

LAW OFFICE OF  
ALLEN M. TRAPP, JR., P.C.  
Telephone: (770) 830-8560  
Fax: (770) 830-8565

Mailing Address:  
P.O. Box 2206  
Carrollton, Ga. 30112

Street Address:  
114B Corporate Drive  
Carrollton, Ga. 30117

***A NEW LOOK AT  
“VALIDATED” NHTSA  
SFST STUDIES:  
HOW THEY HELP YOU***

Website: [www.westgeorgiaduiattorney.com](http://www.westgeorgiaduiattorney.com)

E-mail: [amtjr@iscope.com](mailto:amtjr@iscope.com)

## **Background: A Quick Trip Back in Time**

For several decades law enforcement officers have employed a variety of field sobriety tests to determine if a person is driving under the influence of alcohol. Once upon a time they were looking for drunk drivers, but as we all know it has been over thirty years since Jimmy Carter was President. Twenty or thirty years ago it was common for Georgia State Troopers to have a motorist blow into their "Smokey Bear" hats to guesstimate how much the person had consumed. Needless to say, some in the law enforcement community wanted something more.

Back in the 1970's (and in some jurisdictions to this very day) a wide variety of field sobriety tests were employed, ranging from blowing in the hat to tracing on paper. Beginning in 1975 studies were sponsored by NHTSA through a contract with the Southern California Research Institute to determine which of the field sobriety tests were the most accurate. In other words, SCRI was not charged with developing tests but only with evaluating those already in use.

The National Highway Traffic Safety Administration (NHTSA) is a branch of the Department of Transportation (DOT) and in the mid-1970's, NHTSA requested proposals to conduct research on identifying the best "field sobriety tests" that an officer could use at roadside. Marcelline Burns, a research psychologist and the director of the Southern California Research Institute and her group created a technical and cost proposal, submitted it, and the Southern California Research Institute (SCRI) was awarded a contract by NHTSA to do the research. The first study was reported in 1977. The work by Burns on the 1977 study began in 1975 wherein she, after a literature search of related material, participated in ride-a-longs with police officers all over the United States and developed a list of 16 tests thought to be feasible as potential "sobriety" tests. Using those tests, SCRI conducted some pilot tests with a small group of people, and selected six tests for the 1977 study.

This involved 238 drinking subjects and ten police officers and lasted about one year. The 1977 SCRI study, funded through NHTSA, resulted in the recommendation of the use of three of the six tests, namely, the walk-and-turn, the one-leg stand, and the horizontal gaze nystagmus (referred to as "alcohol gaze nystagmus" in the 1977 final report). The other tests used in the study were the finger to nose, the finger count, and the tracing test. The Romberg test, alphabet test and subtraction tests were interchangeably used. The individuals selected as

subjects for the research were licensed drivers and alcohol consumers. They were instructed not to eat for four hours before they were given measured doses of alcohol; however, the subjects themselves did not know the amount of alcohol they consumed. Their breath alcohol concentrations were measured, and then they were subjected to the six tests listed above. It also should be noted that all of the testing was done in the laboratory setting.

The 1977 study concluded that the Romberg test and the finger to nose test merely reflected the presence of alcohol, but "did not increase the predictive ability of testing." In other words, the finger to nose and Romberg test did not add anything to the predictability of a subject's level of intoxication. It is also interesting to note that the finger count, finger to nose, Romberg, alphabet, and tracing tests were not recommended for use as sobriety tests. Although they were used and were a part of the 1977 NHTSA study, none were selected as being indicators of anything, let alone as indicators of intoxication. Some interesting statistics came out of the 1977 study. Of primary significance was the error rate of the 10 officers involved in the study. ***Their error rate was an astounding 47 percent! You have to read the report, because you will not find this in a NHTSA manual.*** That is to say, in the 1977 study the officers made the decision to "arrest" a total of 101 people. Of those people "arrested", 47 percent had a BAC under 0.10 percent. (Source: 1977 report, page 25) This was totally unacceptable, even according to the author(s) of the study. Marcelline Burns later tried to attribute the high error rate to the inexperience of the officers used in the study.

If this was true, it would seem inexplicable that they would again use inexperienced officers in the 1981 NHTSA study. It is of primary significance to note that approximately **80% of the subjects used in the 1977 study were in their twenties, and about two thirds of them were male.** (Source: 1977 report, page 18, Figure 4) Perhaps it is just the author's experience, but muscle tone and physical dexterity began deteriorating within a year or two following his 10<sup>th</sup> high school reunion.

## **THE 1981 STUDY**

The 1977 study recommended further review, and NHTSA awarded the Southern California Research Institute a second contract for retesting and standardization. This second study resulted in the 1981 NHTSA report. In the 1981 study only the three test battery was used. The 1981 study, like the 1977 study, was done only in a laboratory setting, except for a handful

of experiments conducted at the end of the study. Burns states that the officers again made their decisions to arrest or not to arrest based on the prediction that the subject's BAC was over or under a 0.10. There were 296 subjects in the 1981 study.

Some "divided attention" components were added in midstream during the 1981 study. For example, Burns describes a divided attention component of the walk and turn test as the portion of the test wherein the subject is requested to stand with one foot in front of the other on the line, while listening to the instructions. This is also referred to as the "instructional phase". The standardization aspect of the 1981 study was to establish consistency in the administration guidelines, the instructions, the demonstrations, and the scoring. The objective was to ensure that if an officer in Florida does the three test battery and an officer in Oregon also does it, the two officers should reach the same conclusions. The order in which the tests were given was considered by Ms. Burns to be irrelevant.

In the 1981 study, *out of 118 decisions by the officers to arrest, 32 percent of them were wrong.* (Source: 1981 report, page 27, Table 8) This is only slightly better than the 1977 study which had a 47 percent error rate of false arrests. Also, in the 1981 study, *18 percent of the subjects who had no alcohol in their system were misjudged by the officers to be impaired.* (Source: 1981 report, page 22, Table 4) Ms. Burns takes a rather twisted attempt at explaining this. She opines that the study was done "next to the drug capital of the world.". In other words, she hints that since none of these people were screened for drugs that they may have been impaired on substances other than alcohol. This simply is unsound logic. If one were to accept her logic, perhaps this could be grounds to invalidate the entire study, since none of the subjects, including those who had ingested alcohol, had been screened for drugs. *The officers in the 1981 project believed 31 percent of the people who were at a 0.05 BAC to be impaired.*

However, the most interesting statistics from the 1981 study as discussed by Cole and Nowaczyk, (*The Champion*, August 1995) involve the "dosing differential" of the subjects tested. Most of the subjects (*78 percent*) were dosed with either *high BAC* (about 0.15) *or low BAC* (0.05 and below). (Source: 1981 report, page 15, Table 4) These should have been easy decisions since it should, as a practical matter, be easy for the officers to score an individual as being above a 0.10 BAC when they are 0.15 BAC and above. The same would be true of someone 0.05 and below. *NHTSA claims an overall accuracy rate of 0.80 when using the*

*three-test battery*, however, this overall accuracy rate of .80 is questionable when over two-thirds (**78 percent**) *should be considered "gimmies" (either dosed high or low, hence the "dosing differential")*. In other words, the data of the individuals dosed between 0.05 and 0.15 would undoubtedly have an accuracy rate of much less, however, that data is unavailable. Cole and Nowaczyk opine that one factor in determining the "improval" of the false arrest numbers (47 percent in 1977 down to 32 percent in 1981) could be due in part to the dosing differential.

The number of subjects dosed in the mid-range (**0.05 to 0.15**) went down from 27 percent (Source: 1977 report, page 19, Figure 5) in the 1977 study to **22 percent** in the 1981 study. In other words, only 22 percent of the subjects in the 1981 study were in the more difficult to determine range of between 0.05 to 0.15 BAC. The 1981 study claims a "reliability study" as part of the research in 1981. Reliability basically refers to consistency, or the ability to get the same results each time. The reliability portion consisted of asking 145 of the subjects back for retesting two weeks after the original study. *The "reliability factor" was a 0.77*. This "reliability correlation coefficient" is based on a scale from almost zero to a 1.00. It is interesting to note that a correlation coefficient of 0.9 or above is expected for academic reading tests such as the SAT. This inter-rater reliability coefficient dropped to 0.57 (Source: Page 35, Table 14) when done by different officers. So, when different officers tested the same subjects at the same dose level, the reliability level was very pathetic, and far below scientific acceptability. Dr. Spurgeon Cole states that the scientific community expects reliability coefficients to be in the upper 0.80s or 0.90s for a test to be scientifically reliable. This statistic is quite significant and is one of the reasons that judges should not allow an officer to testify that the accused failed the particular test.

The age and gender of the subjects used in the 1981 project, as with the 1977 study, are highly significant when considering any interpretation of the results. In the 1981 study a whopping 80% of the subjects were between the ages of 21 and 34. Again, as with the 1977 study about two thirds of them were male. (Source: 1981 report, page 14, Table 2) The use of a predominately male population in their twenties means that we should question the applicability of the test results to the population as a whole.

## THE GOOD-AUGSBERGER STUDY

One of the earliest non-NHTSA studies of field sobriety tests was published in the *American Journal of Optometry & Physiological Optics* by two optometrists at Ohio State University, Gregory W. Good and Carol R. Augsberger, in 1986. The article is highly complimentary of the SFST program at the Ohio State Highway Patrol academy and regurgitates NHTSA statistics without any critical analysis. The article dutifully reports that 92% of subjects scoring four “points” or higher on the HGN registered BAC’s above .10.

However, the authors overlooked the bad news, false positives, although their own charts published with the article reveal a startling rate. Fully 81.5% of those with BAC’s under .10 also demonstrated four or more clues. Source: *American Journal of Optometry & Physiological Optics*, Vol. 63, No. 6, page 470, Table 2). Although NHTSA trumpeted that the exercise is “92% accurate in identifying intoxicated people”, there was a concerted effort to ignore the fact that the data says the test is 82% **inaccurate** as applied to INNOCENT PEOPLE.

## THE FIELD VALIDATION STUDIES

The three-test battery of Standardized Field Sobriety Tests has been promoted by NHTSA over the past twenty-plus years and has been adopted by all fifty states. In three highly publicized “validation studies” NHTSA claims to have found the proof that these FST’s are valid measures of BAC. All of the field studies are pretty consistent in terms of low false negative rates. However, the same cannot be said of false positives, and that is what should concern us - the wrongly accused being arrested because of flawed “science.” The reason why so many people over .08 and .10 BAC show 4+ HGN clues is that so many people have 4+ HGN clues at .04, .05, and .06 BAC.

Three SFST validation studies were undertaken between 1995 and 1998.

Colorado - 1995

Florida - 1997

San Diego - 1998

NHTSA says that the Colorado study was the first full field study that utilized law enforcement officers experienced in the use of SFST’s. They also claim that correct arrest decisions were made 93% of the time based on the three-test battery, which was substantially better than the initial study results. (Source: Colorado study, Figure 5)

The Florida SFST field validation study was undertaken in order to answer the question of whether SFST's are valid and reliable indices of the presence of alcohol when used under "present day" traffic and law enforcement conditions. According to NHTSA the correct arrest decisions were made 95% of the time. (Source: Florida study, Figure 5) NHTSA goes on to say that the validation studies have shown that the SFST 3-test battery is the only scientifically validated and reliable method for discriminating between impaired and unimpaired drivers.

The San Diego study was undertaken because NHTSA wanted the SFST's to be recognized as capable of discriminating BAC's above and below .08, as NHTSA and MADD campaigned to reduce per se limits to .08 across the nation. Not surprisingly, the "correct arrest decision" was made 91% of the time at the .08 level and above. (Source: San Diego study, page 18)

### **THE REST OF THE STORY.**

1. The Colorado study found 1 in 8 people under .05 had 4+ HGN clues. (Source: Colorado study, Figure 12)
2. In the Florida study 18% of people below .08 BAC had 5 or 6 HGN clues. (Source: Florida study, Table 4)
3. AND... it suggests that over 50% had at least 4 clues, but it doesn't just come out and say it. NHTSA attempts to conceal these numbers by saying that half of the correctly released drivers had 0 or 2 HGN clues. That suggests that **HALF** of the correctly released drivers (under .08) had **MORE** than 2 clues. (Source: Florida study, Section V, Subsection C, Topic 1)
4. The Florida report also acknowledges that 67% of all incorrect arrests (under .08) had all 6 clues. (Source: Same as above)
5. In the San Diego study, the one most frequently cited, the average BAC of those arrested was .15. (Source: San Diego study, page 16)
6. The same was true in Florida. (Source: Florida study, Section V, Subsection A, Topic 1)

7. In Colorado the average was a bit higher at .152. (Source: Colorado study, Section V, Subsection D)
8. False positives (person arrested had BAC under .08) were six times as common as a false negatives (person not arrested had BAC of .08 or more). (Source: San Diego study, page 18, Figure 4)
9. All police officers participating in the San Diego study were equipped with NHTSA-approved portable breath testing devices.
10. In the Florida study those wrongly arrested averaged 3.6 clues on the WAT and 2 clues on the OLS. (Florida study, Table 4)
11. The Colorado study reported that only 13 errors of administration and 6 errors in instructions were observed in 305 SFST administrations (only 41% were observed). (Source: Colorado study, Section V, Subsection F)
12. No errors were observed in the 313 SFST batteries given in the Florida study, although only two-thirds of the administrations were monitored by civilian employees of the same department. (Florida study, Section V, Subsection E)
13. The data from the San Diego study shows that the FST's miserably fail the "specificity" test when it comes to those with a BAC between .07 and .09. Specificity may be defined as the percentage of true negatives (i.e., under .08) who are correctly classified as such by the test. The rate was only 36%. (Source: Hlastala, Polissar, Oberman, *Statistical Evaluation of Standardized Field Sobriety Tests*, J. Forensic Sci., 2005: Vol. 50, No. 3, Table 3. The rate was only slightly better at 44% for those between .06 and .10 and increased to only 55% for those between .05 and .11.



## WHEN THE JUDGE ALLOWS THE COP TO TESTIFY LIKE AN EXPERT ON STATISTICS

Has it happened to you - waiting to hear that same old "77% accurate at predicting if someone is under the influence" (or the more recent version "in excess of .08") you've been hearing for over a decade - when suddenly a cop blurts out something about 90+ percent? Or has a prosecutor directly asked the officer if recent research has shown the tests to be more than 77% accurate? Even if you object on lack of expertise, hearsay, or inadequate foundation, too many judges will let it in as "part of the officer's training." Here are some suggestions if a prosecutor ever gets to pollute your jury with this toxic waste.

First... even SFST instructors are no more qualified to testify about the field studies than any other officers. If you get a chance to voir dire them or if you only get to cross-examine them after the judge has ruled against you, ask questions like:

How many drivers were examined in the San Diego study?

How many were actually over .08?

How many drivers were arrested or released contrary to the NHTSA standardized clues criteria?

What ratio of drivers under the limit were wrongfully arrested?

What was the base rate of intoxicated individuals in the Florida study?

Better yet... ask the officer if he has read the entire study. When he or she says "No," the judge should rule the officer is not an expert in this area, or if the judge has allowed in this testimony, you have eviscerated the state's expert in the eyes of the jury.

How can an expert testify as such without knowing the rate of error? How can they testify without knowing the degree of confidence with which they can express their opinion? Will this FST instructor know what a "degree of confidence" is? If you need more questions, try these:

Ask him where and when he got his training or experience in statistical analysis.

Ask him whether he is or was a researcher or has ever conducted scientific research.

Ask if he even knows what the *standard deviation* is, and if he claims to know, then ask him to calculate it.

Ask him if he knows what a confidence interval is and how to calculate it.

Ask him what he has done, besides read the study report, to evaluate the alleged results contained in the study report. The list could go on and on, but you should get the idea.

Of course, he will know none of this and after a while, it will become painfully obvious that all he knows is that he read it and the report says something. Your objections should include hearsay and lack of qualifications as an expert. He may be an expert in administering the SFSTs and in teaching others to administer them, but that does not make him an expert in the science of evaluating the statistical accuracy and reliability of them as measures of anything, much less intoxication.

The SFST instructor training has nothing to do with the substance of SFSTs. It is a course on how to teach. Being an instructor no more qualifies someone to testify about the studies than being an astronaut. But, being an instructor sounds good and often gets sold by prosecutors as worth far more than it is. When we buy it, we buy are buying little more than the proverbial swampland.

If you learn the true data from these field studies, you can cross them very effectively. If they are allowed to testify over your objection, eat them alive. For example:

"Officer Expert, isn't it true that over one in three drivers under .08 had at least four HGN clues in the San Diego field study?" (Actually 30 out of 81 or 37%. Source: San Diego study, page 21, figure 5).

"Isn't it true that in the San Diego study over one-half of those with a BAC under .08 had two or more clues on the WAT?" (40 out of 76, or 53%, Source: same as above).

"Isn't it true that in the San Diego study over 40% of the people with a BAC under .08 had two or more clues on the OLS?" (31 of 75, or 41%. Source: same as above.)

"Officer Expert, isn't it true that 70% of everyone under .08 in the Florida study showed two or more clues on the Walk & Turn?" (Note: This 70% number only includes the "correctly released." The actual percentage would be higher if the wrongly arrested were included. Source: Florida study, Section V, Subsection C, Topic 2.

**READ THE DATA FROM THESE STUDIES- IT'S GREAT.**

## **THE ROBUSTNESS OF THE HORIZONTAL GAZE NYSTAGMUS TEST**

This publication was released by NHTSA in September, 2007, with the objective of proving that deviations in the administration of HGN did not really affect the results. Indeed, the report concluded that “The data demonstrate the validity of the HGN test with both standard and varied testing procedures. The variations did not alter the occurrence of, or the observations of, HGN.” Once again, NHTSA’s conclusions are not necessarily consistent with the raw data, so let’s take a look at some of that data.

Even when performed at the correct speed 20 of 26 test subjects (77%) with a BAC under .08 demonstrated four or more clues. (Table 10 on page 14). The NHTSA claim that variations do not affect the validity of the tests is clearly debunked by the data in Table 13 on page 18. When the stimulus was held four inches above eye level (as opposed to the two inch “standard”), 22 of 24 subjects (92%) with a BAC under .08 showed four or more clues.

NHTSA seemed to take great comfort in the fact that officers noted more clues when the stimulus was held ten inches from the nose instead of the standard twelve to fifteen inches. That much is true. However, if you look at the raw data (Table 15, page 21), it shows that officers recorded four or more clues for 22 out of 26 test subjects with a BAC under .08 when the stimulus was twelve to fifteen inches from the nose and four or more clues for 24 out of 26 subjects when the stimulus was ten inches away. In other words, the false positive rate increased from 85% to 92%.

NHTSA’s explanation is that people with BAC’s in the .04 to .08 range can be expected to have four or more clues. In other words, they have attempted to re-define what is a false positive. However, they did get something right. “Analysis revealed that more than 40% of the variability in HGN observations was accounted for by significant differences between participants.” If that is so, how can it be seriously argued the test is valid?

The foregoing are just some casual observations made by a small town country lawyer from Georgia. Please read the studies and see what you can find and then share it with the rest of us. These studies are full of discrepancies. Use them, and you will have more clients who are amazed by your knowledge, more juries who are impressed by your erudition, more judges

whose eyes will be opened and perhaps more respectful of you in the future, and more importantly, more two-word verdicts.

### References

Burns, M. & Moskowitz, H., *Psychophysical Tests for DWI Arrest*. Final Report. DOT-HS-802-424, NHTSA, 1977.

Tharp, V., Burns, M. & Moskowitz, H., *Development and Field Test of Psychophysical Tests for DWI Arrest*. Final Report, DOT-HS-805-864, NHTSA, 1981.

Burns, M. & Anderson, E., *A Colorado Validation Study of the Standardized Field Sobriety Test (SFST) Battery*. Final Report, Project No. 95-408-17-5, NHTSA, 1995.

Burns, M. & Dioquino, T., *A Florida Validation Study of the Standardized Field Sobriety Test (SFST) Battery*. Final Report, Project No. AL-97-05-14-01, NHTSA, 1997.

Stuster, J. & Burns, M., *Validation of the Standardized Field Sobriety Test Battery at BAC's Below 0.10 Percent*. Final Report, NHTSA, 1998.

Cole, S. & Nowaczyk R H., *Field Sobriety Tests: Are They Designed for Failure? Perceptual and Motor Skills*, 1994: 79: 99-104.

Hlastala, Polissar, Oberman, *Statistical Evaluation of Standardized Field Sobriety Tests*, J. Forensic Sci., 2005: Vol. 50, No. 3.

Nowaczyk, R. H. & Cole, S., *Separating Myth From Fact: A Review Of Research On The Field Of Sobriety Tests*, The Champion, Aug., 1995, p. 40.

Good, G. & Augsburger, C., *Use of Horizontal Gaze Nystagmus as a Part of Roadside Sobriety Testing*, American Journal of Optometry & Physiological Optics, 1986, Vol. 63. No. 6, pp. 467-471.

