the claimed 80% accuracy rate, they noted that many subjects had low (<.05%) or high (≥.15%) BACs and thus, presented little challenge to the tests. They noted that the onset of nystagmus at less than 45 degrees of deviation is a liberal application of previous research and invites false positives. Lastly, Nowaczyk and Cole argued the SFSTs are not generally accepted in psychology given the absence of norms and lack of peer review.

In a law review article, Honts and Amato-Henderson (1995) critiqued the evidence presented in *State v. Superior Court* (1986; see below), concluding that much of it did not withstand close scrutiny. Among the problems cited were: (a) the limited reporting of the pilot study in Tharp et al. (1981a) which purportedly ruled out various alternative explanations for observed nystagmus, (b) insufficient exploration of fatigue as a possible contributor to poor SFST performance, and (c) selective reporting of study results. The authors concluded that the scientific foundation of HGN "is at best weak" (p. 694) and principle findings have not been replicated by independent researchers. In another excellent review, Meaney (1996) suggested that jurors may perceive HGN as more objective and precise than it is, with the effect that its probative value may be outweighed by unfair prejudicial impact.

With the SFSTs and their supporting literature increasingly buffeted by criticism, the American Prosecutors

Research Institute (APRI 1999) published an informational guide for prosecutors and judges extolling the virtues of the SFSTs, especially HGN. A substantial portion of the APRI publication addresses the use of the SFSTs in court. Readers are told

> When the HGN test is admitted as a physical observation, the law enforcement officer can establish (its) reliability. The officer would explain that, based on the officer's training and experience in the interpretation and administration of the HGN test to impaired subjects, the officer can accurately identify that a subject is too impaired when he or she performs unsatisfactorily on the HGN test. (p. 7)

Further, readers are told "the officer should take the opportunity to communicate evidence of the HGN test's reliability. Otherwise, the significance of the HGN test as the most reliable of SFST of [*sic*] alcohol impairment will be lost" (p. 7). In sum, this publication appears to confuse the concepts of reliability and accuracy, misstates the evidence of SFST validity, and encourages police officers to testify on scientific issues beyond their training and competence.

Burns (2003) published the first overview of the SFSTs in a peer-reviewed psychology journal. The article described the Finnish and US laboratory and field studies, but did not mention any of the numerous critiques, adverse court opinions, or non-NHTSA SFST studies. Burns also made it clear, if not explicit, that the tests and officers that interpret them are the functional unit of analysis—not the tests themselves. In discussing the SFST test–retest reliability coefficients, which hover around .70, Burns stated that such figures "meet the reliability standard for psychomotor tests (Guilford and Fruchter 1978)" (p. 1192). However, Guilford and Fruchter merely noted modest retest reliabilities for several psychomotor tests (p. 416); they did not advocate a lower standard on this page or elsewhere.

Also in 2003, the American Prosecutors Research Institute issued a compendium of articles that supported the use of HGN. Bobo (2003) stated that habitual drunk drivers often do not show visible signs of intoxication, may practice the Walk and Turn and One Leg Stand, but are unable to defeat HGN in this manner. Citek (2003) noted there are many different types of nystagmus and abnormal eye movements but asserted, "A properly trained police officer will know how to distinguish such eye movements" (p. 20). The source of such optimism is unclear, as NHTSA's training materials contain virtually no material on this subject (NHTSA 2004, 2002) and there is no published evidence on the issue.

Booker (2004) took note of inaccuracies in the SFST literature and charged that "the United States Department of Transportation indulged in deliberate fraud in order to mislead the law enforcement and legal communities into believing the test (HGN) was scientifically meritorious and overvaluing its worth in the context of criminal evidence" (p. 134). Booker proposed three criteria to warrant the serious charge of fraud: (a) lack of competence in research, (b) efforts to hide evidence of flawed research, and (c) deliberate deception and misrepresentation. Five of Langmuir's (1969) six descriptors of "pathological science" ([a] fantastic claims contrary to experience; [b] claims of great accuracy; [c] maximum effect produced by minimally detectable causative agent; [d] magnitude of effect close to limit of detectability; [e] post hoc explanations for failures) were said to "demonstrably apply." Among many other charges, Booker cited omission of unsupportive studies and critiques from materials used to train police officers and prosecutors. As proof of deliberate fraud, he noted the American Prosecutors Research Institute's (1999) citation of three studies as supportive of HGN, when they have nothing to do with sobriety testing.

In a non-forensic chapter for medical readers, Dell'Osso and Darloff (2005) identified 49 types of nystagmus. They criticized the use of nystagmus in sobriety testing, questioning whether a "cursorily trained" police officer could determine if nystagmus is pathologic, noting "such judgments are difficult for experts to make under the best of circumstances" (p. 26).

Case Law and Legal Issues

In the first appellate decision to address admissibility of HGN (*People v. Loomis* 1984), HGN was found inadmissible because the officer's testimony estimating BAC constituted expert medical opinion and HGN was judged not to have gained general acceptance in its field—but no particular field was identified. In *People v. Vega* (1986), the Appellate Court of Illinois, Fourth District, ruled that HGN was evidence beyond the ken of the average individual and required proper foundation, by way of expert testimony, for its introduction. In contrast, the other field sobriety tests were considered non-scientific and deemed to require no specialized knowledge to interpret. In *State v. Nagel* (1986), an Ohio appeals court held that HGN "requires only personal observation of the officer administering it. It is objective in nature and does not require expert interpretation. Objective manifestations of insobriety, personally observed by the officer are always relevant" (p. 286).

Many state supreme courts have addressed issues regarding the SFSTs, particularly HGN. In the first such case, the Arizona Supreme Court (*State v. Superior Court* 1986) ruled that HGN is a scientific test, that it is reliable enough to establish probable cause to arrest, and that behavioral and experimental psychologists, as well as highway safety professionals, are appropriate professionals groups for gauging if HGN has attained general acceptance.

In *Iowa v. Murphy* (1990), the court stated that HGN's ease of administration weighed against regarding it as a scientific test and that a police officer, properly trained in its administration, is qualified to testify about its results. In *State v. Witte* (1992), the Kansas Supreme Court deemed HGN to be scientific evidence, in part because nystagmus is not a commonly recognized sign of intoxication. It also held a police officer could not attribute the presence of nystagmus to alcohol, as there are numerous other judicially-recognized causes[3] and the police officer lacked the knowledge to differentially diagnose the various possibilities.

State supreme court decisions in Texas (*Emerson v. State* 1994) and Maryland (*Schultz v. State* 1995) greatly facilitated admission of HGN evidence in those jurisdictions by taking judicial notice of HGN's purported reliability. As stated in *Emerson v. State*, a court can take judicial notice of any scientific fact that "is capable of accurate and ready verification by resort to sources whose accuracy cannot reasonably be questioned" (p. 764). Referring to previous laboratory studies and the NHTSA research, the court concluded "the effect of alcohol on nystagmus, specifically HGN, is well documented" (p. 766).

In contrast, several states turned a much more skeptical eye to the HGN literature in the mid 1990's. In *People v. Leahy* (1994), the California Supreme Court rejected the premise that a scientific technique should become immune from *Frye* scrutiny "merely by reason of long-standing and persistent use by law enforcement *outside* the laboratory or courtroom" (p. 332), and noted that prior courts had not explained "how police officers are competent to establish general acceptance of HGN testing *in the scientific community*, or how they are qualified to relate the scientific bases underlying the nystagmus test" (p. 334). In *State v. Meador* (1996), testimony by a defense psychologist led Florida judges to cite the low reliability figures for the SFSTs and the absence of evidence that they relate to driving impairment. In *State v. Homan* (2000), the Ohio Supreme Court ruled that an improperly administered SFST was inherently unreliable and could not be introduced even to establish probable cause.

The first federal case to address the SFST's *Daubert* standing (*U.S. v Horn* 2002) held several days of testimony from psychologists, including two involved in test development. The experts detailed numerous shortcomings of the SFSTs and their supporting research. The court explicitly criticized previous decisions, such as *Emerson*, that took judicial notice of HGN based on previous judicial decisions and accepted unsubstantiated claims of general acceptance among psychologists. The court found fault with previous decisions citing acceptance among criminologists, law enforcement, highway safety experts, and prosecutors, questioning whether they possessed the requisite scientific expertise. The judge questioned the claimed accuracy rates for the SFSTs and found no evidence to support the cut-off scores specified for Walk and Turn and One Leg Stand.

Although psychologists have been the primary experts called on the SFSTs, testimony about HGN was taken from an ophthalmologist and neuro-ophthalmologist in *State v. Dahood* (2002a). Both medical experts objected that distinct nystagmus at maximum deviation of the eyes is common and should be eliminated from the test. They also opined that the 2 s allowed for each smooth pursuit pass is too fast, and may result in saccades that could be mistaken for nystagmus. Further, they stated that raising the stimulus above eye level (as NHTSA recommends) involves eye muscles other than those required for lateral eye movement and invalidates the test. The judge gave great weight to these opinions and ruled that HGN did not meet any of the *Daubert* criteria. However, the state supreme court (*State v. Dahood* 2002b) overturned the trial court's ruling. The court took judicial notice of four principles of HGN deemed to have received general acceptance in the relevant (but unidentified) scientific communities:

> (1) HGN occurs in conjunction with alcohol consumption; (2) its onset and distinctness are correlated to BAC; (3) BAC in excess of .10 percent can be estimated with reasonable accuracy from the combination of the eyes' tracking ability, the angle of onset of nystagmus and the degree of nystagmus at maximum deviation; and (4) officers can be trained to observe these phenomena sufficiently to estimate accurately whether BAC is above or below .10 percent. (p. 168)

Discussion

A significant body of research has reported substantial relations between SFST results, particularly for HGN, and estimated BAC. Almost all studies produced statistically significant and sizable effect sizes (see Appendix), but

[3] (a) Problems with the inner ear labyrinth; (b) irrigating the ears with warm or cold water under peculiar weather conditions; (c) influenza; (d) streptococcus infection; (e) vertigo; (f) measles; (g) syphilis; (h) arteriosclerosis; (i) muscular dystrophy; (j) multiple sclerosis; (k) Korsakoff's syndrome; (l) brain hemorrhage; (m) epilepsy; (n) hypertension; (o) motion sickness; (p) sunstroke; (q) eyestrain; (r) eye muscle fatigue; (s) glaucoma; (t) changes in atmospheric pressure; (u) consumption of excessive amounts of caffeine; (v) excessive exposure to nicotine; (w) aspirin; (x) circadian rhythms; (y) acute trauma to the head; (z) chronic trauma to the head; (aa) some prescription drugs, tranquilizers, pain medications, anti-convulsants; (ab) barbiturates; (ac) disorders of the vestibular apparatus and brain stem; (ad) cerebellum dysfunction; (ae) heredity; (af) diet; (33) toxins; (ag) exposure to solvents, PCBs, dry-cleaning fumes, carbon monoxide; (ah) extreme chilling; (ai) lesions; (aj) continuous movement of the visual field past the eyes; and (ak) antihistamine use (*Schultz v. State*).

because of methodological limitations, the empirical support for the SFSTs must be considered circumstantial. In addition to the limitations of most studies noted earlier in the paper, several additional factors must be identified. Laboratory studies were conducted in daytime, volunteer settings and excluded people that might be prone to failing the tests (fatigued, anxious, aged, medically or psychiatrically impaired). The reporting is very incomplete on some important issues, such as potentially confounding variables that were investigated in small, minimally described pilot studies. In the NHTSA field studies, subjects were "selected" because they displayed driving errors, so sensitivity figures in these studies are likely overestimates: Alcohol tolerant drivers with BACs over the limit who did not show driving errors would not have been stopped, and these drivers may well have performed better on the SFSTs than the drivers who were stopped. The effects of climate and other possibly confounding factors (fatigue, anxiety) are dismissed with reassuring statements that imply statistical analyses were done, but were not formally reported (see also Rubenzer 2003a, 2006).

Quantitative Summary of Findings

Because of the problems noted above, the existing studies results cannot be taken at face value. A formal meta analysis, which requires considerable investment of resources (Lipsey and Wilson 2001), does not seem justified at this time given the lack of true double-blind testing in any study. Nonetheless, it may be useful to consider some summary statistics, recognizing that such numbers are likely overly optimistic for the tests used as stand-alone measures.

One accuracy index used below may not be familiar. The *Likelihood ratio* (*LR*) is an index of the diagnostic efficiency at a given cut-off score. It is created by dividing the sensitivity by the false positive rate, thus creating a ratio indicating how much more frequently people with the condition fail than those without it. A test that does not discriminate at all between groups will have a *LR* of 1.00, whereas the *LR* would approach infinity for a perfectly discriminating test. *LR* is independent of the base rate and more intuitive than similar indices, such as the odds ratio (Simon n.d.; Zhou et al. 2002). One can calculate *LR* even when minimal data are reported, which is often the case in SFST studies, and confidence intervals can be calculated to show if the *LR* differs significantly from unity. Although *LR* is not affected by base rate, it will always be easier to discriminate between widely varying BAC levels than those that vary minimally.

With the caveats given above, several patterns are present in the SFST studies to date (see Table 3)[4]. The data are remarkably limited for the SFSTs as a battery. At .10%, two of four available studies included several other sobriety tests that may have influenced accuracy rates. Only two studies are available for .08%. Sensitivities were excellent at .08% (.96, .98), lower at .10% (mean = .74), perhaps due to the limited experience of officers in early studies. For all BAC criteria, mean false positive rates are .20–.24, whereas mean *LR*s range from 3.8 to 4.4. The limited data available suggest the SFSTs are useful for lower BAC levels if the HGN criterion is adjusted. Lastly, it must be remembered that SFST *scores* did not determine judgments of intoxication—police officers using the SFSTs, and other observations, did. Thus, these studies arguably validated the judgment of the police officers in the study rather than the SFST tests.

The pattern of data for HGN is similar to that for the SFSTs, but based on substantially more studies. HGN has almost always fared better than other sobriety tests in head to head comparisons and always showed substantial to large correlations with BAC, averaging .65 across nine studies (range .51–.77). One study (Sussman et al. 1990) reported incremental validity over other observations and tests. The sensitivity of HGN appears quite good, with only one study reporting a figure below .72, even when used for a BAC criterion as low as .04% (a cut-off score of two clues was used). For .08% BAC, an average sensitivity of .88 (range .79–.98) was observed and an average *LR* of 3.6 (range 2.3–6.6) based on six studies. The average false positive rate is approximately .28 (range .13–.37). At .10% BAC the average *LR* is similar though the average false positive rate is considerably higher (.41), based on seven studies. The data at lower BAC limits are very limited but promising, and HGN may be the only sobriety test with any usefulness at .04%. The HGN clue of angle of onset of nystagmus retains sensitivity with very high BAC drinkers who have probably developed tolerance (Goding and Dobie 1986; Lehti 1976) and who often do not show typical, observable indications of intoxication (Sullivan 1987; Urso et al. 1981). Such people nonetheless showed impaired driving at .10% BAC in the only study to examine this issue (Laurell et al. 1990). The effect of tolerance on smooth pursuit performance has not been investigated. Variations in the speed or angle for the smooth pursuit phase do not appear to affect correlation with BAC (McKnight et al. 2002), but the affect on diagnostic statistics has not been examined. The position of the subject appears to have little effect (Citek et al. 2003; Compton 1985), and administration of HGN on a boat provided positive results in two studies (McKnight et al. 1999; Sussman et al. 1990). While lack of blind testing and other subject selection issue might lead to overestimation of diagnostic values, limited training of examiners and range of BAC

[4] Average correlations cited for the SFSTs do not include data from Foss et al. (1990) or Perrine et al. (1993) as they used Somers' *d* rather *r*.

Table 3 Accuracy indices for the SFSTs across studies

Test	Crit.	#	Sens.		FPR		*LR*	
			Mean	Range	Mean	Range	Mean	Range
SFSTs	.04, .05	2	.84	.78–.89	.20	.17–.24	4.2	3.7–4.7
	.08	2	.97	.96–.98	.23	.18–.29	4.4	3.4–5.3
	.10[a]	4	.73	.64–.84	.24	.12–.42	3.8	1.6–5.3
HGN	.04, .05	2	.76	.72–.79	.23	.08–.38	5.5	2.1–9.0
	.08	6	.88	.75–1.00	.28	.13–.37	3.6	2.3–6.6
	.10	7	.87	.50–1.00	.41	.10–.82	3.4	1.2–9.8
WAT	.08	2	.68	.44–.92	.47+[b]	.22–.72+[b]	1.9	1.7–2.0
	.10	4	.59	.50–.73	.27	.15–.49	2.5	1.5–3.3
OLS	.08	2	.69	.46–.92	.25	.09–.41	3.7	2.2–5.1
	.10	5	.59	.50–.67	.17	.08–.27	4.3	2.1–6.5

Note. Numbers in this table reflect behavioral observations beyond test scores. FPR = false positive rate. Only HGN in the standing position data are utilized from Citek et al. (2003). Data for SFSTs at .10% includes data from Burns and Moskowitz (1977), which used Gaze Nystagmus rather than HGN. Data for Cole and Nowaczyk is not reflected in this table because they did not address either individual tests or the SFSTs as a group

[a] Two of the four studies included several sobriety tests beyond the SFST battery, and one utilized Gaze Nystagmus instead of HGN

[b] Includes data from Burns and Dioquino (1997) beyond the number of studies indicated. Because the authors did not report number of subjects who refused the given tests, the figures indicated are lower bound estimates; the actual values may be higher

values in some studies probably lead to conservative estimates. Lastly, HGN shows substantial relations to BAC despite significant problems with scoring reliability. If reliability could be improved through better training or use of roadside implements, accuracy and validity might be further improved. For a discussion of vision science issues pertaining to HGN, see Rubenzer and Stevenson (2007).

Walk and Turn clearly performed less well than HGN, with lower discriminative power and higher false positive rates, particularly at .08%. Only two, unpublished studies (McKnight and Langston 1993, cited in McKnight et al. 1995; Stuster and Burns 1998) reported diagnostic statistics at .08% BAC, yielding a modest mean *LR* of 1.9 and false positive rate of .37. Data from an additional paper (Burns and Dioquino 1997) provides a false positive rate of at least .72. At .10%, four studies provide diagnostic statistics, and for false positive rate and *LR*, the figures are somewhat better: Average sensitivity, false positive rate, and *LR* are .59, .27, and 2.5, respectively. Sensitivity is lower than at .08%, although this (and other differences) is likely unreliable given that the .08% figures are based on only two studies. Lastly and perhaps most problematic, false positive rates near .50 have been observed in subjects at .00% BAC (Perrine et al. 1993). Walk and Turn does correlate substantially with BAC (mean $r = .55$, range .37–.61, four studies), but it appears that its cut-off score is set too low, particularly for older, heavier, and physically inactive or impaired subjects.

One Leg Stand shows a lower average correlation with BAC (r = .45, range = .16–.60, six studies) than Walk and Turn, but surprisingly strong diagnostic discriminative power at .08% (*LR* = 3.7) and .10% (*LR* = 4.3). Like Walk and Turn, there are extremely limited data available at .08%. Based on two studies, average sensitivity was .69 and the false positive rate .25. Data from Burns and Dioquino (1997) suggest a substantial false positive rate but incomplete reporting prevents a precise analysis. At .10% five studies are available and yield respectable diagnostic statistics: sensitivity = .59; false positive rate = .16; and *LR* = 4.5.

Diagnostic statistics such as sensitivity and false positive rate are potentially very informative, but when studies are not conducted blind or are otherwise flawed, their value is much reduced. In such circumstances, various measures of reliability may be a better gauge of a test's functioning and potential. Mundt et al. (1997) reported very low test–retest reliabilities for both psychomotor tests, although the design limitations discussed above made high test–retest reliabilities unlikely. However, similar figures were reported for Walk and Turn when scored by different raters in Tharp et al. (1981a). Interrater differences appear to be a substantial source of variance for all the SFSTs, especially HGN. Tharp et al. reported an interrater reliability for HGN of .62 and Citek et al. (2003) reported figures from .59 to .71 when different officers immediately retested the subject. The various reliability figures reported range from low to clearly unacceptable for tests that provide the basis for arrest and, often, evidence of impairment in legal proceedings. Heilbrum (1992) recommended tests used in forensic contexts have a minimum reliability of .80, while Nunnally and Bernstein (1994, p. 265) cited a "bare minimum" figure of .90. The modest reliability figures give further reason to believe the validity statistics cited above are inflated, since tests that cannot be scored reliably and that yield different results on different occasions are inherently of limited validity (Nunnally and Bernstein 1994).

Conceptual and Legal Issues

Since their inception, the SFSTs have been tied to BAC level as the criterion. To establish probable cause to administer a breath test, this is the logical standard. But for establishing evidence of impairment, the choice of BAC has created several hurdles. In the absence of direct evidence that SFST performance relates to driving impairment, one must look at the SFSTs' ability to estimate BAC and then the ability of BAC to estimate driving impairment. The first link is especially weak, as no study has found the SFSTs (as typically practiced) to predict BAC with better than an average error of estimate of .03%, and most figures have been much larger (Hlastala et al. 2005; Tharp et al. 1981a).

There is only one published study (Kennedy et al. 1994) that reported correlations of SFST scores with measures of cognitive performance. Although the correlations were moderately large, they were smaller than that of HGN with BAC, and the generalization to driving is unclear. Ironically, the original 1977 laboratory study could have examined the relationship of the SFSTs to a variety of behavioral and cognitive tasks, but the authors did not report such analyses. Walk and Turn and One Leg Stand are routinely referred to as "divided attention tests" in NHTSA publications, although empirical analyses have never been presented to justify this label. The obvious importance of gross motor coordination, balance, and short-term memory is rarely mentioned. No studies have examined the SFSTs as measures of physical/behavioral impairment.

This connection is intuitive but may well be wrong. Many people with poor balance or coordination may be perfectly fit to drive—even if the poor balance is due in part to use of alcohol. There is also the issue of calibration: How much impairment on the psychomotor SFSTs translates into substantial impairment in driving ability? The *Meador* and *Horn* courts held this to be a simple matter of common sense, and this may be true for obviously intoxicated suspects. However, it is not at all obvious that raising one's hands 6 in. (15.2 cm) while performing Walk and Turn, or failing to touch heal to toe on one of 18 steps, or even losing one's balance relates to impairment while driving. Informal scoring by jurors who take account of age, weight, and other relevant factors may well prove more accurate than the NHTSA scoring scheme that does not. The need for more rigorous evaluation of both formal and common sense "scoring" is pressing.

The SFSTs have been the subject of several legal debates. Is HGN scientific, technical, or lay evidence? Are the other SFSTs scientific? Is a police officer a technical expert? Can he/she testify as to the correlation between HGN and BAC? In considering HGN to be scientific, a number of dubious arguments have been advanced: that HGN is scientific because (a) it has a scientific name; (b) has a detailed, complex sounding description; (c) is used by physicians; and (d) was developed and researched in medical laboratories. Those arguing that HGN is not scientific claim the test entails observations of an objective indication of intoxication—namely, jerking of the eyes, and requires no special equipment. This argument assumes HGN is easily and objectively scored, but the interrater reliability data reported thus far indicate otherwise: Even SFST instructors are unable to agree adequately on HGN scoring (Citek et al. 2003).

Unlike HGN, the psychomotor SFSTs have been almost universally regarded as lay observations, despite their standardized administration, scoring, and interpretation. It is unclear if reasoning cited in *Meador* shows ignorance of behavioral science methodology, unstated concern over inadequate documentation of cut-off scores, validity, and accuracy (as in *Horn*), or the desire to preserve the doctrine of lay testimony about intoxication. Numerous courts have virtually equated poor performance on the Walk and Turn and One Leg Stand with impairment to drive. There is no doubt that tests like the Walk and Turn and One Leg Stand make excellent courtroom exhibits and are closely associated in the minds of many people, including appellant judges, with impaired functioning. However, given evidence of their prejudicial quality (Cole and Nowaczyk 1995; Perrine et al. 1993), this may be a liability, not an asset.

There are two types of recognized scientific testimony on the SFSTs, as articulated in cases such as *People v. Williams* (1992): that regarding the reliability, validity, and accuracy rates of the SFSTs, and that regarding the physiological mechanisms that underlie their performance. There is no doubt that age and various medical conditions and medications can interfere with SFST performance, and several "psychological" factors also loom very large: test anxiety, fatigue, and psychiatric illness. Anxiety may interfere with performance on psychomotor tasks (Mullen et al. 2005; Noteboom 2001; Pijpers et al. 2005; but see also Hogan 2003) and the neural mechanisms of balance (Balaban 2002; Bolmont 2005). Sleep deprivation may lead to a variety of mood, motor, and cognitive deficits, even when the deficit is 5 h of sleep in the preceding 24 h (Pilcher and Huffcutt 1996). Fatigue or time of day has worsened nystagmus in several studies (Bahill, Iandolo, and Troost 1980; Booker 2001), although sometimes only in association with alcohol (Tharp et al. 1981a, b). Nystagmus has been demonstrated in schizophrenic patients and their first degree relatives (Hartje et al. 1978; Kathmann et al. 2003), manics (Iacono et al. 1982), "neurotic outpatients" (Hartje et al.), sober alcoholics (Kobatake et al. 1983), and adults with Attention Deficit Disorder (Munez et al. 2003; but see Ross et al. 2000).

Daubert Issues

Falsifiability

The theory that alcohol affects SFST performance is clearly subject to falsification if BAC is the primary criterion, and there are numerous studies that correlate SFST performance and BAC level. The proposition that SFSTs are related to driving impairment is also falsifiable but more difficult to test. Whereas impairment on a closed driving course might readily be correlated with SFST performance, some significant performance deficits occur only in response to rare events or in interaction with other vehicles or drivers (e.g., road rage). The theory that SFST performance is related to driving performance is falsifiable, but as yet untested.

Error Rate

In *Daubert vs. Merrell Dow Pharmaceuticals* (1993), the US Supreme Court ruled that courts should consider the error rate of a theory or technique in determining its admissibility. However, the court gave no guidance on *which* error rate to consider or what value would be objectionable (Faigman and Monahan 2005; Foster and Huber 1999). Thus far, no court considering SFST accuracy rates has distinguished among types of error rates. Using a four fold classification table, there are five possibilities: Three of these (overall classification accuracy, percent false positive, percent false negative) are greatly affected by the base rate, whereas the false positive rate (1 − specificity) and false negative rate (1 − sensitivity) are not. Some courts have recognized that high base rates in the NHTSA field studies provide a baseline for improvement over chance. However, there has been no distinction made between related terms such as the false positive rate and percent false positive rate (percentage of "positives" on the test that are wrong). The distinction is important because base rates vary enormously among DWI enforcement situations and the percent false positive rate is linearly dependant on the base rate: If the base rate doubles, so will the percent false positive rate. The NHTSA field studies had base rates from 72–80%. In contrast, when Perrine, Peck, and Fell (1988) assessed base rates at sobriety checkpoints at high risk times (past midnight, Thursday through Saturday), they found only 4.6% of drivers were above .10% BAC, with just over 14% above .05%. Thus, base rates and accuracy indices based on them appear to vary perhaps tenfold across DWI enforcement settings.

Instead of reporting more common indices such as sensitivity and specificity, many SFST studies report the accuracy of "arrest decisions" (percentage of decisions to arrest that are accurate), "release decisions," and overall classification accuracy. As described above, these statistics depend directly on the base rate, and the high figures reported in the NHTSA studies would disappear in low base rate settings. In contrast, sensitivity and specificity are theoretically unaffected by base rates, and therefore provide more meaningful figures for comparisons across studies. Citing an overall, context-free accuracy rate as a guide to decision making is an oversimplification and is potentially misleading to police officers, court personnel, and juries.

Most of the NHTSA SFST studies have used officers' judgments of intoxication as the "test" rather than scores on the SFSTs. Error rates for the tests are available from Burns and Moskowitz (1977), Citek et al. (2003), Compton (1985), Good and Augsburger (1986), McKnight et al. (1995, 1999, 2002), Richman and Jakobowski (1994), and Stuster and Burns (1998), but they are influenced to an unknown degree by observations of nontest behavior and cues. Even so, they are substantial (a complete table is available from the author). These figures all apply to predictions of BAC as above or below a given criterion. No study has examined error rates for assessing mental, physical, or driving impairment.

Peer Review

The issue of peer review has been a contentious one, with prosecutors claiming that NHTSA's in-house review meets this standard while defense experts and attorneys have disagreed. There have been many critiques of the SFSTs and their supporting studies, with lawyers, psychologists, physicians, criminologists, and toxicologists contributing. Most of these have appeared in defense-oriented publications, or those where there is unlikely to be any oversight on scientific issues. Busloff (1993) appears to have produced the first review written by a non-lawyer and published in a peer-reviewed journal, whereas Booker's (2004) review is highly critical. On the other hand, neither Busloff nor Booker is a psychologist, and it is unclear if the peer-review criterion, as articulated in *Daubert*, encompasses review by scientists outside the profession of the author. The present paper is the only review by a psychologist published in a refereed scientific journal.

General Acceptance

This criterion also poses difficulties for a technique that has no natural home within the sciences. HGN derives from medical practice and laboratory research on the effects of drugs on ocular-motor functioning. It was then utilized by police surgeons in Finland to assess suspected drunken drivers, and subsequently adapted by research psychologists for use by police in the United States. The psychomotor tests have been used by police for generations, and were also studied by groups in Europe and the United States. However, there appears to be no scientific group whose general membership is knowledgeable about field sobriety testing: For psychology, optometry, medicine, neurology, ophthalmology, and criminalistics, it is a topic of interest only to a handful of researchers, writers, or practitioners, and is the subject of at most a few publications from each profession.

Critique of Partisan Arguments

Both prosecution and defense writers have made inaccurate assertions about the SFSTs. Prosecutors and other SFST advocates have portrayed the SFSTs as generally accepted in the scientific community without presenting evidence to support such claims. The frequent use of the phrase "alcohol-induced impairment" in reference to HGN is potentially misleading, as it equates impairment in gaze holding and one form of eye movement, under the artificial condition of head immobilization, with a term close to a legal conclusion. Although nystagmus might justifiably be called impairment of an ocular-motor system from a scientific perspective, in a legal setting, such wording is probably prejudicial since contrary to some recent prosecution claims (Abbott 2005), there is little evidence linking nystagmus to behavioral or visual impairment (Leigh and Zee 2006; S. Stevenson, personal communication). Lastly, revisions of the NHTSA student manual no longer advise readers not to not use Walk and Turn and One Leg Stand with older and overweight persons, and contain numerous erroneous claims (see Medley 2005). Claims for a lower reliability standard for psychomotor tests by Burns (2003) and Citek et al. (2003), or that HGN is the most reliable of the SFSTs, are unjustified.

Defense-oriented writers have asserted that persons with BACs above .15% are easily identified without the SFSTs, and thus unfairly contribute to the accuracy rates in the field studies. This may be true, but it is unproven and there is evidence to the contrary (Penttilä et al. 1974; Widmark 1932/1981), particularly for alcohol tolerant drinkers (Sullivan 1987; Urso et al. 1981). Some physicians have asserted that the NHTSA-suggested procedure for HGN is inherently invalid because it deviates from standard medical practice and invokes muscles beyond those required for lateral eye movements. However, the findings of McKnight et al. (2002) and the body of research cited in this paper, despite methodological limitations, challenges this assessment. Critics have cited facts or inconsistencies even when the error disadvantages the SFSTs. Such instances include the underestimation of a 45-degree angle from alignment with the shoulder (for angle of onset of nystagmus) and the possible presence of drug users in the 1981 laboratory study. Pangman's (1987) misstatement about normal endpoint nystagmus has been repeated many times, including by appellate courts.

Booker (2004) charged that NHTSA and the American Prosecutors Research Institute uncritically present supportive statistics while omitting unsupportive ones (e.g., false positive rates). This author's review supports this and many of Booker's other charges, but not all. For example, to show the SFST literature meets Langmuir's second criterion ("Maximum effect is produced by a causative agent of barely detectable intensity, and the magnitude of effect is substantially independent of the cause"), Booker stated that about half of adults have some nystagmus at maximum deviation and some will have nystagmus before 45 degrees. The relation of the argument to the point it seeks to prove is unclear. Booker also concluded that NHTSA and APRI engaged in deliberate fraud based on the erroneous citation of three studies. I don't believe this constitutes the "unambiguous, unimpeachable evidence" (p. 137) that Booker stated is required to substantiate such a charge—carelessness or incompetence are plausible rival hypotheses.

Summary of Findings and Needed Research

There are serious deficiencies in the research that supports the SFSTs. The fundamental question of validity has not been addressed rigorously, nor have the effects of many common, potentially confounding variables likely to be present at many DWI stops been examined. Research is long overdue to:

1. Determine the validity of the SFSTs for predicting BAC under rigorous, double blind conditions.

2. Determine whether the SFSTs are related to driving ability or measures of mental, physical, or visual impairment using double blind studies. If so, to what degree? If not, identify the tests as they are.
3. Determine if the poor reliability of the SFSTs can be improved. Is the problem with HGN primarily due to administration or scoring of the clues? Use of simple equipment (for HGN) and expanding the Yes/No scoring of clues for all the tests to a simple tally of errors may increase score reliability and validity.
4. Investigate the effects of age, gender, physical infirmity, sleepiness, fatigue, time of day, and anxiety/fear of evaluation on performance on the SFSTs, both in conjunction with alcohol consumption and in sober subjects. Norms based on large, representative samples, broken down by relevant factors, should be developed.
5. Identify the frequency of medical and psychiatric conditions that create false positive results. If such factors occur in significant numbers of cases, can they be identified by police officers through screening questions or observations?
6. Investigate the robustness of the SFSTs, particularly HGN, to variations in administration (speed of passes, elevation of stimulus) and environmental factors (e.g., police strobe lights, passing traffic, wind).
7. Determine whether the apparently strong prejudicial effects of the psychomotor tests on police officers and potential jurors are reliable and resistant to correction through jury instructions or deliberations.
8. Distinguish between establishing probable cause and evidence of intoxication in the application of the SFSTs, and then establish separate cut-off scores and diagnostic statistics for each test, cut-off score, and criterion.
9. Establish appropriate cut-off scores for different settings, considering the base rates and their effect on the positive predictive power of test results.
10. Research decision rules that combine the results from the three tests, possibly in conjunction with other behavioral and driving error observations.
11. Determine the reliability, validity, and discriminant validity of individual SFST clues for both BAC level and driving impairment. Determine internal consistency and redundancy.
12. Determine if the SFSTs are able to identify drinkers who have developed high levels of tolerance, and if so, whether SFST performance relates to driving performance.

Given the lack of evidence that the SFSTs are related to driving impairment, it is probably premature to limit investigations to these tests, particularly since the Walk and Turn and One Leg Stand are quite similar in the faculties assessed, depend on balance and an intact support system from the ankle through the spine, are prone to substantial false positive rates, and have been shown redundant in some studies. Technology may make sophisticated psychomotor and reaction time tests available roadside.

If the SFSTs are used as evidence to corroborate impairment, it would be desirable to quantify the level of evidence a test result constitutes. Six clues on Walk and Turn constitute stronger evidence than two clues, even though both are failing scores under current guidelines. A continuum of scores, with associated values for sensitivity and specificity for different criteria, would provide much more information.

Conclusion

This review of the SFST empirical research finds many deficiencies and unanswered questions. The SFSTs are not validated as tests of impaired driving or as indicators of loss of normal physical functioning: I could not identify a single study, published or not, that that has ever addressed these issues. There is only one peer-reviewed study that reported moderate correlations of SFST performance with decrements in cognitive performance. The SFSTs do show substantial correlations with BAC in most studies, subject to the limitations cited throughout this paper. HGN has repeatedly demonstrated higher correlations with BAC than the psychomotor tests, with some of the supportive findings published in peer-reviewed journals by authors not associated with NHTSA. However, the SFSTs cannot be used to estimate BAC in court and lack specificity for alcohol. The limited reliability data suggest that variations in administration or scoring from one police officer to another will be a substantial source of error, regardless whether BAC or behavioral impairment is the criterion.

The SFSTs were introduced into widespread use before thorough testing was completed and the results independently replicated. Even more than 20 years later, many basic questions concerning their use have not been answered. The effects of many variables, including medical conditions, fatigue, and fear, have not been examined. Further research is needed to justify their widespread use and to establish whether, in light of the current legal environment, the current SFST battery is the best available, or good enough, for distinguishing those who are impaired from those who are not.

Acknowledgements The author would like to thank H. Anthony Semone, H. D. Kirkpatrick, Jorey Krawczyn, Robert McIntyre, Joel Wiesen, J. Ray Hays, and Troy McKinney who provided their support and feedback on this article.

Appendix

Summary of Studies that Examined Either the SFSTs or Constituent Elements

Study	Setting	Test	Crit.	BAC Crit.	*N*	BR	Accur.	Sens.	FPR	*LR*	*d*	Corr.
Pentilla et al. (1974)	F	SP[a]	T	.10	408	.85	.84	.88	.43	2.1	–	–
Tenhu and Pentilla (1976)	F	SP	T	.10	1952	.81	.85	.90	.36	2.5	–	–
Lehti (1976)	L	AON	T		–							–.76[b]
Burns and Moskowitz (1977)	L	SFSTs[c]	J	.10	238	.27	.76	.84	.27	3.1	–	.67
	L	GN	T	.10	231	.27	.82	.68	.13	5.2	–	.67
	L	WAT	T	.10	229	.27	.75	.60	.19	3.1	–	.55
	L	OLS	T	.10	229	.27	.76	.65	.20	3.2	–	.48
Tharp et al. (1981a, b)	L	SFSTs	J	.10	441[d]	.28	.81	.64	.12	5.3	–	.65
	L	AON	T		42							–.78[b] L –.74[b] R
	L	AON	T		438	.28	.78, .88[e]	–	–	–	–	–.72[b] L –.71[b] R
Anderson et al. (1983)	L, F	HGN	T	.10	1072[f]	–	.77,[g].82	–	–	–	–	–
	L, F	WAT	T	.10	1072[f]	–	.68,[g] .80	–	–	–	–	–
	L, F	OLS	T	.10	1072[f]	–	.65,[g] .78	–	–	–	–	–
	L, F	2 Test[h]	T	.10	1072[f]	–	.80,[g] .83	–	–	–	–	–
Giguire (1983)	DC	HGN	T	.10	23	.43	.52	.90	.77	1.2[i]	–	–
	DC	WAT	T	.10	23	.43	.70	.50	.15	3.3[i]	–	–
	DC	OLS	T	.10	23	.43	.74	.50	.08	6.5[i]	–	–
Compton (1985)	SCP	HGN[j]	T	.10	300[k]	.21	.74	.95	.31	3.0	–	–
	SCP	HGN and Obs.[j,l]	J	.10	300[k]	.21	.74	.95	.32	3.0	–	–
Norris (1985)	L	AON	T		38, 88[m]							.73, .90[m]
Good and Augsburger (1986)	F	HGN[n]	T	.10	2429	.92	.90	.96	.82	1.2	0.9[o]	–
Golding and Dobie (1986)	ER	AON[p]	T	.10	46	–	–	.96	–	–	–	–
Golding and Dobie (1986)	F	AON	T	.10	149[q]	–	–	–	–	–	–	–.89[b]
Streff et al. (1989)	SG	HGN	T	.10	70	.23	.76	.50	.17	3.0	–	.51[r]
		OLS	T	.10	70	.21	.84	.67	.11	6.1	–	.60
Sussman et al. (1990)	F–B	SFSTs[s]	J	.10	96	.29	.82	.75	.14	5.2	–	.70
Foss et al. (1990)	F	SFST	J	.10	190	.13[t]	.59	.68	.42	1.6	–	.20[u]
Perrine et al. (1993)	F	HGN	T	.05	185	.41[t]	–	.72	.08	9.0	–	.49[u]
	F	HGN	T	.08	185	–	–	.79	.17	4.6	–	[v]
	F	HGN	T	.10	185	.24[t]	–	.85	.23	3.7	–	[v]
	F	WAT	T	.10	185	.24[t]	–	.73	.49	1.5	–	.13[u]
	F	OLS	T	.10	185	.24[t]	–	.56	.27	2.1	–	.16[u]
McKnight and Langston (1993)	–	WAT	T	.08	–	–	–	.44	.22	2.0	0.6	–
	–	WAT	T	.10	–	–	–	.54	.25	2.2	0.8	–
	–	OLS	T	.08	–	–	–	.46	.09	5.1	1.2	–
	–	OLS	T	.10	–	–	–	.58	.17	3.4	1.2	–
Richman and Jakoblowski (1994)	L	HGN	T	.08	210[k]	.43	.87	.88	.13	6.6	–	–
Cole and Nowaczyk (1994)	L	WAT and OLS[w]	J	.10	294[k]	.00	.54		.46			
Kennedy et al. (1994)	L	HGN	T		67[x]							.77
	L	WAT	T		67							.54
	L	OLS	T		67							.57
Kennedy et al. (1995)	L	HGN	T		62[k]							.64
	L	WAT	T		62[k]							.37
	L	OLS	T		62[k]							.16[y]
Burns and Anderson (1995)[z]	F	SFSTs	J	.05	234	.79	.86	.89	.24	3.7	–	–
Burns and Dioquino (1997)	F	SFSTs	J	.08	256	.80	.93	.96	.18	5.3	–	–
Stuster and Burns (1998)[aa]	F	SFSTs	J	.04	83	.78	.80	.78	.17	4.7	–	–

continued

Study	Setting	Test	Crit.	BAC Crit.	*N*	BR	Accur.	Sens.	FPR	*LR*	*d*	Corr.
	F	SFSTs	J	.08	297	.72	.91	.98	.29	3.4	–	.69
	F	HGN	T	.08	290	.72	.88	.98	.37	2.6	–	.65
	F	WAT	T	.08	271	.72	.79	.92	.53	1.7	–	.61
	F	OLS	T	.08	273	.73	.83	.92	.41	2.2	–	.45[ab]
McKnight et al. (1999)	L	HGN	T	.10	78[k]	–	–	–	–	–	–	.69
McKnight et al. (1999)	F–B	HGN	T	.10	60[k]	.35	.93	1.00	.10	9.8	4.5	.77
Burns et al. (2000)	L	GN	T	.08	48	.51	.93	1.00	.14	7.0	–	–
McKnight et al. (2002)	L	HGN	T		–							.55
McKnight et al. (2002)	F	HGN	T	.04	240[k]	–	–	.79	.38	2.1	1.3	.56
Citek et al. (2003)[ac]	L	HGN	T	.08	284[k,ad]	–	.73–.77	.80–.89	.29–.56	1.9–2.8	1.3–1.6	.63
Citek et al. (2003)[ab]	L	HGN	T	.10	284[k,ac]	–	.61–.69	.89–.97	.41–.46	1.7–2.2	1.4–1.7	[u]

Note. Values were not given or could not be calculated for some cells. This is indicated by a dash ("–"). Values were not applicable for some cells, which are left blank. DNMD = distinct nystagmus at maximum deviation; AON = angle of onset of nystagmus. Setting = setting in which data were collected: F = field (actual traffic stop), L = laboratory, DC = driving course, SCP = simulated check point, F–B = field, boating stop, SG = social gathering, ER = emergency room; Crit. = Criterion: whether decision was based on test score (*T*) or examiner's judgment (J); BR = Base rate; *N* = Number of observations, some studies were repeated measures designs; Sens. = Sensitivity; Spec. = Specificity; % Corr. = percentage of correct decisions overall; FPR = false positive rate; *d* = Cohen's *d*. All correlations and *LR*s were statistically significant at $p < .05$ unless noted. Data from Hlastala *et al.* not reported because they are presented earlier and are derived from and partially redundant with Stuster and Burns (1998). Under "Corr.," L = left; R = right

[a] Based on "coarsely divided" nystagmus

[b] Correlations for AON with BAC are negative, as increased alcohol leads to onset of nystagmus at lower angles of deviation from the visual midline

[c] This study used Gaze Nystagmus rather than HGN, and judgments were based on a total of six tests. However, the results for a discriminant analysis using Gaze Nystagmus, Walk and Turn, and One Leg Stand were reported to be "essentially the same" as for the six test battery

[d] Included subjects who were retested

[e] First figure is for judgment of AON of 45 degrees or less; second figure is for continuous measure; both made with aid of an apparatus

[f] This is the total number of subjects tested; some may have refused individual tests and the actual number may be considerably less

[g] Calculated from data in Tharp et al. based on *N* of 441

[h] Two test combination of HGN and WAT

[i] 95% CI includes 1.0, not significant. Author did not report any statistics in the original article and did not disclose anything about the raters or their training

[j] HGN was performed through a car window with subject seated

[k] Repeated measures design

[l] Included divided attention sobriety tests

[m] The first figure is the correlation with BAC, the second with the BAC estimate from a breath test

[n] Officers apparently administered other field sobriety tests as well, which likely contributed to arrest decisions

[o] Reported by Citek et al. (2003)

[p] HGN in this study apparently did not include DNMD and held the stimulus closer than specified by NHTSA standards. Although the smooth pursuit phase was administered, analyses were based only on AON

[q] Two or more officers made observations on 149 subjects. The article does not state if their observations were pooled or treated separately

[r] Very limited training probably led to lower correlation with BAC than in other studies

[s] Included tests not part of the SFST battery

[t] Base rates were not reported and were estimated from the full data set

[u] This coefficient is Somer's *d*, not *r*

[v] Same value as for the cell above

[w] Included tests in addition to WAT and OLS

[x] Number of observations inferred from number of subjects and description of design

[y] This correlation not significant at $p < .05$

[z] Used HGN cutoff score of two clues. This study examined accuracy at .10% but did not report the results

[aa] Officers in this study had access to portable breath tests and there were no observers to ensure they were not used

[ab] This figure is probably reduced by atypical scoring by some officers

[ac] Range of numbers reflects figures for HGN administered in standing, sitting, and supine positions. This applies to ranges for all figures reported for this study

[ad] This number refers to the number of BAC readings obtained. However, each subject was tested by two or three officers

References

References marked with an asterisk were used in the quantitative analyses.

Abbott, W. C. (2005). SFSTs: A blessing or a curse? *The Texas Prosecutor, 35*(5), 8–10.

American Optometric Association (1993). House of Delegates Resolution 6, Horizontal Gaze Nystagmus as a field sobriety test. *Journal of the American Optometric Association, 64*, 653.

American Prosecutors Research Institute (1999). Horizontal Gaze Nystagmus—The science and the law: A resource guide for judges, prosecutors, and law enforcement. Alexandria, Virginia: Author. Retrieved August 2, 2006, from http://www.ndaa.org/pdf/sci_law2.pdf.

American Prosecutors Research Institute (2003). Horizontal Gaze Nystagmus state law summary and chart. Retrieved August 2, 2006, from http://www.ndaa.org/pdf/hgn_state_case_law_chart_2003.pdf.

Anderson, T., Schwerz, R., & Snyder, M. (1983). *Field evaluation of a behavioral test battery for DWI.* Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration (DOT HS-806-475).

Aschan (1958). Different types of alcohol nystagmus. *Acta otolaryngologica, 140*, 169–178.

Bahill, A. T., Iandolo, M. J., & Troost, B. T. (1980). Smooth pursuit eye movement in response to unpredictable target waveforms. *Vision Research, 20*(11), 923–931.

Balaban, C. D. (2002). Neural substrates linking balance control and anxiety. *Physiology & Behavior, 77*, 469–475.

Bobo, J. (2003). Introduction: The best field sobriety test. *Admissibility of Horizontal Gaze Nystagmus evidence: Targeting hardcore impaired drivers.* Alexandria, Virginia: American Prosecutors Research Institute. Retrieved August 2, 2006, from http://www.ndaa.org/pdf/admissibility%20of%20hgn_april_2003.pdf.

Bolmont, B. (2005). Role and influence of moods including anxiety on motor control. In Clark (Ed.), *Causes, role and influence of mood states* (pp. 57–73). Nova Biomedical Books.

Booker, J. L. (2001). End-position nystagmus as an indictor of ethanol intoxication. *Science and Justice, 41*(2), 113–116.

Booker, J. L. (2004). The Horizontal Gaze Nystagmus test: Fraudulent science in the American courts. *Science and Justice, 44*(3), 133–139.

Bramness, J. G., Skurtveit, S., & Mørland, J. (2003). Testing for benzodiazepine inebriation—Relationship between benzodiazepine concentration and simple clinical tests for impairment in a sample of drugged drivers. *European Journal of Clinical Pharmacology, 59*(8–9), 593–601.

Brick, J., & Carpenter, J. A. (2001). The identification of alcohol intoxication by police. *Alcoholism: Clinical and Experimental Research, 25*(6), 850–855.

Burns, M. (2003). An overview of field sobriety test research. *Perceptual and Motor Skills, 97*, 1187–1199.

Burns, M. (in press). *The robustness of the Horizontal Gaze Nystagmus (HGN) test.* Washington, DC: U.S. Department of Transportation.

*Burns, M., & Anderson, E. W. (1995). *A Colorado validation study of the Standardized Field Sobriety Tests (SFST) battery.* Final Report to the Colorado DOT. Los Angeles, CA: Southern California Research Institute; Aspen CO: Pitkin County Sheriff's Office.

Burns, M., & Dioquino, T. (1997). *A Florida validation study of the standardized field sobriety test (SFST) battery.* Jacksonville, FL: Florida Department of Health.

*Burns, M., & Moskowitz, H. (1977). *Psychophysiological tests for DWI arrest*, final report, DOT-HS-802–424.

Burns, M., Fiorentino, D., & Stuster, J. (2000). *The observational threshold of Horizontal Gaze Nystagmus.* Unpublished manuscript. Retrieved August 2, 2006, from http://www.icadts.org/proceedings/2000/icadts2000–037.pdf.

Busloff, S. E. (1993). Can your eyes be used against you? The use of Horizontal Gaze Nystagmus in the courtroom. *The Journal of Criminal Law & Criminology, 84*(1), 203–238.

Citek, K. (2003). HGN and the role of the optometrist. *Admissibility of Horizontal Gaze Nystagmus evidence: Targeting hardcore impaired drivers.* Alexandria, Virginia: American Prosecutors Research Institute. Retrieved August 2, 2006, from http://www.ndaa.org/pdf/admissibility%20of%20hgn_april_2003.pdf.

*Citek, K., Ball, B., & Rutledge, D. A. (2003). Nystagmus testing in intoxicated individuals. *Optometry, 74*(11), 695–710. Available: http://www.ndaa-apri.org/pdf/nystagmus_testing.pdf.

Citron, J. (n.d.). How to be your own expert witness. Unpublished manuscript.

Coffey, M. (2004). DWI: Modern day Salem witch hunts. *The Champion, 28*(9), 51–54, 63. Retrieved August 2, 2006, from http://www.nacdl.org/public.nsf/01c1e7698280d20385256d0b00789923/0ce16f3b9551615c85256f6a00558f3a?OpenDocument.

Cohen, H. M. (2002). Field sobriety tests (FSTs). In R. E. Erwin (Ed.), *Defense of drunk driving cases, criminal/civil.* New York: Matthew Bender & Co., Inc.

Cole, R. M., & Cole, S. N. (1991). New proof that field sobriety tests are "failure-designed." *DWI Journal: Law & Science, 6*(2), 1–5.

Cole, S., & Nowaczyk, R. H. (1994). Field sobriety tests: Are they designed for failure? *Perceptual & Motor Skills, 79*, 99–104.

Cole, S., & Nowaczyk, R. H. (1995). The .10% solution. *The Champion, 19*(7), 40–43.

*Compton, R. P. (1985). Pilot test of selected DWI detection procedures for use at sobriety checkpoints. National Highway Traffic Safety Administration, DOT-HS-806–724.

Cowen J. D., & Jaffee, S. G. (1989). Proof and disproof of alcohol-induced driving impairment through evidence of observable intoxication and coordination testing. *American Jurisprudence Proof of Facts, 3d, 9*, 459–596.

Cowen, J. D., & Jaffee, S. G. (1990). Field sobriety tests: The flimsy scientific underpinnings. *DWI Journal: Law & Science, 5*(12), 1–7.

Daubert vs. Merrell Dow Pharmaceuticals (1993). 509 U.S. 579, 113 S.Ct. 2786, 125 L.Ed. 2d.469.

Dell'Osso, L. F., & Darloff, R. B. (2005). Nystagmus and saccadic intrusions and oscillations. In W. Tasman, & E. A. Jaeger (Eds.), *Duane's clinical ophthalmology.* Philadelphia: Lippencott, Williams & Wilkins (Vol. 2, Rev. ed., Chap. 11).

Emerson v. State (1994). 880 S.W.2d 759 (Tx. Crim. App.).

Essen, R. J., & Levenstein, M. (1989). Roadside sobriety tests: Both "scientific" and unreliable. *DWI Journal: Law & Science, 4*(11), 6–10.

Ettinger, U., Kumari, V., Crawford, T. J., Davis, R. E., Sharma, T., & Corr, P. J. (2003). Reliability of smooth pursuit, fixation, and saccadic eye movements. *Psychophysiology, 40*, 620–628.

Faigman, D. L., & Monahan, J. (2005). Psychological evidence at the dawn of law's scientific age. *Annual Review of Psychology, 56*, 631–659.

Forkiotis, C. J. (1987). Optometric expertise: The scientific basis for alcohol gaze nystagmus. *Optometric Extension Program Foundation, 59*(7), 1–15.

*Foss, R. D., Voas, R. B., & Perrine, M. W. (1990). Technological and behavioral predictors of blood alcohol concentration. In M. W. Perrine (Ed.), *Alcohol, drugs and traffic safety–T89. Proceedings of the 11th International Conference on Alcohol, Drugs and Traffic Safety*, Oct 24–27, Chicago.

Foster, K. R., & Huber, P. W. (1999). *Judging science: Scientific knowledge and the federal courts.* Cambridge: MIT Press.

*Giguire, W. (1983). Impairment caused by moderate blood alcohol levels in a closed course: Preliminary demonstration. In. S. Kaye & G. Meier (Eds.), *Alcohol, drugs and traffic safety*–T'95, Vol. 1. *Proceedings of the 9th International Conference on Alcohol, Drugs, and Traffic Safety*, pp. 529–542. Washington, DC: U.S. Dept. of Transportation.

Glass, G. (2003). Field sobriety test: Problem with Horizontal Gaze Nystagmus as an assumption of blood alcohol level of .08% or higher. *Voice for the Defense*, 16–21.

*Goding, G. S., & Dobie, R. (1986). Gaze nystagmus blood alcohol. *Laryngoscope, 96*, 713–717.

*Good, G. W., & Augsburger, C. R. (1986). Use of Horizontal Gaze Nystagmus as a part of roadside field sobriety testing. *American Journal of Optometry & Physiological Optics, 63*(6), 467–471.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education*. Auckland: McGraw Hill.

Halperin, E., & Yolton, R. L. (1986). Is the driver drunk? Oculomotor sobriety testing. *Journal of the American Optometric Association, 57*, 654–657.

Hartje, W., Steinäuser, D., & Kerschensteiner, M. (1978). Diagnostic value of saccadic pursuit eye movement in screening for organic cerebral dysfunction, *Journal of Neurology, 217*, 253–260.

Heilbrum, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*(3), 257–272.

Hlastala, M. P., Polissar, N. L., & Oberman, S. (2005). Statistical evaluation of standardized field sobriety tests. *Journal of Forensic Science, 50*(3), 1–8.

Hogan, M. J. (2003). Divided attention in older but not younger adults is impaired by anxiety. *Experimental Aging Research, 29*(2), 111–136.

Honts, C. R., & Amato-Henderson, S. L. (1995). Horizontal Gaze Nystagmus test: The state of the science in 1995. *North Dakota Law Review, 71*, 671–700.

Hlastala, M. P. (1998). The alcohol breath test—A review. *Journal of Applied Physiology, 84*(2), 401–408.

Hlastala, M. P., & Anderson, J. C. (2007). The impact of breathing pattern and lung size on the alcohol breath test. *Annals of Biomedical Engineering, 35*(2), 264–272.

Iacono, W. G., Pelquin, W. J., Lumry, A. E., Valentine, R. H., & Tuason, V. B. (1982). Eye tracking in patients with unipolar and bipolar affective disorders in remission. *Journal of Abnormal Psychology, 91*(1), 35–44.

Iowa v. Murphy (1990). 451 N.W.2d 154 (Iowa).

Jones, A. W., Beylich, K. M., Bjorneboe, A., Ingum J., & Morland, J. (1992). Measuring ethanol in blood and breath for legal purposes: Variability between laboratories and between breath-test instruments. *Clinical Chemistry, 38*, 743–747.

Kathmann, N., Hochrein, A., Uwer, R., & Bondy, B. (2003). Deficits in gain of smooth pursuit eye movements in schizophrenia and affective disorder patients and their unaffected relatives. *American Journal of Psychiatry, 160*, 696–702.

Kelly, T. H., Foltin, R. W., Serpick, E., & Fischman, M. W. (1997). Behavioral effects of alprazolam in humans. *Behavioral Pharmacology, 8*(1), 47–58.

*Kennedy, R. S., Turnage, J. J., Dunlap, W. P., & Drexler, J. M. (1995). Calibration of a portable, automated, posture assessment system using graded blood alcohol level: Comparison with the standardized field sobriety and cognitive tests. In C. N. Kloeden & A. J. McLean (Eds.), *Alcohol, drugs and traffic safety*–T'95, Vol. 1. *Proceedings of the 13th International Conference on Alcohol, Drugs, and Traffic Safety*, pp. 455–462. Adelaide, Australia: NHMRC Road Accident Research Unit, Univ. of Adelaide 5005.

Kennedy, R. S., Turnage, J. J., Rugotzke, G. G., & Dunlap, W. P. (1994). Indexing cognitive tests to alcohol dosage and comparison to standardized field sobriety tests. Journal of Studies on Alcohol, 55*, 615–628.

Kobatake, K., Yoshii, F., Shinohara, Y., Nomura, K., & Takagi, S. (1983). Impairment of smooth pursuit eye movements in chronic alcoholics. *European Neurology, 22*, 392–396.

Langenbucher, J., & Nathan, P. E. (1983). Psychology, public policy, and the evidence for alcohol intoxication. *American Psychologist, 38*(10), 1070–1077.

Langmuir, I. (1969). Pathological science. *Physics Today, 42*, 36–48.

Laurrell, H., McLean, A. J., & Kloeden, C. N. (1990). The effect of blood alcohol concentration on light and heavy drinkers in a realistic night driving situation. Unpublished manuscript, NHMRC Road Accident Research Unit, University of Adelaide.

Leggett, A., McCloskey, S., Malfer, R., & Kerrigan, S. (2007). Driver behavior and impairment symptoms in cannabinoid positive subjects arrested for driving under the influence of drugs (DUID). *Poster Session Presented at the Annual American Academy of Forensic Sciences Meeting*, San Antonio, TX.

*Lehti, H. M. J. (1976). The effect of blood alcohol concentration on the onset of gaze nystagmus. *Blutalkohol, 13*, 411–414.

Leigh, R. J., & Zee, D. S. (2006). *The neurology of eye movements* (4th ed.). New York: Oxford University Press.

Lipset, M. W., & Wilson, D. B. (2001). *Practical meta-analysis, Applied social research methods series, Vol 49*. Thousand Oaks: Sage.

McCloskey, S., Leggett, A, Malfer; R., & Kerrigan, S. (2007). Driving under the influence of methamphetamine: Comparison of driving behavior and impairment symptoms in subjects arrested for driving while intoxicated. *Poster Session Presented at the Annual American Academy of Forensic Sciences Meeting*, San Antonio, TX.

McKnight, A. J., & Langston, E. A. (1993). The use of video in training for standard field sobriety testing (SFST)/Evaluation of the Training for Standardized Field Sobriety Testing (SFST). Report prepared under NHTSA Contract DTNH 22-91-C-05109.22.

*McKnight, A. J., Langston, E. A., Lange, J. E., & McKnight, A. S. (1995). Development of standardized field sobriety tests for lower BAC limits. Report prepared under NHTSA Contract no. DTNH 22-92-C-00700. Washington, DC: US Department of Transportation, National Highway Traffic Safety Administration.

*McKnight, A. J., Lange, J. E., & McKnight, A. S. (1999). Development of a standardized boating sobriety test. *Accident Analysis and Prevention, 31*(1–2), 147–152.

*McKnight, A. J., Langston, E. A., McKnight, A. S., & Lange, J. E. (2002). Sobriety tests for low alcohol blood concentrations. *Accident Analysis and Prevention, 34*(3), 305–311.

Meaney, R. (1996). Horizontal Gaze Nystagmus: A closer look. *Jurimetrics Journal, 36*, 383–407.

Medley, J. M. (2005). Standardized field sobriety exercises: Screening tests for arrest or forensic evidence? Unpublished manuscript.

Mullen, R., Hardy, L., & Tattersall, A. (2005). The effects of anxiety on motor performance: A test of the conscious processing hypothesis. *Journal of Sport & Exercise Psychology, 27*(2), 212–225.

Mundt, J. C., Perrine, M. W., & Searles, J. S. (1997). Individual difference in alcohol responsivity: Physiological, psychomotor, and subjective response domains. *Journal of Studies on Alcohol, 58*(2), 130–140.

Munez, D. P., Armstrong, I. T., Hampton, K. A., & Moore, K. D. (2003). Alerted control of visual fixation and saccadic eye movement in attention-deficit hyperactivity disorder. *Journal of Neurophysiology, 90*, 503–514.

National Highway Traffic Safety Administration, U.S. Dept. of Transportation (2002). *DWI Detection and Standardized Field Sobriety Testing, Instructors Manual* (DOT HS 178 R1/02).

National Highway Traffic Safety Administration, U.S. Dept. of Transportation (2004). *DWI Detection and Standardized Field Sobriety Testing, Participant Manual* (HS 178 R9/04).

National Highway Traffic Safety Administration, U.S. Dept. of Transportation (1984). *Improved sobriety testing* (DOT HS 806 512).

Nichols, D. H. (1998). *Drinking/driving litigation: Criminal and civil.* St. Paul: West Group.

Norris, J. (1985). The correlation of angle of onset of nystagmus with blood alcohol level: Report of a field trial. *California Ass'n Criminalistics Newsletter*, 21.

Noteboom, J. T. (2001). Acute stressors activate the arousal response and impair performance of simple motor tasks. *Dissertation Abstracts International: Section B: Sciences and Engineering, 61*(12-B), 6428.

Nowaczyk, R. H., & Cole, S. (1995). Separating myth from fact: A review of research on the field sobriety tests. *The Champion, 19*(7), 40–43.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.

O'Keefe, M. (2001). Standardized field sobriety tests: A survey of police surgeons in Strathclyde. *Journal of Clinical Forensic Medicine, 8*, 57–65.

Pagano, M. R., & Taylor, S. P. (1979). Police perceptions of alcohol intoxication. *Journal of Applied Psychology, 10*(2), 166–174.

Pangman, W. A. (1987). Horizontal Gaze Nystagmus: Voodoo science. *DWI Journal: Law & Science, 2*(3), 1–6.

Papafotiou, K., Carter, J. D., & Stough, C. (2005). An evaluation of the sensitivity of the Standardised Field Sobriety Tests (SFSTs) to detect impairment due to marijuana intoxication. *Psychopharmacology, 180*(1), 107–114.

Penner, D. W., & Coldwell, B. B. (1958). Car driving and alcohol consumption: Medical observations on an experiment. *Canadian Medical Association Journal, 79*, 793–800.

Pentillä, A., Tenhu, M., & Kataia, M. (1971). *Clinical examination for intoxication in cases of suspected drunken driving*. Helsinki, Fin.: Statistical and Research Bureau of Talja.

Pentillä, A., Tenhu, M., & Kataia, M. (1974). *Clinical examination for intoxication in cases of suspected drunken driving: II.* Helsinki, Fin.: Liikenneturva.

People v. Leahy (1994). 8 Cal 4th 587, 34 Cal. Rptr. 2d 663, 882 P.2d 321.

People v. Loomis (1984). 156 Cal. App. 3d 1, 203 Cal. Rptr. 767 (Cal. Super. 1984).

People v. Vega (1986). 145 Ill. App. 3d 996.

People v. Williams (1992). 3 Cal. Rptr. 2d 130 Cal. App. 5 Dist.

*Perrine, M. W., Foss, R. D., Meyers, A. R., Voas, R. B., & Velez, C. (1993). Field sobriety tests: Reliability and validity. In H.-D. Utzelmann, G. Berghaus, & G. Kroj (Eds.), *Proceedings of the 12th International Conference on Alcohol, Drugs, and Traffic Safety*, pp. 1133–1138. Cologne: Verlag TUV Rheinland.

*Perrine, M. W., Peck, R. C., & Fell, J. C. (1988). Epidemiologic perspectives on drunk driving. Paper presented at the U.S. Surgeon General's Workshop on Drunk Driving, Washington, DC: Government Printing Office.

Pierce, F. A. (1984). *Evaluation of pilot programs in standardized field sobriety testing and drug recognition expert concepts.* Michigan Department of State Police, Traffic Services Division. Interim Report, Project MAL.

Pijpers, J. R., Oudejans, R. R. D., & Bakker, F. C. (2005). Anxiety-induced changes in movement behaviour during the execution of a complex whole-body task. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 58*(3), 421–445.

Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep, 19*(4), 318–326.

Pisoni, D. B., & Martin, C. S. (1989). Effects of alcohol on the acoustic phonetic properties of speech: Perceptual and acoustic analysis. *Alcohol: Clinical and Experimental Research, 13*(4), 577–587.

Price, P. B. (1996). Field sobriety testing. *The Champion, 20*(7), 46–50.

Price, P. B., & Cole, S. (2001). NHTSA field sobriety tests: Validation and invalidation. *The Champion, 25*(3), 25, 37–42.

*Richman, J. E., & Jakobowski, J. (1994). The competency and accuracy of police academy recruits in the use of the Horizontal Gaze Nystagmus test for detecting alcohol impairment. *New England Journal of Optometry, 47*(1), 5–8.

Rouleau, M. (1990). Unreliability of the Horizontal Gaze Nystagmus test. *American Jurisprudence Proof of Facts 3d, 4*, 439–495.

Ross R. G., Olincy, A, Harrisb, J. G., Sullivan, B, & Radant, A. (2000) Smooth pursuit eye movements in schizophrenia and attentional dysfunction: Adults with schizophrenia, ADHD, and a normal comparison group. *Biological Psychiatry, 48*(3), 197–203.

Rubenzer, S. J. (2003a). The psychometrics and science of the standardized field sobriety tests, Part 1. *The Champion, 27*(4), 48–54.

Rubenzer, S. J. (2003b). The psychometrics and science of the standardized field sobriety tests, Part 2. *The Champion, 27*(5), 40–44.

Rubenzer, S. J. (2006). A history and review of the standardized field sobriety tests. Unpublished manuscript.

Rubenzer, S. J., & Stevenson, S. (2007). Horizontal Gaze Nystagmus: A review of vision science issues (submitted for publication).

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

Senter, R. J. (1969). *Analysis of data: Introductory statistics for the behavioral sciences.* Glenview, IL: Scott, Foresman and Company.

Silber, B. Y., Papafotiou K., Croft R. J., & Stough, C. K. K. (2005). An evaluation of the sensitivity of the standardised field sobriety tests to detect the presence of amphetamine. *Psychopharmacology, 182*(1), 153–159.

Simon, S. (n.d.). Steve's attempt to teach statistics. Retrieved August 2, 2006, from http://www.childrens-mercy.org/stats/journal/oddsratio.asp.

Simpson, G. (1988). Attacking NHTSA's three-test field sobriety assessment. *DWI Journal: Law and Science, 3*(5), 9–12.

State v. Dahood (2002a). #96-JT-707 (Concorde District Court).

State v. Dahood (2002b). 814 A.2d 159 (NH).

State v. Homan (2000). 89 Ohio St.3d 421, 732 N.E.2d 952.

State v. Meador (1996). 674 So.2d 826 (Fla.App 4 Dist.).

State v. Murphy (1990). 451 N.W.2d 154 (Iowa).

State v. Nagel (1986). 30 Ohio App.3d 80, 506 N.E.2d 285.

State v. Superior Court (1986). 149 Ariz. 269, 718 P.2d 171.

State v. Witte (1992). 251 Kan. 313, 836 P.2d 1110.

*Streff, F. M, Geller, E. S., & Russ, N. W. (1989). Evaluation of field sobriety tests for use by social hosts. In B. Perrine (Ed.), *Proceedings of the 11th International Conference on Alcohol, Drugs, and Traffic Safety*, pp. 450–455. Chicago, IL: National Safety Council.

*Stuster, J., & Burns, M. (1998). *Validation of the Standardized Field Sobriety Test battery at BACs below .10 percent.* Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration (DOT-HS-808–839).

Sullivan, J. B., Hauptman, M., & Bronstein, A. C. (1987). Assessing degree of intoxication in humans with high plasma alcohol concentrations. *Journal of Forensic Sciences, 32*(67), 1650–1665.

*Sussman, E. D., Needalman, A., & Mengert, P. H. (1990). *An experimental evaluation of a field sobriety test battery in the marine environment*. Report prepared under Contract no. DOT-CG-D-04-90. Washington, DC: US Department of Transportation, US Coast Guard.

Taubenslag, W. W., & Taubenslag, M. J. (1975). *Selective Traffic Enforcement Program* (STEP). Final Report Purchase Order No. 813430, NHTSA, Washington, DC.

Taylor, L., & Oberman, S. (2005). *Drunk driving defense* (6th ed.). Aspen Publishers.

Tenhu, M., & Pentillä, A. (1976). The value of nystagmus in the practical examination of suspected drunken drivers. *Forensic Science, 8*, 199–200.

*Tharp, V., Burns, M., & Moskowitz, H. (1981a). *Development and field test of psychophysical tests for DWI arrest*. Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration (DOT-HS-805-864).

Tharp, V., Moskowitz, H. A., & Burns, M. M. (1981b). Circadian effects on alcohol gaze nystagmus. *Psychophysiology, 18*(2), 5–8.

Tiffany, D. V. (1986). Optometric expert testimony: Foundations for the horizontal gaze nystagmus test. *Journal of Optometry and the Law, 57*(9), 705–708.

Toglia, J. U. (1976). *Electronystagmography: Technical aspects and atlas*. Springfield, IL: Charles C. Thomas.

U.S. v. Horn (2002). 185 F.Supp.2d 530 (D.Md.).

Urso, T., Gavaler, J. S., & Van Thiel, D. H. (1981). Blood ethanol levels in sober alcohol users seen in an emergency room. *Life Sciences, 23*, 1053–1056.

Vingilis, E. (1983). Drinking divers and alcoholics: Are they from the same population? *Research Advances in Alcohol and Drug Problems, 7*, 299–342.

Widmark, E. M. P. (1981). *Principles and application of medicolegal alcohol determination* (R. C. Baselt, Trans.). Davis, CA: Biomedical Publications (Original work published 1932).

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York: Wiley-Interscience.