The Champion

June 2003 , Page 40

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

By Steve Rubenzer

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

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

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

1. Screening questions for possible medical problems and conditions should be standardized and validated.

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

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

3. SFST training does not emphasize rigorous adherence to the standardized instructions.

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

4. SFST training materials do not address how instructions are to be delivered (attitude, speed, and tone).

Should the officer be polite? Authoritative? Commanding? Is it all right to be impatient, surly, and condescending? How does this affect performance? What about speed of delivery? Should the officer's demeanor facilitate maximum performance? That is the usual standard for neuropsychological tests.[4] In contrast, some officers appear to make the tests harder by delivering instructions in a rapid, bored, monotone voice.

It is unlikely that the officers in the laboratory studies, using volunteers and monitored by the researchers, adopted the hostile, impatient demeanor sometimes displayed by officers during SFST administrations. To the extent that arresting officers behave differently than the

officers in the NHTSA studies (which was not recorded), the validation evidence is diminished.

5. For the Walk-and-Turn, a variety of line situations are permitted.

There is no research on the effect of using an imaginary line, a crooked line, an offset line, or one that the line creates an uneven surface.

6. What constitutes "demonstrates understanding"?

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

7. Scoring rules are often inadequately specified.

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

8. It is unclear, both in the studies and the Student Manual, what the criteria are for failing the SFST battery.

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

9. Officers are not specifically directed to record their observations immediately.

Failure to do so encourages a tendency to assign scores consistent with the officer's arrest decision and, for example, to remember seeing a particular clue in both eyes rather than one. As the authors of the 1981 laboratory study stated, "…many of the advantages of standardized scoring are lost when the scoring is left to memory."[5]

Reliability and validity problems

1. The SFSTs have not been subjected to a rigorous "blind" assessment of their validity.

As discussed above, none of the studies of the SFSTs have been truly double blind, as expected in medical research. The laboratory studies came close; the field studies do not. (Standard 1.17)

2. The effects of fatigue, drowsiness, circadian rhythm, driver stiffness or roadside conditions on SFST performance have not been adequately investigated. (Standard 10.1)

The angle of onset of nystagmus was found to advance five degrees in the hours after midnight, while the other laboratory studies were conducted during daytime hours.[6] In the 1981 study, the authors stated that exercise, sleep loss, elevated temperatures, and antihistamines are associated with increased body sway.7 Strobe and emergency lights, gusts of wind from passing traffic—all have unknown effects on SFST performance and validity given the limitations of the field studies.

3. Drivers suspected of DWI and subjected to the SFSTs may be highly anxious, which alone or in combination with small amounts of alcohol, may influence their performance.

In the laboratory studies, subjects were volunteers who had no reason to be anxious, aside from possible self-consciousness. There are theoretical reasons to believe that fear, anxiety, or stress may affect performance on the WAT and OLS,[8] and no study has demonstrated these factors are not relevant.

4. The clues for the WAT and OLS lack documentation of their individual validity and reliability.

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

5. Reliability data are lacking or below accepted standards for psychological tests used for making decision about individuals.

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

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

6. Standard errors of measurement (SEM) are not provided. (Standard 6.5)

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

7. SFSTs have not been normed on sober people.

As acknowledged in the 1981 study, "Balance tests of various sorts show large individual differences in the performance of sober individuals…."[12] When most psychological tests are developed, they are tested on a large sample to determine what is "normal." The Personality Assessment Inventory is a self-report test designed to assess psychopathology. Before it was published, the author administered it to some twelve hundred psychiatric patients — the intended population for the test. But he also administered it to over twelve hundred volunteers from around the country. Then, volunteers were dropped in order to obtain a census-projected nationally representative sample in terms of age, race, and education.[13] The SFSTs have never been administered to a large, representative group of sober people. There is no "normal" score.

8. There is very limited data on the SFSTs for people under 21 or over 50-55. (Standard 3.6)

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

9. SFSTs have questionable validity for those who are elderly, in poor physical condition, or overweight.

If the SFSTs are of questionable validity for people more than 50 pounds overweight,[14] what about short people who are 45 or 40 pounds over the ideal? Proportionately, a person who is 4'8" and 40 pounds overweight is likely to be more physically impaired than someone 6'3" and 51 pounds overweight. Why does the test suddenly become invalid when one goes from 50 to 51 pounds over the ideal? Obviously, the impediment due to weight is likely to be gradual. The same issue applies to people in their late 50's versus the arbitrary cutoff of 60[15] or 65.[16] Physical health and condition are likely to be more important than age. (Standards 7.2, 7.3, 10.1)

10. Even NHTSA claims the SFSTs, when optimally used, are only 80 percent accurate.[17]

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

Another perspective on SFST accuracy is provided by using a bathroom scale as an analogy. Even a cheap scale might be expected accurate within a few pounds. Yet, the NHTSA authors state, " [I]t is unrealistic to attempt to use behavioral tests to discriminate BACs in a +.02% margin around a given level."[18] This is equivalent to a 100-pound woman stepping on a scale, seeing a reading of 120, and being told the scale is functioning within its design limits. And this is under ideal conditions. But how well can police officers actually estimate individuals' BACs? In the 1981 laboratory study, police officers' estimates of BAC (measured by Intoximeters) were incorrect by an average of .03 percent[19] — meaning approximately half the errors were larger than this.

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

Let's return to the analogy of a 100-pound woman stepping on a bathroom scale using the SFST BAC estimation errors. Using the most conservative average error reported (.03 percent), and using standard tools to create a confidence interval,[20] we find that a 100-pound woman would observe a scale reading of between 25 and 175 pounds on 95 of 100 trials. The other five percent of readings would be more inaccurate. In the 1981 field study, officers' average BAC estimates were off by an incredible .077 percent before training and a whopping .0537 percent after training.[21] Creating a 95 percent confidence interval from the "before training" figure (.077 percent) means our 100-pound woman will weigh anywhere from -93 to 293 pounds on our SFST bathroom scale — 95 percent of the time.

Miscellaneous Issues

1. The SFSTs have been evaluated primarily by NHTSA-supported researchers, with no rigorous evaluation by disinterested researchers in a field settings.

Replication by impartial researchers is the sine qua non of reliable scientific knowledge.

2. SFSTs have usually been evaluated in high base rate settings where up to 92 percent of the persons tested were legally intoxicated.

Base-rates have a major effect on the confidence that can be given to a test result.[22] In both the laboratory and field studies, the majority of subjects or drivers tested were intoxicated so generalization to settings (sobriety checkpoints or daytime stops) where the incidence of DUI is much lower is not warranted. An earlier NHTSA study[23] showed high rates of false positives when the frequency of intoxicated (BAC > .10 percent) drivers was experimentally set to 48 percent. HGN, either alone or in combination with observations of driver behavior and appearance, showed false positive rates of up to 75 percent for those with a BAC between .05 percent and .09 percent. Officers who received only three hours of training in administration of HGN[24] assessed 24 percent of those in the .00-.04 percent BAC range as impaired — and the majority of these were probably completely sober.

3. SFST scoring is potentially biased by the officer's suspicion of intoxication.

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

4. The SFSTs may be harder than driving.

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

5. Although the SFSTs were not designed as indications of driving impairment and have undergone little validation for this purpose, they are still frequently admitted as evidence for establishing the driver was impaired.

The SFSTs were expressly developed and validated to distinguish between BACs of above and below .10 percent — not driving impairment. Marcellin Burns has emphasized this distinction,[29] but NHTSA materials[30] and court decisions[31] wrongly equate the two terms. While the SFSTs attempt to gauge BAC, NHTSA plainly states "Impairment varies widely among individuals with the same BAC level."[32]

Only a couple of studies have attempted to correlate SFST scores with driving impairment. In one of the original NHTSA laboratory studies, subjects were tested on both the SFSTs and adivided attention performance test designed to simulate the demands of driving. Each SFST correlated about .30 with the different performance measures. When the results of the tests were combined statistically, the two psychomotor tests (WAT and OLS) carried all the weight — HGN added nothing.[33] Other NHTSA-associated researchers stated "there is no evidence that the eye movements that constitute Nystagmus seriously impair the visual processes involved in driving or operating a boat."[34] A third group of NHTSA researchers evaluated dozens of behavioral tests to determine their potential to assess driver impairment and other desired qualities. These authors recommended a completely different battery than the SFSTs,[35] and the SFSTs received low ratings of relevance to driving skills: HGN received a rating of 0 (zero on a scale of 0-100) for its value in assessing driver impairment, while WAT and OLS received ratings of 40 and 20, respectively.[36]

6. SFSTs, particularly HGN, arguably are more prejudicial than probative on the issue of impairment.

When a person suspected of DWI/DUI has difficulty performing a field sobriety test, the jury viewing the performance may logically assume the suspect is drunk. Given the circumstances, this is the natural interpretation. A study published by Cole and Nowacyk[37] had 21 completely sober people perform the sobriety tests (not including HGN) and other tasks. Police officers perceived 46 percent of the subjects performing sobriety tests as drunk and worthy of arrest. If prejudicial value = high and probative value = low or medium, then high > low or medium = nonadmissible. Meaney argued that a negative HGN is more probative than a positive finding because "no study suggests the possibility of intoxication without nystagmus."[38]

Conclusion

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

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

Notes

National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000) at VIII-3.
Emerson v. State, 880 S.W.2d 759 (Tx.Cr.-App., 1994), Homan v. State, 732 N.E.2d 952 (Ohio 2000).
Crime to Court: Rappin' Up the DUI (instructional video), South Carolina ETV, in cooperation with the South Carolina Criminal Justice Academy, SC Law Enforcement Division (1995). The Truth Is in the Eyes (instructional video) cited in New Hampshire v. Dahood, supra.

Muriel D. Lezak, Neuropsychological Assessment 139-140 (3rd ed., 1995).
V. Tharp et al., Development and Field Test of Psychophysiological Tests for DWI Arrest, Final Report, DOT-HS-805-864 (1981) at 70.
Id. at 17.
Id. at 83.
Phillip B. Price, Sr., Fear and Sobriety Testing (2000) (unpublished manuscript, available from Mr. Price's office, 217 Randolph Avenue, Huntsville, AL 35801, 256-536-6000, dwilawyer @aol.com).
Jum C. Nunnally & Ira H. Bernstein, Psychometric Theory 265 (3rd ed. 1994).
K.J. Snapper et al., An Assessment of Behavioral Tests to Detect Impaired Drivers, Final Report, DOT-HS-806-211, (1981) at 3-34 to 3-37.

Jum C. Nunnally & Ira H. Bernstein, Psychometric Theory 214 (3rd ed.1994). Hoi K. Suen, Principles of Test Theories 141 (1990).
Tharp et al., supra at 83.
Leslie Morey, Personality Assessment Inventory Professional Manual (1989).
National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000)
Improved Sobriety Testing, DOT-HS-806-512 (1984) at 7.
National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000).
Id. at VIII-12.
Marcelline Burns & Herbert Moskowitz, Psychophysiological Tests for DWI Arrest, Final Report, DOT-HS-802-424 (1977) at 41.
V. Tharp et al., supra, at 72.
A confidence interval is set using the standard deviation rather than the average deviation. For a normal distribution, the standard deviation equals 1.25332 * average deviation. (see R. J. Senter, Analysis of Data: Introductory Statistics for the Behavioral Sciences 92 (1969). A 95 percent confidence interval is set by taking the mean +/- twice the standard deviation.
V. Tharp et al., supra, at 63.
See Louis M. Hsu, Diagnostic Validity Statistics and the MCMI-III, 14 Psych. Assess. 410, 410-411 (2002).
Richard P. Compton, Pilot Test of Selected DWI Detection Procedures for Use at Sobriety Checkpoints, National Highway Traffic Safety Administration, DOT-HS-806-724 (1985).
Three hours of training in administration of HGN may not be atypical — most of the 24 hour NHTSA student course is devoted to topics other than administration of the SFSTs. The Texas A&M University System TEEX — Law Enforcement Training Division, Texas Standardized Field Sobriety Testing Program Instructor Manual (2002).
Marcelline Burns, supra.
K.J. Snapper et al., supra.
National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000) at VII-3.
Michael O'Keefe, Drug Driving — Standardized Field Sobriety Tests: A Survey of Police Surgeons in Strathclyde. 8 J. Forensic Med. 57, 60-61 (2001).
Marcelline Burns, supra.
National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000). Horizontal Gaze Nystagmus: The Science & The Law (A Resource Guide for Judges, Prosecutors, and Law Enforcement), National Highway Transportation Safety Administration, http://www.ntsa.dot.gov/peopole/injury/nystagmus/hgntxt.html
State v. Baue, 258 Neb. 968 (2000), U.S. v. Horn, 185 F.Supp.2d 530, 561 (D.Md. 2002).
National Highway Traffic Safety Adm., U.S. Dept. of Transp., HS 178 R2/00, DWI Detection and Standardized Field Sobriety Testing, Student Manual (2000) at VII-6.
Marcelline Burns & Herbert Moskowitz, Psychophysiological Tests for DWI Arrest, Final Report, DOT-HS-802-424 (1977) supra, at 54.
A. James McKnight et al., Development of a Standardized Boating Sobriety Test, 31 Accid. Anal. & Prev. 147 (1999).
K.J. Snapper et al., supra, at 4-2.
Id. 3-34 to 3-37.
Spurgeon Cole & Ronald H. Nowaczyk, Field Sobriety Tests: Are They Designed for Failure? 79 Percep. & Motor Skills, 99 (1994).
Joseph R. Meaney, Horizontal Gaze Nystagmus: A Closer Look, 36 Jurimetrics J. 383, 406 (1996).

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