IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MARYLAND

	*	
UNITED STATES OF AMERICA	*	
v.	*	**Case No. 00-946PWG**
ERIC D. HORN	*	

* * * * *

MOTION IN LIMINE TO EXCLUDE THE GOVERNMENT'S FIELD SOBRIETY TEST EVIDENCE AND REQUEST FOR A HEARING

Defendant Eric Horn, by and through counsel, James Wyda, Federal Public Defender for the District of Maryland, and Sasha Natapoff, Assistant Federal Public Defender, respectfully moves in limine pursuant to Rule 104 and Rule 702, Fed. R. Evid., to exclude any and all expert testimony and evidence regarding the field sobriety tests (FSTs) administered to him because the tests are unreliable and the information provided by the tests is overly prejudicial. In support of his motion, Mr. Horn alleges as follows:

1. According to police reports, on June 28, 2000, Mr. Horn was stopped by Officer Daniel Jarrell at the Harford Gate of Aberdeen Proving Ground. Officer Jarrell performed several so-called field sobriety tests, or FSTs, on Mr. Horn. Specifically, Officer Jarrell performed the horizontal gaze nystagmus test (HGN), the walk and turn test (WAT), and the one-leg stand test (OLS). Officer Jarrell also asked Mr. Horn to perform a "finger dexterity test" and to recite the alphabet. Mr. Horn was subsequently charged with driving under the influence of alcohol in violation of Md. Code Ann., Transp. § 21-902.
2. The defense expects the government to introduce evidence of those field sobriety tests at

trial.

3. Field sobriety tests are technical tests administered under special conditions by persons with specialized training in the administration and interpretation of those tests. They are therefore "technical or other specialized knowledge" under *Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137, 147 (1999), and must satisfy the Supreme Court's test for reliability and relevance established in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993). *See* attached Memorandum of Law.

4. The field sobriety tests administered to Mr. Horn are methodologically unreliable and therefore should be excluded under Fed. R. Evid.702, and *Kumho Tire.*

5. The results of FSTs are only tenuously related to the issue of intoxication and their admission as evidence would be unduly prejudicial. In the event that one or more of the tests are considered non-technical evidence, they should be excluded under Fed. R. Evid. 701 and 403 as prejudicial and unhelpful lay testimony. *See* attached Memorandum of Law.

6. The question of whether FSTs are scientifically reliable is a complex question that would benefit greatly from expert testimony and oral argument. In addition, the government bears the burden of establishing that Officer Jarrell is a qualified expert whose "testimony is based upon sufficient facts or data" and who "applied the principles and methods [of FSTs] reliably to the facts of the case." Fed. R. Evid. 702. Accordingly, the defendant requests a hearing to address the scientific reliability, relevance, and admissibility of the field sobriety tests administered to Mr. Horn.

Respectfully submitted,

JAMES WYDA
Federal Public Defender
for the District of Maryland

_____
Sasha Natapoff
Assistant Federal Public Defender
100 S. Charles Street
Tower II, Suite 1100
Baltimore, Maryland 21201
(410) 962-3962

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that on this ___ day of February, 2001, a copy of the foregoing Motion in Limine to Exclude the Government's Field Sobriety Test Evidence and Request for a Hearing was delivered to Paul Marone, Special Assistant United States Attorney, U.S. Army Garrison, Building 310, Wing 10, Aberdeen Proving Ground, Maryland, 21001.

_____
Sasha Natapoff
Assistant Federal Public Defender

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MARYLAND

	*	
UNITED STATES OF AMERICA	*	
v.	*	**Case No. 00-946PWG**
ERIC D. HORN	*	

* * * * *

MEMORANDUM OF LAW IN SUPPORT OF DEFENDANT'S MOTION IN LIMINE TO EXCLUDE THE GOVERNMENT'S FIELD SOBRIETY TEST EVIDENCE AND REQUEST FOR A HEARING

INTRODUCTION

Field sobriety tests, or FSTs, are psychomotor tests that attempt to measure a person's physical coordination and/or ability to perform more than one task at a time, so-called "divided attention" tests. Because alcohol can impair these functions, police use FSTs to assist them in determining whether a person's cognitive and motor skills may be impaired by alcohol consumption.

The National Highway Traffic Safety Administration (NHTSA) has developed standardized procedures for the administration of the three FSTs which NHTSA considers the most reliable. *See* NHTSA Manual, Ex. B. These standardized FSTs (SFSTs) are taught to and used by police officers across the country and were administered to Mr. Horn in the instant case. The three standardized FSTs are: the horizontal gaze nystagmus test (HGN), the walk-and-turn test (WAT), and the one-leg stand test (OLS).

There are also many other FSTs that have not been studied or standardized by NHTSA. In this case, Officer Jarrell instructed Mr. Horn to perform a "finger dexterity test" and told him

to recite a portion of the alphabet. Because there is absolutely no documented scientific validity to the non-standardized FSTs, this Memorandum focuses on the SFSTs recommended for use by NHTSA.

The SFSTs administered to Mr. Horn are designed to be used by police officers to establish probable cause to arrest individuals who are under suspicion of driving while intoxicated and to support the administration of a breathalyzer test which measures more directly a person's blood alcohol content (BAC). As direct, independent evidence of intoxication, however, SFSTs are extremely unreliable and have an immense margin of error. Furthermore, individual officers often administer the tests differently or under non-ideal testing circumstances, further reducing their reliability. While some courts have admitted FSTs results into evidence, the recent *Daubert/Kumho* line of cases and the newly amended Fed. R. Evid. 702 now forbid reliance on those old, lax standards. FSTs do not meet the new, more rigorous standards of *Kumho Tire* and Rule 702, and therefore the government should not be permitted to introduce them – either the details of their administration or their results – into evidence in criminal trials.

Even if the Court were to treat some or all of the FSTs as non-technical evidence, they are sufficiently unreliable and prejudicial as to warrant exclusion under Fed. R. Evid. 701 and 403. The mere fact that an arresting officer testifies that the person "failed" a particular field sobriety test is likely to prejudice the defendant. In light of the error rates and unreliability of FSTs, the administration and results of FSTs should be excluded as unhelpful and unduly prejudicial.

ARGUMENT

I. FIELD SOBRIETY TESTS INVOLVE TECHNICAL AND SPECIALIZED EXPERTISE AND ARE THEREFORE SUBJECT TO THE DAUBERT TEST

In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), the Supreme Court held that scientific testimony must satisfy certain criteria of reliability and relevance in order to be admissible in federal court. District courts must inquire, *inter alia*, whether the evidence is susceptible of testing, whether it has a known error rate, whether it has been subject to peer review, and whether it is generally accepted by the relevant scientific community. *Id.* at 593-94. Six years later, in *Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137 (1999), the Court expanded the scope of *Daubert* to include not only "scientific" evidence but any technical or specialized knowledge as well. Amended Fed. R. Evid. 702 tracks this development and requires that "scientific, technical, or other specialized knowledge" meet the rigorous standards laid out in *Daubert*.

Prior to *Kumho Tire*, some courts treated some FSTs as mere observations that could be admitted without scientific foundation. Such reasoning is no longer available in federal court: as the following discussion demonstrates, each of the three SFSTs are the sort of specialized knowledge that *Kumho Tire* brought within the purview of *Daubert* and therefore are inadmissible without proper foundation. *See, e.g., Volk v. United States*, 57 F. Supp.2d 888, 894 n.3 (N. D. Cal. 1999) (FSTs are "specialized knowledge" subject to *Kumho Tire* and Fed. R. Evid. 702).

A. Horizontal Gaze Nystagmus Test

"Horizontal gaze nystagmus" (HGN) is the involuntary jerking of the eye that occurs

naturally when the eyes move from side to side. NHTSA Manual at VIII-12. The onset of HGN can occur earlier in the field of vision as a result of alcohol or other central nervous system depressants or illnesses. In the HGN test, a police officer instructs the subject to follow a moving object (such as a penlight) with their eyes from left to right. If the subject's eyes "jerk" prior to 45 degrees or "lack smooth pursuit," meaning that they do not follow the object smoothly, the officer may infer that alcohol or some other cause is affecting the subject's HGN. Due to the highly technical characteristics of HGN, officers require specialized training in its administration and interpretation. *See* NHTSA Manual at VIII-12-18.

Even before *Kumho Tire*, the majority of courts to consider the issue held that the HGN test is a scientific test that requires an evidentiary foundation. *See, e.g., State v. Witte*, 251 Kan. 313, 320, 836 P.2d 1110, 1114 (1992) (listing cases) (Ex. F); *see also Schultz v. State*, 106 Md. App. 145, 664 A.2d 60 (1995) (following 17 other states courts in holding that HGN is scientific test requiring foundation). Accordingly, and because of the highly specialized nature of the test, HGN should be assessed under the *Daubert/Kumho*/Rule 702 requirements for reliability and usefulness.

B. Walk and Turn and One Leg Stand Tests

In the late 1970s and early 1980s, NHTSA commissioned three studies which identified HGN, the walk-and-turn test (WAT) and the one-leg-stand (OLS) as the three most reliable FSTs. NHTSA Manual at VIII-1-7. WAT and OLS require a subject to perform several unfamiliar physical tasks – walking heel to toe or standing with one leg elevated – while listening to instructions, counting, or otherwise performing divided attention tasks. The administering officer is supposed to watch for certain predetermined technical "clues" such as missing a heel to

toe, or lowering a leg. Not every mistake is considered a clue, however, and the officer is trained to count only those clues identified through testing and validation. If a person misses more than two "clues" on either test, the officer may consider that as "evidence" that the subject's blood alcohol level (BAC) exceeds .10. NHTSA Manual at VIII-7-11.

The validity of the WAT and the OLS rests on the theory that alcohol impairs a person's motor skills and their ability to perform divided attention tasks. In its training manual, NHTSA emphasizes numerous times that the WAT and OLS must be performed only under certain conditions, i.e., on "a dry, hard, level, unslippery surface," NHTSA Manual at VIII-21, and interpreted only based on the predetermined "clues" in order to retain the validity that NHTSA assigns to them. NHTSA Manual at VIII-8, 9, 10, 11, 12. As NHTSA explains:

> [I]t is also necessary to emphasize one final and major point. This validation applies <u>ONLY</u> WHEN THE TESTS ARE ADMINISTERED IN THE PRESCRIBED, STANDARDIZED MANNER; AND <u>ONLY</u> WHEN THE STANDARDIZED CLUES ARE USED TO ASSESS THE SUSPECT'S PERFORMANCE; AND <u>ONLY</u> WHEN THE STANDARDIZED CRITERIA ARE EMPLOYED TO INTERPRET THAT PERFORMANCE.
>
> IF ANY ONE OF THE STANDARDIZED FIELD SOBRIETY TEST ELEMENTS IS CHANGED, THE VALIDITY IS COMPROMISED.

NHTSA Manual at VIII-12 (all capitalization and emphases in original).

NHTSA did not include other FSTs – e.g., finger count, touching finger to nose – as part of its approved battery of SFSTs, finding that these non-standardized FSTs did not contribute to the determination of intoxication. *See* NHTSA Manual at VIII-2-3. Accordingly, there are no standardized procedures for administering or interpreting other FSTs, even though Officer Jarrell administered them to Mr. Horn in this case.

Some pre-*Kumho* cases treats the WAT and the OLS as non-technical field observations

that can be admitted into evidence with no scientific foundation or other form of objective validation. *See, e.g.*, *United States v. Everett*, 972 F. Supp. 1313, 1320 (D. Nev. 1997) (FSTs are technical, non-scientific observations that do not require *Daubert* analysis); *Crampton v. State*, 71 Md. App. 375, 387-88, 525 A.2d 1087, 1093-94 (1987) (one-leg-stand, walk-and-turn, and reciting alphabet tests require no scientific foundation). After *Kumho Tire*, however, the tests must be subjected to the *Daubert* analysis in federal court, both under the reasoning of *Kumho* itself and under the newly amended Fed. R. Evid. 702. Field sobriety tests are technical: they rest on technical theories of human physical and neurological response to alcohol. They are specialized: officers not only must receive training in order to perform and evaluate the tests, but the NHTSA Manual emphasizes that where FSTs are not performed in accord with this training and specifications, they lose their validity. *See* NHTSA Manual at VIII-12. NHTSA itself rejected other, non-standardized FSTs because they did not contribute significantly to the intoxication inquiry. *See* NHTSA Manual at VIII-2-3.

Finally, the significance of FSTs is not readily transparent to the average layperson. An average juror or even judge will not know whether an FST was properly administered or whether its validity has been compromised by adverse field conditions. An average factfinder will not know the significance of missing a "clue," or even what a "clue" might be. Rather, like all specialized areas of expertise, the testifying officer must demonstrate his or her expertise and training in the area and explain the bases of the tests, their administration and their results in the particular case, in order for them to be legally meaningful. For all these reasons, WAT and OLS should, like HGN, be subject to the *Daubert*/Rule 702 analysis. *See Volk v. United States*, 57 F. Supp.2d at 894 n.3.

In addition, FSTs should be subject to *Daubert* and Rule 702 to ensure that testifying police officers properly substantiate their expertise and training in the area and demonstrate whether they properly administered the tests . Rule 702 instructs that "technical" or "specialized" evidence is admissible only if "the witness has applied the principles and methods reliably to the facts of the case." FSTs are a paradigmatic example of a specialized test whose validity depends heavily on the method of its application. *See* NHTSA Manual at VIII-12. Yet police officers receive a wide and unpredictable range of training in the administration of FSTs. Some may have extensive NHTSA-sponsored training, while others may have merely sat through a brief seminar put on by the local police force. As a result, different officers may administer and interpret the tests in different ways even while using the same language to describe the process and result. The standards laid down by *Daubert* and Rule 702 will ensure that such variations are properly addressed.

II. FST METHODOLOGY IS UNRELIABLE

A. The Legal Standard

Under *Kumho Tire*, specialized and technical knowledge as well as more traditional "scientific" knowledge is subject to the rigors of the *Daubert* analysis. Likewise, Rule 702 now provides:

> If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.[1]

---

[1] The amended version of Rule 702 went into effect December 1, 2000. The defendant assumes that the amended rule governs the resolution of this motion even though the

The rule's use of the conjunctive "and" indicates that all three factors – sufficient data, reliable principles and methodology, and reliable application – must be present. The burden of production lies with the party proffering the expert evidence, in this case the government, to provide the court with a factual basis from which it could conclude that the expert testimony is reliable. *Maryland Casualty Co. v. Thermo-Disc, Inc.*, 137 F.3d 780, 783 (4th Cir. 1997).

As the *Daubert* Court explained, the new test makes "gatekeepers" of federal judges, who must independently assess the factual basis, scientific validity, and application of technical methodologies to ensure that only reliable information is introduced into evidence. In this brave new evidentiary world, technical evidence is not admissible simply because it has been admitted by courts or used by experts in the past. Rather, courts must reassess such factors as whether the method is susceptible of testing, whether it has been the subject of peer review, whether it has an acceptable margin of error, whether it has gained general acceptance, and whether the method has legitimate uses outside of litigation. *See Samuel v. Ford Motor Co.*, 96 F. Supp.2d 491, 493 (D. Md. 2000) (enumerating non-exclusive list of factors that court should consider). If, in light of these and other factors that the court deems relevant, the information is found to be reliable as well as helpful, then and only then may it be admitted. The fact that courts may have admitted the data as evidence in previous cases is irrelevant; the court must assess the information afresh. Likewise, the fact that the information is widely used in law enforcement is also irrelevant; as this Court recently commented, "[t]he fact that an entire industry may use a test of insufficient reliability does not make it admissible into evidence." *Samuel*, 96 F. Supp.2d at 500.

---

conduct arose before the promulgation of the rule. *See Landgraf v. USI Film*, 511 U.S. 244, 275 (1994) ("Changes in procedural rules may often be applied in suits arising before their enactment without raising concerns about retroactivity.").

B. NHTSA's Studies Do Not Establish FST Reliability

The primary source for information about the validity of FSTs comes from NHTSA, the government agency charged with improving traffic safety. *See* 49 U.S.C. § 30101. With respect to drunk driving, NHTSA functions not as an independent scientific research institution but as a species of law enforcement agency. Its "research objectives" in conducting its three FST studies were to create law enforcement tools, namely, "to complete the development and validation of the sobriety test battery," NHTSA Manual at VIII-3, and "to develop standardized, practical and effective procedures for police officers to use in reaching arrest/no arrest decisions." *Id.* at VIII-6. As the Ninth Circuit has explained, "[o]ne very significant fact to be considered [under *Daubert*] is whether the experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for the purposes of testifying." *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 43 F.3d 1311, 1316 (9th Cir. 1995).[2] NHTSA's research has been developed primarily for the purpose of arrest and prosecution of intoxicated drivers, and police acquire expertise in FSTs expressly for those purposes. Accordingly, NHTSA's research does not deserve the weight of an independent scientific research agenda, and NHTSA's scientific conclusions about FSTs should be appropriately discounted to account for its mandate.

Even if NHTSA's conclusions about FSTs are taken at face value, NHTSA itself has

---

[2] The Ninth Circuit has distinguished certain law enforcement tools – DNA analysis, fingerprint and voice recognition – as scientific tools that "have the courtroom as the principle theater of operations" but are nevertheless reliable, *Daubert,* 43 F.3d at 1317 n.5. But FSTs have inherent flaws that DNA, fingerprint and voice recognition lack. FSTs depend on the subjective perceptions of the arresting police officer at the moment of arrest, rather than an independent expert with no stake in the outcome. Moreover, unlike other law enforcement tests, FST results cannot be checked or duplicated after the fact.

documented FST unreliability and their large margins of error. According to the three NHTSA studies, when administered in perfect accordance with the standardized conditions and procedures, HGN is 77 percent accurate, walk-and-turn is 68 percent accurate, and one-leg-stand is 65 percent accurate. NHTSA Manual at VIII-11. By "accurate," NHTSA means that the test leads police officers to correctly classify a subject as having a BAC above or below .10. NHTSA Manual at VIII-5. Thus a police officer using HGN will wrongly estimate a person's BAC 23 percent of the time; with WAT he will be wrong 32 percent of the time; and with OLS he will be wrong 35 percent of the time. Using HGN and WAT together, he will be wrong 20 percent of the time. NHTSA Manual at VIII-11. The studies tell us nothing whatsoever about FST accuracy at BAC levels below .10.[3] Error rates between 20-35 percent far exceed error rates found acceptable in other areas of scientific evidence. *See, e.g.,United States v. Chischilly*, 30 F.3d 1144, 1154 (9th Cir. 1994) (finding DNA error rates of 1-4 percent acceptable); *United States v. Galbreth*, 908 F. Supp. 877, 891 (D. N.M. 1995) (admitting polygraph evidence where error rate found to be 5-10 percent).

C. Dr. Spurgeon Cole

NHTSA's error rates, large as they are, underestimate the true error rates of FSTs. According to Dr. Spurgeon Cole, NHTSA's studies were conducted under flawed conditions without proper scientific controls, and tend to inflate the apparent reliability of the FSTs. *See* Cole, S. & Nowaczyk, R., *Field Sobriety Tests: Are They Designed For Failure*? Percept. & Motor Skills 99-104 (1994), Ex. C. In the 1977 NHTSA study, 47 percent of the subjects were

---

[3] The non-standardized FSTs are completely undocumented and therefore lack any validity or measurable error rate at all.

wrongly identified. *Id.* at 100. The 1981 study found reliability coefficients that were "below accepted levels for standardized clinical tests" among officers, meaning that officers came to inconsistent conclusions when assessing the same subjects. *Id.* The 1983 study was based on after-the-fact analysis of police stops and arrests of DWI suspects, *i.e.,* people already suspected of being intoxicated. The FSTs were thus being tested on a sample group of subjects who were more likely than the average person to be intoxicated, which would in turn make the FSTs appear more accurate than they really are. Dr. Cole also identified "lack of standardization across many of the field sobriety test studies" as a further source of concern. *Id.*

Even more disturbingly, Dr. Cole's independent research produced startlingly different results. Under controlled conditions, police officers were told to assess whether subjects were intoxicated based on their performance on the walk-and-turn and the one-leg-stand FSTs. Although *none* of the subjects had any alcohol, *forty-six percent* of the officers' decisions were that the person "had too much to drink to drive." *Id.* at 102. Dr. Cole hypothesizes that subjects often miss one or more clues merely because they are unfamiliar with the tests, and that FSTs lead police to conclude that subjects are impaired even when they are not. *Id.* at 102-03. Dr. Cole concludes that "[t]his study brings the validity of field sobriety tests into question," *id.* at 103, a conclusion consistent with NHTSA's own findings that a significant percentage of police assessments based on FSTs are incorrect.

D. Judicial Assessments of HGN Unreliability

Several courts have questioned HGN's reliability. In an exhaustive analysis and citing numerous scientific studies, the Kansas Supreme Court concluded that "[t]he reliability of the HGN test is not currently a settled proposition in the scientific community." *State v. Witte*, 251

Kan. at 329, 836 P.2d at 1120; *see also People v. Leahy*, 882 P.2d 321 (Cal. 1994) (following *Witte*). The Kansas Court noted that HGN has many causes other than alcohol:

> Nystagmus can be caused by problems in an individual's inner ear labyrinth. . . . Physiological problems such as certain kinds of diseases may also result in gaze nystagmus. Influenza, streptococcus infections, vertigo, measles, syphilis, arteriosclerosis, muscular dystrophy, multiple sclerosis, Korsakoff's Syndrome, brain hemorrhage, epilepsy, and other psychogenic disorders all have been shown to cause nystagmus. Furthermore, conditions such as hypertension, motion sickness, sunstroke, eyestrain, eye muscle fatigue, glaucoma, and changes in atmospheric pressure may result in gaze nystagmus. The consumption of common substances such as caffeine, nicotine, or aspirin also lead to nystagmus almost identical to that caused by alcohol consumption. Temporary nystagmus can occur when lighting conditions are poor. An individual's circadian rhythms (biorhythms) can affect nystagmus readings – the body reacts differently to alcohol at different times of the day.

State v. Witte, 251 Kan. at 326, 836 P.2d at 1120 (internal citations omitted). The Court also worried that HGN is unreliable in practice because of the difficulty in estimating a 45 degree angle. *Id*. at 328, 836 P.2d at 1120 ("A visual estimation of the angle would seem to cause inaccurate and inconsistent results."). The Court concluded that the government had not met its burden of establishing HGN reliability.

The scholarly literature likewise reflects the unreliability of HGN. One commentator concludes that NHTSA's claims for HGN reliability "are not supported by field study. . . . No study establishes the accuracy, margin of error, or reliability of trained police officers performing the roadside HGN test. Officers in the field have not shown that they can correctly classify those individuals with actual BACs in the critical range (0.05%-0.15% BAC)." Joseph Meaney, *Horizontal Gaze Nystagmus: A Closer Look*, 36 Jurimetrics J. 383, 398 (1996), Ex. D. Meaney further concludes that "HGN's potential rate of error is unknown," "peer review of SCRI's HGN work is limited," and "NHTSA['s] . . . claims exceed the available data." *Id*. at 401.

E. Dr. Yale Caplan

Dr. Yale Caplan, former Chief Toxicologist for the State of Maryland and former Scientific Director of the Maryland Alcohol Testing Program, while more sanguine about FST reliability than Dr. Cole, has serious reservations about their reliability when used as evidence of alcohol intoxication. Based on 30 years of experience in the field, Dr. Caplan concludes that "field sobriety tests alone were never designed for or demonstrated to be unequivocally capable of indicating alcohol impairment." Affidavit of Dr. Yale Caplan ("Caplan Aff.") at 1., Ex. E. Rather, FSTs at best are capable of indicating "physiological impairment [which] can be the result of alcohol, drugs, or medical conditions." Caplan Aff. at 1. If FSTs suggest the presence of any impairment, "the causative factor needs to be further identified by subsequent tests for the presence of alcohol, drugs, or impairing medical conditions." *Id.* "Field sobriety tests alone can not be used to establish alcohol impairment with absolute certainty." *Id.* at 2.

In other words, according to the State of Maryland's own professional expert, SFSTs by themselves have no bearing on whether a person is intoxicated. Rather, they should only be performed in association with a chemical breathalyzer test to determine the cause of any possible impairment.[4]

F. Applying the *Daubert* Factors

While there is no *per se* acceptable error rate under *Daubert*, courts have admitted scientific evidence with error rates of between 1-5, or 5-10 percent, *see United States v.*

---

[4] Based on Dr. Caplan's expert testimony alone, a court should find as a matter of law that where evidence of intoxication consists only of field sobriety tests unconfirmed by any chemical analysis, there is insufficient evidence to convict a person of driving under the influence of alcohol.

Chischilly, 30 F.3d at 1154 (DNA); *United States v. Galbreth*, 908 F. Supp. at 891 (polygraph), while excluding methodologies that had 50 percent or higher or indeterminate error rates. *See Flores v. Johnson*, 210 F.3d 456, 465 (5th Cir. 2000) (excluding expert testimony predicting "future dangerousness" which had error rate of fifty percent or higher). NHTSA's research indicates that under perfect conditions SFST margins of error range from 23 percent to as high as 35 percent. Dr. Cole's research indicates that it is closer to 50 percent, *i.e.*, approximately as accurate as flipping a coin. And even NHTSA acknowledges that under adverse field conditions where standardized procedures and conditions are not followed, the error rate will be higher still. While HGN appears to be the most reliable of the tests, HGN's reliability has been seriously questioned by at least two state supreme courts as well as numerous scientists and scholars.

With respect to peer review, only the HGN data appears to have been peer reviewed at all. Reliability data on the WAT and OLS come exclusively from NHTSA, *i.e.*, a law enforcement agency charged with deploying FSTs to reduce drunk driving. Such instrumental validation hardly constitutes the sort of rigorous intellectual crucible contemplated by *Daubert*. FSTs are theoretically susceptible of testing but the divided literature makes clear that they have yet to be adequately tested.

Finally, HGN has not been generally accepted in the relevant scientific community, *see State v. Witte*, 251 Kan. at 329, 836 P.2d at 1120; *see also People v. Leahy*, 882 P.2d 321 (Cal. 1994), and there is so little research on WAT and OLS that it can hardly be said whether anyone accepts them at all.[5] In sum, because the errors rates for FSTs are either undetermined or

---

[5] The non-standardized FSTs administered to Mr. Horn – finger dexterity and alphabet recital – are likewise completely undocumented and therefore have no determined level of reliability at all.

unacceptably high, because there is little if any peer review or testing, and because there is no consensus within any scientific community as to their validity, none of the three SFSTs meet the rigorous standards of *Daubert* or Rule 702.[6]

The *Daubert* factors are non-exclusive and do not limit the Court's ability to assess FST reliability based on additional criteria. FSTs should therefore be considered, not merely from the perspective of scientific testing and validation, but common sense. FSTs are intuitively suspect. Not only are they flat out wrong much of the time, but they are administered under highly suspect circumstances, namely, by a law-enforcement officer whose job is not only to arrest suspected drunk drivers but eventually to defend those arrest decisions in court using the very FST evidence at issue. FTS results are usually witnessed only by the officer: a subject performing FSTs cannot easily contradict an officer's testimony that he or she missed a particular clue because the person cannot observe his or her own performance. FST results, moreover, are as ephemeral as the blood alcohol they purport to measure: they cannot be replicated or verified afterwards. FSTs are also extremely easy to fail. As Dr. Cole has pointed out, a person could easily miss two or more clues due to nervousness, unfamiliarity with the procedures, or simply because they are standing by the side of the road in the dark. *See* Cole at 103.

Finally, the question remains whether FSTs are valid at all, *i.e.*, whether they actually measure anything relevant to the ultimate issue in DWI case. The fact that a person twice fails to place his heel precisely in front of his toe, given eighteen opportunities to do so, hardly

---

[6] Rule 702 also demands a showing that "the testimony is based upon sufficient facts or data . . .[and that] . . . the witness has applied the principles and methods reliably to the facts of the case." These inquiries are specific to the facts of the case and will require an evidentiary hearing and testimony from the individual officer.

constitutes compelling evidence of anything. As one court has puzzled, "it is difficult to imagine how defendant's performance on the walk and turn test could have bolstered a finding of intoxication." *Volk*, 57 F. Supp.2d at 895-96. Rather, as Dr. Caplan explains, at best, FSTs might indicate impairment from some unknown source, and Dr. Cole's research indicates that FSTs may not even do that. In sum, FSTs may well be legally irrelevant.

For all these reasons, FSTs should be found unreliable as a matter of law.

III. FST ARE SUITABLE ONLY FOR PROBABLE CAUSE DETERMINATIONS AND NOT AS SUBSTANTIVE EVIDENCE

The reason SFTSs do not meet *Daubert*'s rigorous standards is that they were never meant to. As Dr. Caplan explains, "field sobriety tests alone were never designed for or demonstrated to be unequivocally capable of indicating alcohol impairment." Caplan Aff. at 1. Rather, SFSTs were designed as tools to assist in arrest decisions, *i.e.*, probable cause determinations. NHTSA Manual at VIII-6. SFSTs thus strongly resemble portable breathalyzer tests (PBTs), which are small-scale breathalyzer machines that police use roadside to assess whether probable cause to arrest exists. Courts have unanimously held that the results of PBTs are inadmissible as substantive evidence because they are scientifically unreliable. *See United States v. Iron Cloud*, 171 F.3d 587, 591 & n.5 (8th Cir. 1999) (holding PBT results inadmissible and listing state case decisions holding same). In the same vein, "[p]olygraph examinations widely are accepted and used in employment, law enforcement and security contexts, yet this fact does not make them admissible as evidence in trials." *Samuels*, 96 F. Supp.2d at 500.

The same reasoning applies to the three SFSTs administered to Mr. Horn. While such tests may provide probable cause, they do not meet the much higher standards of reliability

required for admissibility as substantive evidence. The transformation of the SFST from a quick roadside probable cause determination into evidence in a federal criminal trial is thus unwarranted.[7]

IV. EVEN IF THE WALK-AND-TURN AND ONE-LEG-STAND TESTS ARE NON-TECHNICAL EVIDENCE, THEY ARE UNRELIABLE, UNHELPFUL, AND HIGHLY PREJUDICIAL, AND SHOULD BE EXCLUDED UNDER RULES 701 AND 403.

Rule 701 provides:

> If a witness is not testifying as an expert, the witness's testimony in the form of opinions or inferences is limited to those opinions or inferences which are (a) rationally based on the perception of the witness and (b) helpful to a clear understanding of the witness' testimony or the determination of a fact in issue.

Rule 403 states that "evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading of the jury . . . ." Fed. R. Evid. 403. If the Court were to decide to treat one or more of the FSTs as non-scientific evidence not governed by *Daubert*, they should nevertheless be excluded because they are unreliable and therefore unhelpful in determining the ultimate issue of intoxication, and because the probative value of officer testimony regarding FSTs is substantially outweighed by the danger that it will prejudice the defendant.

The above discussion demonstrates that SFSTs are unreliable. Dr. Caplan has explained that SFSTs do not measure alcohol intoxication at all but, at best, merely indicate whether someone may be impaired for some reason. Between their unreliability and their tenuous relation to the issue of whether a person has been drinking, FSTs are not helpful to the determination of

---

[7] Mr. Horn does not concede that the non-standard FSTS administered to him would contribute to or establish probable cause.

whether a person is under the influence of alcohol.[8]

Moreover, police officer testimony regarding the administration and results of FTSs is highly prejudicial. Expert testimony by law enforcement has "an aura of special reliability and trustworthiness," *United States v. Webb*, 115 F.3d 711, 721 (9th Cir. 1997), and the technical and conclusory language of FSTs – "pass," "fail," "impaired," "missed clues" – gives them an appearance of authority that could unduly sway a jury. On balance, the substantial danger of prejudice outweighs the limited evidentiary value of the FST evidence.

CONCLUSION

For the above reasons, Mr. Horn moves that the all evidence of the field sobriety tests administered to him be excluded.

Respectfully submitted,

JAMES WYDA
Federal Public Defender
for the District of Maryland

_____
SASHA NATAPOFF
Assistant Federal Public Defender
100 S. Charles Street
Tower II, Suite 1100
Baltimore, Maryland 21201
(410) 962-3962

[8] This argument applies with even more force to the non-standardized, untested, unvalidated FSTs administered to Mr. Horn.

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that on this ___ day of February, 2001, a copy of the foregoing Memorandum of Law in Support of Defendant's Motion in Limine to Exclude the Government's Field Sobriety Test Evidence was delivered to Paul Marone, Special Assistant United States Attorney, U.S. Army Garrison, Building 310, Wing 10, Aberdeen Proving Ground, Maryland, 21001.

_____
Sasha Natapoff
Assistant Federal Public Defender

Exhibit A
Daubert/Kumho Worksheet

1. Name of Expert Challenged: Officer Daniel Jarrell

2. Brief summary of opinion(s) challenged (if more than one, designate separately), including reference to the source of the opinion (i.e., Rule 26(a)(2)(B) disclosure, deposition transcript references, interrogatory answers). Attach highlighted copy of source materials as exhibit:

   Officer Jarrell performed three sobriety tests on Mr. Horn and concluded that he was intoxicated. *See* Ex. G (police report and Alcohol Influence Report)

3. Briefly describe methodology/reasoning used by expert to reach each opinion which is challenged. Include reference to source of challenged methodology/reasoning, and attach a highlighted copy as an exhibit:

   Upon information and belief, Officer Jarrell relied on the methodology of field sobriety testing contained in the NHTSA training manual, attached as Ex. B.

4. Briefly explain the basis for the challenge to the reasoning/methodology used by the expert (for example, methodology unreliable; methodology reliable, but not valid for application to this case; failure to use standardized or accepted methodology (for example, with a standardized test); etc.) Attach a highlighted copy of affidavit or other source material supporting challenge to methodology/reasoning as an exhibit:

   a. Mr. Horn challenges the underlying methodology of FSTs as unreliable indicators of alcohol intoxication. *Source materials*: Dr. Spurgeon Cole (Ex. C); Dr. Yale Caplan (Ex. E); Jurimetrics article (Ex. D); case law (Ex. F).

   b. Mr. Horn also challenges Officer Jarrell's application of the methodology in this case and contends that Officer Jarrell failed to apply even the standardized or accepted FST methodology in this case. *See* Ex. G (police report).

5. Is the challenged methodology/reasoning subject to a known or potential error rate? If so, briefly describe it, and attach a highlighted copy of any relevant source material as an exhibit:

   The error rates for FSTs are unknown.

6. Summarize relevant peer review materials relating to methodology/reasoning challenged, and attach a highlighted copy of any relevant source material as an exhibit:

   a. With respect to the walk and turn and one leg stand FSTs, Dr. Cole has published a peer reviewed article (Ex. C) explaining that most of the FST literature is not peer reviewed but rather consists of government-sponsored reports. Cole's research indicates that the WAT and OLS tests are unreliable indicators of impairment and that police officers consistently misidentify subjects as impaired when they are not using those tests.

   b. In *Horizontal Gaze Nystagmus: A Closer Look*, 36 Jurimetrics J. 383 (1996) (Ex. D), Joseph Meaney argues that "HGN's potential rate of error is unknown," "peer review of [the central HGN study] is limited," and "NHTSA['s] . . . claims exceed the available data."

7. If the challenge to the opinion is based upon a contention that the methodology/reasoning has not been generally accepted within the relevant scientific or technical community, briefly explain the basis for this contention. Attach highlighted copy of any relevant supporting materials as an exhibit:

   Most of the research on FSTs come from NHTSA: there has been some independent research on HGN and none on walk-and-turn and one-leg-stand. Two state supreme courts as well as numerous scholars conclude that HGN is not generally accepted in the scientific community. *See State v. Witte*, 251 Kan. 313, 320, 836 P.2d 1110, 1114 (1992) (Ex. F) (describing cases and scholarship). Dr. Cole summarizes the literature on walk and turn and one leg stand and concludes that there is no general acceptance. See Ex. C.

Ex. B
NHTSA Manual

Ex. C
Cole article

Ex. D
Jurimetrics article

Ex. E
Caplan affidavit and resume

Ex. F
State v. Witte

Ex. G
police report, alcohol influence report

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MARYLAND

	*	
UNITED STATES OF AMERICA	*	
v.	*	**Case No. 00-946PWG**
ERIC D. HORN	*	

* * * * *

MOTION FOR DISCOVERY OF GOVERNMENT EXPERT WITNESSES UNDER FED. R. CRIM. P. 16(a)(1)(a)

Defendant Eric Horn, by and through counsel, James Wyda, Federal Public Defender for the District of Maryland, and Sasha Natapoff, Assistant Federal Public Defender, moves for discovery of any and all expert witnesses that the government intends to call at trial and/or at the *Daubert* hearing requested by the defendant by motion filed this same day. In particular, Mr. Horn seeks discovery including but not limited to materials regarding any police officer who the government intends to call as an expert witness. In support of his motion Mr. Horn alleges the following:

1. According to police reports, on June 28, 2000, Mr. Horn was stopped by Officer Daniel Jarrell at the Harford Gate of Aberdeen Proving Ground. Officer Jarrell performed several so-called field sobriety tests, or FSTs, on Mr. Horn. Mr Horn was subsequently charged with driving under the influence of alcohol in violation of Md. Code Ann., Transp. § 21-902.
2. The defense expects the government to call Officer Jarrell as an expert witness at trial to testify about the administration of the field sobriety tests, the results he observed, and his

opinion as to whether Mr. Horn was intoxicated.

3. In addition, the defendant has moved in limine for a *Daubert* hearing to address the scientific reliability, relevance, and admissibility of the field sobriety tests administered to Mr. Horn. The government may seek to call Officer Jarrell, as well as other experts, to testify at that hearing.

4. Under Rule 16(a)(1)(E), Fed. R. Crim. P., a defendant is entitled to a written summary of expert testimony that the government intends to use under Rule 702, 703 or 705, including a description of the witnesses' opinions, the bases and reasons therefore, and the witnesses' qualifications. A defendant is entitled to discovery at any stage of the proceeding, including pretrial motions.

5. Accordingly, Mr. Horn is entitled to and requests all discovery related to Officer Jarrell and any other expert witnesses that the government intends to call at trial or at the *Daubert* hearing. In particular, Mr. Horn requests documentation of all police officer witnesses', including Officer Jarrell's, qualifications to administer and interpret FSTs, their training in field sobriety tests, copies of any manuals used in their training, descriptions of any and all courses taken by them, all training materials and/or manuals used in that training, the names of any and all instructors who provided that training, any evaluations of or scores given to Officer Jarrell or other officers in the course of that training, and any and all policy statements, protocols, manuals, or any other materials issued by or relied on by the military police department by which Officer Jarrell or other officers are employed that address training requirements related to FSTs.

Respectfully submitted,

JAMES WYDA
Federal Public Defender
for the District of Maryland

_____
SASHA NATAPOFF
Assistant Federal Public Defender
100 S. Charles Street
Tower II, Suite 1100
Baltimore, Maryland 21201
(410) 962-3962

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that on this ___ day of February, 2001, a copy of the foregoing Motion for Discovery of Government Expert Witnesses was delivered to Paul Marone, Special Assistant United States Attorney, U.S. Army Garrison, Building 310, Wing 10, Aberdeen Proving Ground, Maryland, 21001.

_____
Sasha Natapoff
Assistant Federal Public Defender

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MARYLAND

	*	
UNITED STATES OF AMERICA	*	
v.	*	**Case No. 00-946PWG**
ERIC D. HORN	*	

* * * * *

REPLY TO GOVERNMENT'S
RESPONSE TO DEFENDANT'S MOTION IN LIMINE

The government's studies and expert opinions demonstrate at best that standardized field sobriety tests (SFSTs) are useful law enforcement tools for establishing probable cause to arrest a driver. The government's submission does not establish the much more difficult proposition: whether SFSTs meet the rigorous standards of Daubert and qualify as valid, admissible evidence in a criminal trial. Indeed, with respect to the one-leg-stand (OLS) and the walk-and-turn (WAT) tests, the government offers not a single peer reviewed article or independent scientific expert. Generally speaking, the government submissions show at best that the SFST battery is approximately as reliable as a portable breath test (PBT), which is itself unreliable and inadmissible. Since the government has failed to establish the scientific validity of its own evidence, the results of the SFSTs should be excluded.

I. THE GOVERNMENT'S AND DEFENDANT'S RESPONSES

In its response to defendant's motion, the government proffers a resource guide, two affidavits, and five studies. The guide, "Horizontal Gaze Nystagmus: The Science and the Law," is a resource guide compiled by a law enforcement advocacy organization for use by prosecutors

and police. One affidavit is from Lieutenant Colonel Jeff Rabin, an Army optometrist, who opines that there is a "very good correlation between the results of the horizontal gaze nystagmus and breath analysis for intoxication." The second affidavit is from Detective Daniel L. Jarrell, the arresting officer in the instant case whose expert testimony is the subject of defendant's challenge. His affidavit chronicles his training in SFST administration, and his administration of the tests to Mr. Horn in the instant case. Finally, the five studies are validation studies sponsored by NHTSA and/or other government transportation agencies, designed to establish the validity of SFSTs for use by law enforcement.

In response to the government's submission, defendant offers the opinions and analyses of Dr. Spurgeon Cole, Mr. Harold Brull, and Dr. Joel Wiesen, as well as a peer-reviewed study by Dr. James L. Booker. Cole, Brull and Wiesen each independently reviewed the scientific basis for the SFSTs offered by the government. Their conclusions are attached in the form of affidavits and/or published articles.

Dr. Spurgeon Cole is Professor Emeritus of Psychology at Clemson University. He holds a Ph.D. in clinical psychology. He has published numerous peer reviewed articles in the field of behavioral psychology and testing, including Cole, S. & Nowaczyk, R., Field Sobriety Tests: Are They Designed For Failure? Percept. & Motor Skills 99-104 (1994) (hereinafter "Cole Study 1994") (attached as Ex. 2), and Nowaczyk & Cole, Separating Myth from Fact: A Review of Research on the Field Sobriety Tests, in HANDLING TRAFFIC CASES IN SOUTH CAROLINA, Ch. 33 (1994) (hereinafter "Cole Research Review 1994"); see also Cole Resumé, Ex. 3. He concludes that although the NHTSA laboratory tests were well conducted, their results indicate high SFST unreliability, and that the field studies were improperly conducted, misleading, and inconclusive.

Dr. Cole's own independent peer-reviewed research indicates that the WAT and OLS are unreliable.

Mr. Brull is an expert in the design and evaluation of human behavior and performance tests. He is Senior Vice President of Personnel Decisions International (PDI), one of the world's largest industrial psychology consulting organizations which specializes in the measurement and testing of human attributes, particularly in the employment setting. Mr. Brull has designed and evaluated thousands of human behavior/performance tests. He has also worked with over 1,000 law enforcement agencies in the area of performance testing. He has a masters degree in educational psychology, a bachelors degree in biochemistry, and he has taught at Cornell University, the University of Minnesota, St. Olaf College, and the Southern Police Institute. See Brull Aff., Ex. 4; Brull Resumé, Ex. 5. Mr. Brull concludes that the two NHTSA laboratory tests indicate potential usefulness for the SFST battery but that they are highly unreliable in certain areas, that the field studies are incomplete and inconclusive, and that overall the studies are scientifically unreliable.

Dr. Joel Wiesen is an industrial psychologist specializing in the development and evaluation of human behavioral tests. He holds a Ph.D. in psychology, and is a published test author, having developed a test of mechanical aptitude which is now used nationwide. He is currently an independent consultant in the area of human performance test development and validation: past and current clients include Bell Atlantic, T.J. Maxx, Maryland, Massachusetts, Pennsylvania, and Virginia. See Wiesen Aff., Ex. 6; Wiesen Resumé, Ex. 7. He concludes that the lab studies, although flawed, were overall well-designed, that their results indicate reliability problems with FSTs, that the field studies are inadequate and do not meet the standards of the

professional testing community, and that overall the SFSTs do not meet the reliability standards of the professional testing community.

Dr. James L. Booker is a forensic scientist. He holds a Ph.D. in chemistry, and has worked in law enforcement as well as the private sector and in the academy. See Booker Resumé, Ex. 9. He has published numerous articles in the areas of scientific testing methodology. His study, End-position nystagmus as an Indicator of Ethanol Intoxication, 41 Science & Justice 113 (2001) (hereinafter "Booker Study"), is published in the peer-reviewed journal issued by one of the largest forensic science organizations in the world. See Ex. 8. Based on independent experimentation, Booker's study concludes that end-point nystagmus is present in over fifty-percent of non-drinking subjects, that there is a strong correlation between nystagmus and fatigue, that the vast majority of police officers do not administer the HGN test properly, and that therefore the HGN test is not a reliable indicator of alcohol intoxication. See Booker Study, Ex. 8 .

II THE LEGAL STANDARD

As discussed at length in defendant's original motion, the legal standard for admissibility for each field sobriety test is governed by Daubert. A court must consider, inter alia, whether the evidence is susceptible of testing, whether it has a known error rate, whether it has been subject to peer review, whether it is generally accepted by the relevant scientific community, and whether it has legitimate uses outside of litigation. See Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 593-94 (1993); United States v. Cordoba, 194 F.3d 1053 (9th Cir. 1999); Samuel v. Ford Motor Co., 96 F. Supp.2d 491, 493 (D. Md. 2000). The government

bears the burden of establishing reliability.

The ability to discern an actual error rate is particularly important. In Cordoba, the Ninth Circuit affirmed the exclusion of polygraph test results, noting that while the tests were subject to testing and peer review, test administration in practice varied widely, that laboratory-quality error rates were "not transferrable to real life exams," and that therefore "the error rate of real-life polygraph testing is not known and not particularly capable of analyzing." Cordoba, 194 F.3d at 1059. Similarly, this Court has excluded scientific evidence based in part on the lack of a "fit" between laboratory tests and real-life application of the principles. Samuel, 96 F. Supp.2d at 502.

III. THE GOVERNMENT'S SUBMISSIONS DO NOT ESTABLISH THE SCIENTIFIC VALIDITY OF THE STANDARDIZED FIELD SOBRIETY TESTS

At the outset, it should be noted that the government's submission does not even purport to establish a correlation between SFST results and driving impairment. With only one exception, the studies and affidavits attempt to correlate SFST results with BAC, i.e., the presence of alcohol in the blood.[9] While BAC levels of .08 and above are now per se illegal in Maryland,[10] the relationship between BAC and actual driving impairment is assumed, not shown.

---

[9] The 1977 study attempted a simple correlation experiment between FSTs and driving skills, using a crude apparatus designed to measure tracking, reaction time, and driving errors. Only tracking was found to correlate significantly with the FSTs, and no other studies purport to have established a definite relationship between FSTs and driving ability. 1977 Study at 51-57.

[10] Since the inception of this case, the Maryland legislature lowered the legal limit for the offense of driving under the influence of alcohol per se from .10 to .08 BAC. Md. Code Ann., § 27-388A, Md. Code Ann., Transp. § 21-902(a)(2) (effective Sept. 30, 2001). The legislature also substituted the term "impaired by" for "under the influence" in Md. Code Ann.,

As Dr. Cole points out, "[i]n one of NHTSA's own reports, the following statement is made: "...even valid, behavioral tests are likely to be poor predictors either of actual behind-the-wheel driving . . . or of accidents." Cole, Research Review 1994 at 3, Ex. 1.

The statute and case law make it illegal to drive while "when an individual's normal judgment, perception, and/or coordination [is] adversely affected; that is, made worse to any extent by the consumption of an alcoholic beverage." United States v. Sauls, 981 F.Supp. 909, 918 (D. Md. 1997). The studies assume, without showing, that the presence of nystagmus, or a person's inability to take eighteen steps, heel to toe, in a straight line, or to hold one foot aloft for 30 seconds, correlates with a relevant impairment of judgment, perception, or coordination. But that is not a transparent proposition. Alcohol consumption may also impair a person's ability to knit, or perform mathematical calculations, but the burden remains on the government to show that those impairments correlate meaningfully with a person's driving ability. The government has not done so.

The HGN Resource Guide and Rabin affidavit assert that HGN correlates with the presence of some alcohol in the blood. See, e.g., Rabin Aff. at 2 ("[A]lcohol consumption affects smooth pursuit movements and triggers nystagmoid movements at blood alcohol levels of 0.03-0.04% . . . ."). Assuming its truth, this proposition does not establish the reliability of the police-administered HGN roadside test. Simply because there is a relationship between HGN and alcohol does not mean that the test reliably reveals the presence of an impairing level of alcohol.

Transp. § 21-902(b). For completeness, the reliability inquiry thus considers the .08 limit and the new terminology of "impairment" as well as the former law, although Mr. Horn's conduct is governed by the earlier .10 standard. See Lynce v. Mathis, 519 U.S. 433, 440-41 (1997) (defendant's conduct is governed by the law in effect at time conduct is committed).

The reliability of the test depends on its design and administration, which is addressed in the five studies. But the NHTSA-sponsored studies are not themselves reliable, procedures vary widely within the studies, and the government's submissions establish that small variations in technique and interpretation can dramatically alter results. The government's submission also fails to establish that HGN correlates with illegal impairment or intoxication, since HGN can occur at BAC levels below impairment levels, and does not vary with quantity of alcohol consumed. Rabin Aff. at 2.

In contrast to the government's non-peer reviewed submissions, the peer-reviewed Booker study reveals significant flaws in aspects of the HGN test which render it unreliable. The Booker study also found that officers almost never administer the test properly so as to obtain valid results. Booker Study at 116, Ex. 8.

The five NHTSA/DOT studies are the only evidence offered in support of the WAT and the OLS field sobriety tests and they are grossly inadequate to establish scientific reliability. None of them are peer reviewed. Margins of error are high -- when they can be discerned at all -- and vary across tests. Laboratory standards and procedures differ widely from those used in the field. Methodologies vary from test to test, or are entirely missing from the analysis. There is no analysis of the tests in relation to any established scientific standards or communities of expertise – the studies simply stand alone. Accordingly, they do not meet the Daubert standard.

A. The Horizontal Gaze Nystagmus Test

The HGN field sobriety test is made up of three components: smooth pursuit, nystagmus at maximum deviation, and angle of onset. There is significant scientific debate over whether

each of these three inquiries correlates reliably with impairing levels of alcohol. The government proffers the Rabin affidavit and the HGN Resource Guide as sources of scientific validation for the HGN test. Defendant provides the peer-reviewed study: End-position nystagmus as an indicator of ethanol intoxication, attached as Exhibit 8, in response. The NHTSA studies are discussed separately.

1. Rabin Affidavit

Lt. Col. Jeff Rabin is an Army optometrist who reviewed unnamed pieces of literature regarding HGN and its correlation with alcohol ingestion. He has no particular expertise in the design or administration of the HGN test under actual field conditions; his expertise is medical and general. Indeed, although he claims to have formally presented on the effects of alcohol on eye movements and testified as an expert on HGN, his resumé does not list any HGN or alcohol related publications or presentations.

Rabin admits that lack of smooth pursuit and nystagmus can occur at BAC levels as low at 0.03-0.04%. Rabin Aff. at 2. He concludes, nevertheless, based on a limited literature review, that "there is a very good correlation between the results of the horizontal gaze nystagmus test and breath analysis for intoxication." Rabin Aff. at 3. This conclusion is based in part on Rabin's own routine practice of administering nystagmus tests. Rabin Aff. at 1-2. He believes that no medical training is required to administer a nystagmus test, and surmises that "a police officer may be trained accurately to administer the horizontal gaze nystagmus test and to interpret test results." Rabin Aff. at 2, 3.

The Rabin affidavit thus stands for the proposition that there is a correlation between alcohol ingestion and nystagmus, sometimes at perfectly legal levels of BAC, and that in theory a

police officer with proper training could discern nystagmus. This is a far cry from establishing that the roadside HGN test, performed under widely varying conditions, administered by police officers with varying degrees of training, reliably indicates illegal intoxication. In particular, it tells us nothing about the error rate of the actual test. See Cordoba, 194 F.3d at 1059 (theoretical validity of properly conducted polygraph did not translate into real life exams).

2. HGN: Resource Guide

The HGN Resource Guide (hereinafter the "Guide") is a compilation of information for judges, prosecutors, and law enforcement. It adds nothing by way of independent validation or expertise to its sources. It is not peer reviewed. Its author, James J. Dietrich, is a staff attorney at the American Prosecutors Research Institute, who brings no more scientific or other expertise to bear on the matter of HGN reliability than undersigned counsel.

The Guide is an admittedly biased document. Its aim is not to explore, even-handedly, all aspects of the HGN reliability question, but rather to "short circuit the inaccurate and self-serving view of HGN that is propounded by defense counsel." Guide at 2. The Guide aims to help prosecutors "lay the foundation for the admissibility of the HGN test" and to "encourage judges to accept the results . . . ." Guide at 4. The Guide does not purport to present a balanced viewpoint or competing evidence and indeed, its scientific assertions are one-sided.

For example, the Guide asserts that the "NHTSA studies show that fatigue has no significant effect on the manifestation of HGN." Guide at 9. In support of that proposition, the Guide cites to the 1981 NHTSA study. That study, however, acknowledges that "possible effects of fatigue or circadian rhythms on gaze nystagmus could be significant." 1981 Study at 9. The study authors tested one element of the HGN test – the correlation between BAC and the angle of

nystagmus onset – at different times of day and night, and found a significant correlation between alcohol ingestion and angle onset as the day gets later. It cannot be concluded from this narrow finding that there is no correlation between HGN and fatigue. Rather, it affirmatively suggests that HGN angle onset correlates with time of day.

The government's own expert, Lt. Col. Rabin, acknowledges that smooth pursuit is affected at levels "two times less than the legal limit of intoxication." Rabin Aff. at 2. He also admits that end-point nystagmus "can occur normally." Id. Dr. Booker points out that the 1981 study suggests a correlation between nystagmus onset angles and fatigue. "Considering . . . the SCRI developers produced experimental data showing nystagmus onset to be a function of the time of day of the test, it is remarkable that no investigation was conducted into the possibility that the prevalence of non-alcohol induced end-position nystagmus might be a function of time of day." Booker Study at 116, Ex. 8. Finally, the measurement of angle onset is an extremely difficult measurement to perform accurately and has a large impact on estimations of BAC. See 1981 NHTSA Study at 30; Cole Research Review 1994 at 545 ("The task for the officer to detect such small changes [of two of three degrees] is daunting, if not impossible."); State v. Witte, 251 Kan. 313, 320, 836 P.2d 1110, 1114 (1992) (listing concerns about police ability to measure angle onset and citing authorities).

In sum, every one of the HGN components is either controverted, or very difficult to perform under actual field conditions.

3. Response: The Booker Study

As reported in his peer-reviewed article, Dr. Booker tested the effects of fatigue on end-

position nystagmus, one of the HGN test components. His results were as follows:

1. 55% of non-drinking subjects exhibited nystagmus after being awake for an average of 24.5 hours
2. 19% of well-rested subjects exhibited nystagmus prior to being dosed with alcohol
3. 62% of subjects exhibited nystagmus at BAC levels of 0.00 when tested immediately after their blood cleared of alcohol
4. the dose-response relationship between alcohol and end-position nystagmus varied widely (37% to 68%) depending on whether the subject's BAC was rising or falling

Dr. Booker also examined the HGN procedures. In order to accurately assess the presence of HGN, officers must hold the stylus still for four seconds, four times, to properly assess whether end-point nystagmus is present. See NHTSA Manual at VIII-15 (instructing officers to assess maximum deviation nystagmus after four seconds) (attached to Def. Mem. as Ex. B). The entire HGN test should take a minimum of 48 seconds to conduct properly. Dr. Booker then reviewed fifty-two arrest tapes in which police officers administered the HGN test. He found that only 15% of officers ever held the stylus still for four seconds even once, and that only one officer conducted the entire test properly.

Dr. Booker concludes that the HGN test is routinely administered in "situations where a high incidence of false positives is to be expected." He describes the NHTSA assertions of 77% accuracy as "inflated and erroneous." He also concludes, based on his observation that 98% of HGN tests are improperly administered, that either test protocols or training procedures are "inadequate to assure proper administration." Booker Study at 116.

Neither the Guide nor the Rabin affidavit address the scientific concerns presented in the

Booker study. They do not establish either the reliability of the test, or its general acceptance in the scientific community. They should therefore be discounted as an incomplete and insufficient basis for a finding of scientific reliability. See Young v. City of Brookhaven, 693 So.2d 1355, 1360 (Miss. 1997) (finding HGN test not generally accepted within the scientific community and cautioning that only proper use of HGN is to establish probable cause due to the "high degree of likelihood that the jury would confuse the proper weight to be given the test results").

B. Jarrell Affidavit

The government offers the affidavit of Detective Jarrell in support of the admissibility of the FSTs that he administered to Mr. Horn. The Jarrell affidavit, however, is irrelevant at this stage of the proceedings. Det. Jarrell has no independent expertise or knowledge which would contribute to the reliability inquiry. The fact that he may have administered the tests many times, and that he was trained to do so, has no bearing whatsoever on the question of whether the tests themselves are reliable. Indeed, Det. Jarrell may have been very well trained in, and be very good at administering unreliable tests.

To put it another way, this hearing is centrally about the question of whether police officers – who have no independent scientific background – can nevertheless testify at trial as field sobriety experts by relying on the reliability of NHTSA studies and NHTSA-sponsored training. The Supreme Court of New Mexico addressed this precise issue in State v. Torres, 976 P.2d 20, 127 N.M. 20 (1999). The Court held that a police officer, although trained in HGN methodology, lacked the independent scientific expertise to lay the foundation for the admission of the HGN test itself, and that therefore "only a scientific expert may testify as to [the HGN]

results." 127 N.M. at 33, 976 P.2d at 33. Only after an independent scientific foundation is laid may police officers testify about their training and administration of the test. Id.; see also Barrett v. Atlantic Richfield Co., 95 F.3d 375, 382 (5th Cir. 1996) (animal behaviorist not qualified to testify about the cause of his observations of chromosomal changes in rats because the causes lay beyond his expertise).

Det. Jarrell should not be permitted to bootstrap his own contested expertise into evidence. Defendant thus respectfully requests that the Jarrell affidavit be stricken.

D. The Five NHTSA/DOT Studies

The five NHTSA/DOT studies constitute the only evidence submitted by the government in support of the validity of the WAT and the OLS field sobriety tests. The admissibility of those two tests thus turns exclusively on the scientific acceptability and sufficiency of the five studies. The studies also purport to establish the validity of the standardized HGN test as a predictor of intoxication.

Daubert requires a court to consider, inter alia, whether scientific evidence is susceptible of testing, the error rate, whether the evidence has been peer reviewed, and whether the results are generally accepted in the scientific community. In this case, although the SFSTs are susceptible of testing, they meet no other Daubert criteria. They have not been adequately tested – either in the lab or in the field – to establish reliability or validity, and the results from the tests that have been done indicate high levels of unreliability. The government's scant evidence indicates that error rates are either indeterminate or unacceptably high. None of the NHTSA studies are peer reviewed, and the only peer reviewed article to assess WAT and OLS concludes

that they are unreliable. See Cole Study 1994, Ex. 2. Finally, the government offers no evidence that the SFST battery is generally accepted by any scientific community whatsoever. In sum, nearly every aspect of the Daubert inquiry indicates that this evidence should be excluded.

1. The lab studies

Dr. Cole, Mr. Brull, and Dr. Wiesen each independently reviewed the 1977 and 1981 NHTSA laboratory studies. Their conclusions are summarized below.

All three defense experts agreed that the 1977 and 1981 lab studies appeared to have been performed in a scientifically acceptable manner. They also agreed that the results of those lab studies presented serious concerns about FST reliability, concerns that the study authors themselves recognized and documented.

The fundamental problem with the lab tests is that they do not replicate the uncontrolled, highly variable conditions in the field and therefore overstate the accuracy of the tests. Lighting, weather conditions, slope of the ground, differences in officer training and administration, not to mention the fear and stress attending the civilian-police encounter, all potentially worsen SFST scores, yet are not accounted for in the lab setting. To put it another way, the lab studies do not accurately measure what is at issue in this case – the reliability of actual SFSTs administered under real-life conditions. See Samuel, 96 F. Supp.2d at 502 (rejecting vehicle roll-over test because "the 'fit' between the test and the issues in this case is not a good one" (citing Daubert, 509 U.S. at 591)).[11]

---

[11] Dr. Marcelline Burns, the principle author of every government study save one, distances herself from her own lab studies by asserting that "the laboratory data are only indirectly enlightening about current roadside use of the tests." Colorado Validation Study at 1.

The exaggerated reliability of the lab studies is exacerbated by the fact that the two lab studies used different techniques from each other as well as the field studies. In performing the HGN, for example, the 1977 study used a chin rest and told participants to cover one eye. 1977 Study at 13, 48. In 1981, no chin rest was used and both eyes were open. The use of the chin rest exaggerates the accuracy of the HGN because small deviations in angle and perception make large differences in the HGN test results. See Cole Research Review 1994 at 545; State v. Witte, 251 Kan. at 320, 836 P.2d at 1114. The NHTSA standardized HGN test does not include covering one eye; indeed, covering one eye can actually cause nystagmus. Rabin Aff. at 2. The 1977 results therefore do not reflect the actual accuracy of the HGN as eventually tested and implemented.

The lab tests contained additional problems. For example, "borderline cases [were] assumed to fall into the non-error category," 1977 Study at 28, 31, thereby inflating the assessment of arrest accuracy. Wiesen Aff. at 4. The studies were performed by the same principle author, Marcelline Burns, under contract with a government agency with a specific research agenda. They have not been peer reviewed, and the government provided no evidence that the results have ever been replicated by other researchers. Brull Aff. at 5; see also Wiesen Aff. at 4 (detailing other inherent flaws).

Even assuming the validity of the lab studies, those studies conclude that there are significant problems with the reliability and validity of the SFSTs.

FSTs are inherently inaccurate. As the authors put it, "[q]uite simply, there are no

---

This statement is misleading. The lab studies are highly enlightening in that they demonstrate the inherent limitations of accurate field sobriety testing even under the best of circumstances.

behavioral cue [sic] which differentiate infallibly in a +/- .02% BAC margin." 1977 Study at 27, 41. The mean absolute error rate of 0.03% in the 1981 study also indicates high unreliability. Wiesen Aff. at 5.

FSTs generate a high false arrest rate. Out of 101 arrests, 47 were of people with BACs below 0.10. As the authors admitted, "[o]bviously, an error rate of 47% in making arrests is not acceptable." 1977 Study at 25. See Brull Aff. at 7.

Low inter-rater reliability. Inter-rater reliability for the arrest decision was .59. 1981 Study at 32. In other words, when the same subject under the same conditions was rated by different officers, the officers agreed on the arrest decision only 59% of the time. Brull Aff. at 6. Since the predicate of reliability is repeatability, the fact that different officers using FSTs come to different conclusions regarding the same person indicates high FST unreliability. Cole Study 1994 at 100, Ex. 2. Also troubling, as Brull explains, this statistic suggests that officers are likely using FSTs as "proof" of their arrest decisions, while basing their decisions on numerous other, subjective factors. Brull Aff. at 7.

Low test/retest reliability. 145 participants returned to be retested at the same alcohol doses. 1981 Study at 34. Officer test-retest reliability was only .57. In other words, officers agreed with their original decision only 57 percent of the time. Wiesen and Brull identify this as a particularly low, and therefore unreliable, test-retest score. Wiesen Aff. at 6; Brull at 7. The study authors also appear to consider their test-retest scores to be low. 1981 Study at 34 ("Test-retest reliabilities for psychomotor tests are typically on the order of 0.7.").

Peer-reviewed research shows the OLS and WAT to be unreliable. In his peer-reviewed 1994 study, Dr. Cole performed an experiment designed to test the hypothesis presented by

Burns, et al., namely, that the WAT and OLS accurately indicate intoxication. Officers observed 21 videotaped subjects performing the two tests, none of whom had ingested any alcohol. Dr. Cole's results suggest that the Burns' studies significantly overstate SFST accuracy. Out of 21 subjects, officers indicated that only three were totally unimpaired, and would have arrested 46 percent as having had too much to drink. Brull independently reviewed the Cole study. Finding the results "startling," Brull opines that the study "seriously undermines the confidence in FSTs as a predictor of alcohol impairment." Brull Aff. at 11.

In sum, under ideal laboratory conditions, NHTSA's non-peer reviewed studies indicate high levels of FSTs unreliability, Brull Aff. at 8, while Dr. Cole's peer-reviewed work indicates that FSTs are even more unreliable than NHTSA's work suggests.

2. The field studies

The field studies offered by the government are flawed and deeply misleading as to the actual scientific reliability of FSTs. Brull, Wiesen and Cole exhaustively document the flaws in the field studies. Their conclusions are summarized below:

The studies rely on highly biased subject sample. The field studies were performed on people who were stopped on suspicion of drunk driving. The average BAC of drivers arrested for driving under the influence at the time of the studies was .17. 1981 Study at 60. In other words, the police were stopping, and performing FSTs, on a highly biased sample. Officers who perform FSTs on these subjects and conclude that the person is intoxicated are likely to be correct most of the time, not because FSTs are particularly accurate, but because the likelihood that the subject has been drinking is very high. See Brull Aff. at 9. In addition, the 1981 study

authors admitted that the study was skewed because sobriety tests were only given to subjects who appeared intoxicated. 1981 Study at 63 (subjects "represent a subset of this population biased toward high BAC").

The studies indicated high margins of error. The 1981 authors report that after training, officers in the field erred on average by .05 in their estimations of BAC, p. 63-64; Brull Aff. at 9; Wiesen Aff. at 8. In other words, an officer estimation of .10 could in fact be as low as .05 or as high as .15 BAC. Wiesen Aff. at 8. Officer error margins in the lab were lower – .03 – but still troubling. 1981 Study at 62.

The field studies relied on PBT results as the criteria for FST accuracy. While the lab studies evaluated FST results against actual BAC as measured by Intoximeter breathalyzer readings, most of the field studies compare FST results to PBT results. The PBT, however, is itself unreliable and inadmissible. See United States v. Iron Cloud, 171 F.3d 587, 598 (8th Cir. 1999). Where FST results are calibrated only to PBT results, the studies can establish at best that FSTs are no more reliable than PBTs. Indeed, the 1983 report concluded that:

> [T]he test battery appears to be about as effective as the use of PBTs in improving the BAC distribution of those arrested (e.g., a reduction in false positives). 1983 Study at 11.

In the Daubert context, this conclusion is tantamount to conceding inadmissibility.

Police already knew the answer. Many of the police already knew the results of the PBT administration, thus affecting their evaluation and reporting of FST results. Wiesen Aff. at 10. The 1981 study authors admit that "most of the officers' BAC estimates were invalid" for this reason, 1981 Study at 63, and in the 1983 study, the authors likewise warn that the officers probably calibrated their results to match PBT results, thereby rendering the data invalid. 1983

Study at 9; Wiesen Aff. at 10; Brull Aff. at 10.

Insufficient monitoring of testing. The 1983 study authors admit that "no statement can be made as to how closely the requested data collection procedures were followed." 1983 Study at 6. Wiesen Aff. at 10.

Complete lack of statistical analysis. Data reported in the field studies were insufficient to support basic statistical analysis or provide a meaningful error rate. The 1981 field study authors report that their own "data are not appropriate for significance testing." 1981 Study at 54; see Wiesen Aff. at 7. For the same reason, the 1983 field study is "suspect." Wiesen Aff. at 10.

The Florida and Colorado Studies are incomplete and thus an inappropriate basis for a reliability finding. The government submitted the conclusions of the Florida and Colorado studies without submitting their underlying data or methodology. Particularly since the field study context naturally inflates the reliability for FSTs for all the reasons stated above, without data or methodology the bare assertions of these two studies cannot be evaluated or relied on. Brull Aff. at 9-10; Wiesen Aff. at 11-12.

In sum, the 1977 and 1981 laboratory studies present the best possible scenario for FST reliability, and those studies indicate high margins of error and unreliability. The 1981 and 1983 field studies are not reliable or conclusive, due to flaws in data collection and methodology, and the Florida and Colorado studies as presented are simply incomplete. The overall conclusion to be drawn is that there is insufficient data to support the claim that the three standardized field sobriety tests are scientifically reliable indicators of intoxication.

CONCLUSION

State courts are split every which way on the question of whether standardized field

sobriety tests are scientific evidence and whether they are admissible in court. No federal court has definitely answered the question.[12] The fact that some courts have taken judicial notice of the reliability of the HGN while others exclude it altogether make a rigorous inquiry based on first principles in this case all the more important. Daubert places the burden squarely on the propounding party, in this case the government, to support its evidence with independent scientific validation. Here, the government has not done so. It relies on biased, un-peer-reviewed research, compilations by interested advocates, and the testimony of the very officer whose expertise is at issue. The government offers only a single affidavit from an independent scientist, Lt. Col. Rabin, and his expertise is only generally related to the question. By contrast, the defendant's independent experts and peer-reviewed studies, not to mention the raging dispute that plagues the courts, cast more than enough doubt on the reliability of the SFSTs to warrant exclusion.

Like PBTs and polygraphs, SFSTs may have many appropriate applications. They are useful tools in the probable cause inquiry and, like the PBT, appear to provide a reasonable basis to arrest a driver. But "probable cause" is a "commonsense, nontechnical conception[] that deal[s] with 'the factual and practical considerations of everyday life on which reasonable and prudent men, not legal technicians, act.'" Ornelas v. United States, 517 690 695 (1996) (quoting Illinois v. Gates, 462 U.S. 213, 231 (1983)); see also United States v. Williams, 10 F.3d 1070, 1074 (4th Cir. 1993) ("[P]robable cause is a practical, nontechnical concept based on probabilities

---

[12] Only one federal court appears to have addressed the question at all. In Volk v. United States, 57 F. Supp.2d 888 (N.D. Cal. 1999), the district court found no abuse of discretion where the magistrate judge admitted FST evidence, based on the lower court's finding that the officer's specialized experience and training made the evidence reliable. No Daubert hearing was held, and no independent or competing evidence was submitted on the question of reliability.

and common sense."). The Daubert reliability inquiry demands the very opposite approach, namely, a non-commonsense, technical, rigorous, scientific analysis of expert evidence. Daubert, 509 U.S. at 588 ("[I]n order to qualify as 'scientific knowledge,' an inference or assertion must be derived by the scientific method. Proposed testimony must be supported by appropriate validation . . . ."). Mr. Horn respectfully submits that the government has not met this substantial burden of showing that SFSTs are the sort of information designed or suitable for admission as evidence in federal court, and that therefore the results of such tests should be excluded.

Respectfully submitted,

JAMES WYDA
Federal Public Defender
for the District of Maryland

_____
Sasha Natapoff
Assistant Federal Public Defender
100 S. Charles Street
Tower II, Suite 1100
Baltimore, Maryland 21201
(410) 962-3962

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that on this ___ day of November, 2001, a copy of the foregoing Reply was delivered to Paul Marone, Special Assistant United States Attorney, U.S. Army Garrison, Building 310, Wing 10, Aberdeen Proving Ground, Maryland, 21001.

_____
Sasha Natapoff
Assistant Federal Public Defender

Ex. 1: Cole 1994 Champion article

Ex. 2: Cole Psych. Motor article

Ex. 3: Cole resume

Ex. 4: Brull affidavit

Ex. 5: Brull Resume

Ex. 6: Wiesen affidavit

Ex. 7: Wiesen resume

Ex. 8: Booker article

Ex. 9: Booker resume

Ex. 10: Forensic Science description

AFFIDAVIT OF JOEL P. WIESEN, Ph.D.

I, Joel P. Wiesen, do hereby affirm and state as follows:

1. Education and Experience.

I am an industrial psychologist, specializing in the development of fair, valid tests of human abilities. I was awarded a Ph.D in Psychology from Lehigh University in 1975. My major field of doctoral study was experimental psychology and my minor field of study was psychometrics and statistics. My graduate studies included courses in both psychology and mathematics. I have taught undergraduate and graduate-level courses in statistics and research methods at Northeastern University and elsewhere.

For over ten years I worked for the Division of Personnel Administration, which is the agency of the Commonwealth of Massachusetts responsible for administering the civil service examination program for both the state and municipal civil service employees, covering some 70,000 state employees and some 200 cities and towns. My responsibilities included the development and validation of examinations, supervision and management of a staff of examiners who developed civil service examinations, as well as the oversight and review of examinations prepared by various consultants hired for this purpose. I also advised the agency and served as an expert in various matters related to test development and validation.

For the past 10 years I have been an independent consultant and have specialized in the development and validation of tests, mainly tests used for personnel selection purposes. Since 1980, I have done work for and advised private and public organizations in the area of test development and validation. Some of these organizations are: Cummins Engine Company, Bell Atlantic (now Verizon), T.J. Maxx, the Commonwealth of Pennsylvania, the Commonwealth of Virginia, the Commonwealth of Massachusetts, the state of Maryland, the city of Oklahoma City, the city of Springfield, Massachusetts, the city of Orlando, and the U.S. Department of Justice.

I am also a published test author, having developed a test of mechanical aptitude which is now used nationwide in some Fortune 250 companies as well as many smaller companies. Although I develop and use mostly written tests, I have worked with and developed human performance tests, including tests of physical abilities for jobs, especially for the job of fire fighter.

I am a member of the following professional societies and organizations: American Psychological Association, American Psychological Society ("Founding Fellow"), the Society for Industrial and Organizational Psychology, the Personnel Testing Council of Metropolitan Washington, the American Statistical Association, the Assessment Council of the International Personnel Management Association, and the New England Society for Applied Psychology. I was elected and served as president of the last two organizations.

I have also served as a reviewer for professional societies, including journal reviewer for the International Personnel Management Association, and reviewer for several annual conferences of the Society of Industrial and Organizational Psychology and of the Assessment Council of the International Personnel Management Association. In this role, I reviewed manuscripts submitted for acceptance for the journal or for presentation at annual conferences.

In addition, I make presentations at national conferences and other professional meetings on various aspects of testing, including such topics as: test development, test validation, and test fairness. These conferences include: the American Psychological Association, the Society of Industrial and Organizational Psychology, and the Assessment Council of the International Personnel Management Association.

I am a licensed psychologist in Massachusetts and Pennsylvania.

2. My Charge

I was asked by the Office of the Federal Public Defender to review certain publications and, based on those publications, to evaluate the Field Sobriety Test (FST) as I would evaluate any other test of human capacity, report on its quality and validity as a test, and offer my opinion as to whether the FST meets the scientific standards of my profession.

3. Criteria for Evaluating Tests and Testing Research

New tests of human performance must live up to certain professional criteria prior to being accepted by psychologists as valid and useful measures. Over 50 years ago, the American Psychological Association developed and published a set of guidelines for psychological testing, and these are periodically updated.

In 1999, a 15-chapter book entitled, "Standards for Educational and Psychological Testing," was jointly issued by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. These standards are accepted in and followed by the professional testing community, although each standard may not apply to every test or testing situation. The book defines "test" as:

> "An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process." (p.183)

FSTs fall under this definition of a test since they involve measuring specific behaviors of people in a standardized manner.

In the field of industrial psychology, as in the other fields of psychology which use tests, these 1999 standards are used by test users (the person or agency responsible for the choice and

administration of a test, and the interpretation of test scores), test publishers, and test authors as criteria for the evaluation of tests and testing practices. To the extent that the applicable standards are not followed or met, a test user should tend to avoid using a given test, especially for high-consequence decision making. To the extent that a test does not meet these standards, it is also less likely the test will be published or used by testing professionals. If tests are used which do not meet the applicable standards, the test results will be treated as less valid.

4. Summary

My opinions on the scientific acceptability of the FST are based on my review and analysis of the following five publications:

1. Burns and Moskowitz, 1977, "Psychophysical Tests for DWI Arrest"
2. Tharp, Burns, and Moskowitz, 1981, "Development and Field Test of Psychophysical Tests for DWI Arrest" (volume 1 only)
3. Anderson, Schweitz, & Snyder, 1983, "Field Evaluation of a Behavioral Test Battery for DWI"
4. Burns & Anderson, 1995, "A Colorado Validation Study of the Standardized Field Sobriety Test (SFST) Battery"
5. Burns & Dioquino, undated, "A Florida Validation Study of the Standardized Field Sobriety Test (S.F.S.T.) Battery"

In addition, I reviewed parts of Chapters VI, VII, and VIII of the "DWI Detection and Standardized Field Sobriety Testing", an undated publication of the National Highway Safety Administration. I did not evaluate this manual, but did note the procedures described for the FST on some of the pages in Chapter VIII.

These publications, singly and taken together, show only that the FST may have promise as a psychological test. The five studies fall short of meeting professional standards in several important areas related to testing and related to behavioral science research. More and better research is needed before the scientific community can be assured that the FST is a fair, reliable, valid predictor of intoxication. If any of these studies were submitted for publication in a peer-reviewed research publication, in my opinion they would be rejected due to their serious shortcomings in methodology and data analysis.

5. Burns and Moskowitz (1977)

This report is flawed in several very serious ways. Considered as a whole, this report does not meet the professional standards of the testing community. Some of the major shortcomings of the report include:

a. The test studied and evaluated is different from the test used in the field.

In Burns and Moskowitz (1977) chin-rest and angle indicating equipment was used for the nystagmus test (p.13, next to last ¶; p. 14; p. 48, fourth ¶), and this equipment was said to be the reason that their data showed "a substantially larger BAC-nystagmus correlation than reported in the data from Finland" (p.48, second ¶). However, later reports indicate that this equipment is not provided for use by police officers in the field. As a result, the accuracy of the FST in the field will be significantly below that reported in the 1977 study.

b. Overt bias in the evaluation of test accuracy.

In evaluating the FST accuracy, Burns and Moskowitz (1977) report that "borderline cases are assumed to fall into the non-error category" (p. 28, last sentence). In plain language, the authors artificially inflated the accuracy of the test by this method of dealing with people who fall at the borderline. Thus, the accuracy for the FST is less than they report.

c. The evaluation of accuracy capitalizes on chance.

The authors both develop the criterion score based on the data they collect, and then evaluate the accuracy of the categorizations based on this same set of data (see last ¶ on p. 28). It is well known in the field that this type of approach artificially inflates the estimate of the accuracy. A better approach involves what is called "cross validation" where the evaluation is done with a second set of data (sometimes "held out" from the original analysis). There is no simple way to evaluate the extent to which the results are biased by the method Burns and Moskowitz chose for this part of their data analysis, but it is clear based on their methodology that the FST accuracy is less than they report.

d. The test is not neutral with respect to age and gender.

The authors report that older people and women will tend to have higher scores and therefore be categorized as intoxicated more often than younger people or men (p. 34, fourth ¶; p. 119, third ¶; and p. 121). This lack of neutrality is not explored in detail in their report. This type of bias is a serious threat to the valid use of any test.

e. The officers were being watched.

The officers in this study were being watched by a member of the authors' staff (1977, p. 16, first ¶). As a result of the ever-present "trained observers", the police officers may have been more motivated than police officers in the field to carefully follow the test administration and scoring procedures. Therefore, the accuracy of the test seen in this study is likely to be a maximum, rather than to be representative of the FST accuracy

when used by police officers in the field.

f. The study is unacceptable for journal publication.

Peer-reviewed professional research journals commonly reject for publication reports with deficiencies such as those described above. Due to its errors and shortcomings, it is highly unlikely that the Burns and Moskowitz (1977) report would have been accepted for publication by the Journal of Applied Psychology, or by a similar professional research journal, had it been submitted for publication.

6. Tharp, Burns and Moskowitz (1981): The Laboratory Study

This report describes two studies: a laboratory evaluation (described in Chapter 2) and a field evaluation (described in Chapters 3 and 4). I will separately consider these two parts of the report. The laboratory evaluation of the report is flawed in several very serious ways. Considered as a whole, this part of the report does not live up to the professional standards of the testing community. Some of the major shortcomings of the report include:

a. Many false positives.

Of the people tested who had no alcohol, about 20% were classified as too impaired to drive (known as "false positives"); 18% were so classified by officers and 21% by observers, that is, the authors' staff (p. 20, second ¶; p.22, the first two entries in column 3). This is a high rate of incorrect classification of absolutely sober people.

b. The "mean absolute" error is high.

The authors calculated the difference between the actual blood alcohol content (BAC) and the BAC estimated by the police officer who administered the FST, and then found the average of these differences, ignoring the direction of the difference (they refer to this as the "mean absolute value," p. 21, Table 3). They report the average difference to be .030% (p. 20, first ¶). Although the authors do not give the distribution of these errors, it is reasonable to think that about half of the officers' BAC estimates based on the FST are wrong by more than .03%. So, for example, half the time the FST predicted a BAC of .10% the actual BAC would be either less than .07% or more than .13%. This amount of error is high in relation to the range of BAC being considered.

c. Test results vary with time of day and scoring does not account for time of day.

The test score for the horizontal gaze nystagmus (HGN) test depends, in part, on the "angle of onset" (p. 87, line C2). The authors report a statistically significant decrease in the angle of onset for people in the alcohol group tested after midnight (p.9, last ¶). This means that the test score varies based on the time of day the test is administered. The report does not address the implications of this statistically significant finding.

d. Over-reliance on pilot work.

"Pilot work" usually refers to a small-scale investigation intended to refine a study's data collection methods. Usually pilot work is done with relatively few people, and the exact procedure used and results obtained may not be reported. In contrast, usually a "study" is done with a sufficient number of people to reach scientifically sound conclusions, and a full report of the data collection methodology and the data analysis is provided.

The authors used "pilot work with gaze nystagmus" to "rule out a number of unimportant variables" including: stimulus brightness, room brightness, fixation distance, velocity of the stimulus movement, monocular versus binocular fixation, instructions to inhibit nystagmus, and vertical positioning of the eye. These seven variables are all potentially important, since they are likely to occur often in real-life applications. Most of this pilot work is not reported in any detail (p.7, fourth ¶). Without a full study clarifying the effect of such variables, the standardization of the test is called into question.

e. Agreement between officers is low.

The 1981 study included a retest of 145 participants who returned a second time to be tested under the same alcohol dose (p. 34, fourth ¶). That the dose was the same for the two sessions is seen in the correlations of .96 to .97 reported in Table 14 (p. 35). The degree of agreement between raters for the total FST score is reported in terms of test-retest reliability to be .57 or .62, depending on whether officers' or observers' data are considered (rightmost column, p. 35). Usually inter-rater reliability of .8 (or even .9) or more is achievable. Reliability around .6, as in this study, is extremely low.

f. Test administration procedure changed over time.

In the 1981 report, the test-taker follows the visual stimulus with both eyes (1981, p. 85, last ¶). In the 1977 report, the test-taker was instructed to cover one eye when taking the test (1977, p. 90, ¶ 2). This may constitute a new version of the test. The studies do not tell us to what extent the evaluations of the earlier versions of the test accurately describe the new version.

g. Police Officers did not follow the decision criteria.

The authors give the decision criteria in Appendix B, but also state that they "were not necessarily followed by the testers" (i.e., by the police officers, p. 19, first ¶). In other words, police officers did not necessarily use the FST results to decide whether the person tested was too impaired to drive and to estimate the BAC. Not only does this mean that

the test results (correct or incorrect arrest decisions) cannot be attributed to the FSTs alone, but it indicates that officers in the field will not follow the decision guidelines.

h. False positive rates calculated on people tested on two days.

The authors report false positive rates in Table 8 (p. 27) which are based on 441 testings. But only 296 people were tested (p.15), so Table 8 includes data from 145 people who returned on another day and tested a second time. Table 4 (p. 22) shows a much lower error rate for the placebo dose people on the second day of testing, as compared to their first day of testing. In the real world people are not called back on another day, given the same dose of alcohol, and then retested. This means that the false positive rates reported in Table 8 are artificially low.

7. Evaluation of Tharp, Burns and Moskowitz (1981): The Field Study

The field evaluation of the 1981 report is flawed in several very serious ways. Considered as a whole, this part of the report does not meet the professional standards of the testing community. Some of the major shortcomings of the report include:

a. Authors say the data are not appropriate for statistical significance testing.

The authors say "the data are not appropriate for significance testing" (p. 54, last ¶). This is a very serious and worrisome statement. Tests of statistical significance are fundamental to this type of research, since they are the main method by which hypotheses are tested and conclusions drawn. That the data cannot be tested with statistical tests is a fundamental flaw in the study.

b. Authors report that the data were biased.

The authors report that the "data obtained during the ride-alongs may be biased" (p. 57, number 2, second ¶). Specifically, they say that most officers waited until the end of their shifts to fill out the data forms, by which time they probably knew the BAC levels based on the breath tests (p. 63, ¶b). The only field data the authors consider valid are for 73 arrestees who were given blood or urine tests, and these are reported to be a "biased sample" in part because about one third of them were suspected of being under the influence of drugs other than alcohol (p. 63, ¶b and ¶c). For this reason, the accuracy of the test as reported in this study is artificially inflated, rather than representative of the FST accuracy when used by police officers in the field with people who are not on drugs other than alcohol.

c. No analysis of the data by ethnic group.

Some physiological measures vary by ethnic group. Although the authors collected ethnic group identification (p. 44, first line; p. 52, section 3), and although the 1977 report indicated gender and age differences in FST performance, the authors failed to report data by ethnic group (p. 58). A reviewer thus cannot tell if the test operates equally across ethnic groups.

d. The "mean absolute" error is high.

The authors calculated the difference between the actual BAC and the BAC estimated by the police officer who administered the FST, and then found the average of these differences, ignoring the direction of the difference. They report that, after training, the officers' average difference is .0537% (p. 63, last ¶, and p. 64). Although the authors do not give the distribution of these errors, the implication is that about half of the officers' BAC estimates based on the FST are off by more than .0537%. This is high in relation to the range of BAC being considered, which would in turn lead to a high proportion of false arrests. This is reflected in the authors' report that only half of the people with a BAC of .10% to .149% would be arrested, and that 28.6% of the people with BAC of .05 to .099 (i.e., legal drivers) would be arrested (p. 66). Both the low detection rate and the high number of false positives are based on data collected after the police officers were trained (p. 66).

e. An unspecified number of police officers had problems scoring the tests.

The authors report that most officers had "little problem" scoring the balance test, but do not report how many did have problems, nor what the problems were (p. 42, first ¶). The authors report that by the end of training "very few questions remained" but do not report how many or what these questions were (p. 42, end of third ¶). If the officers had trouble learning the procedure when trained by the authors' staff, then it may be that officers in operational settings will have even less clarity about how to administer and score the FST.

f. Sample of police officers is biased.

The authors started the field evaluation study with 20 police officers, but only used data from 11 of them, because the other 9 did not provide data which the authors deemed useable (p. 54, last ¶; p. 64). This sample is both small and biased through self-selection. The authors say that 5 of the 9 officers who did not provide useable data had a "poor attitude" or showed "lack of cooperation" (p. 54, last ¶). Since the laboratory study showed considerable difference between officers in their success in using the FST (see, e.g., p. 26), the sample of more motivated or more cooperative officers may not be representative. For this reason, the accuracy of the test as reported in this study is

artificially inflated, rather than representative of the FST accuracy when used by police officers in the field.

g. The test scoring system changed over time.

The field evaluation part of the 1981 report presents a scoring system for the FST (p. 44, table 17). This system has 9 "checkmarks" or points for the walk and turn (WAT), 5 checkmarks for the one legged stand (OLS), and 8 for the HGN, for a total of 22 possible points. However in Appendix B another scoring system is presented (p. 87-88), with 10 "checkmarks" or points for the WAT, 7 checkmarks for the OLS, and 8 for the HGN, for a total of 25 possible points. Further, the scoring system "decision criteria" described by the authors (p. 88) uses scores from the individual tests, and therefore deviates from the total number of points approach used in the 1977 report (1977, p. 28, section C). To the extent that the test administration scoring system changed, we have a new version of the test. This is true even across the two parts of the 1981 report itself, as just described. As a result, the scores on the changed test may be higher or lower, or the accuracy or correlation with criteria of interest may have changed. Since the new and old versions of the test were not compared, the evaluations of the earlier versions of the test may not be applicable to the new version.

h. Test administration and scoring in the field is uneven in quality.

The authors report that in the field some police officers (number not given) "forgot or ignored most of the administration procedures" other than for the nystagmus test, but the officers did not recognize they forgot (p. 70, first ¶). They also indicate that officers are reluctant to use any scoring system (p. 69, next to last ¶). Both of these are serious threats to the validity of the FST as used in the field. Even the report by Anderson, Schweitz, and Snyder states that Tharp, Burns and Moskowitz "did not use a standardized procedure for combining [the test] results and reaching an arrest/no arrest decision" (1983, p. 3, second ¶). To the extent that the combining of test results was left to the judgment of the individual officers, the FST scoring was not standardized.

8. Evaluation of Anderson, Schweitz, and Snyder (1983)

This report describes a field study in which FSTs were administered by police officers to drivers stopped for suspicion of driving while intoxicated. One might expect this study to be more objective and better than the previous reports, since it was conducted by different researchers. Unfortunately, this report too is flawed in several very serious ways. Considered as a whole, this report does not meet the professional standards of the testing community. Some of the major shortcomings of the report include:

a. Data collection procedures were unmonitored and so cannot be trusted.

The data collection procedures were designed to "minimize the possibility that knowledge of PBT [breath test] results would be available to officers before administering or recording battery scores" (p. 6, third ¶), but the authors report that "no statements can be made as to how closely the requested data collection procedures were followed" (p. 6, third ¶). If the PBT was administered before the FST, the scoring of the FST would likely be intentionally or unintentionally biased in favor of the accuracy of the FST. As a result, it is not possible to trust the results of this study.

b. The arrest decisions were made based on breath analysis as well as FST.

The criterion for this study was the accuracy of the police officers' arrest decisions. However, the authors report that "most arrest decisions were based on PBT [breath test] data, rather than just test battery data" (p. 9, ¶ 2). To the extent that the FST was not individually evaluated, the study can make no statement as to the accuracy or usefulness of the FST.

c. The relevant data (from North Carolina) are not presented in full.

A little more than one quarter of the data collected on the FST came from North Carolina, the only jurisdiction which did not administer the PBT (p. 7, third ¶; p. 9, third ¶). The authors do not report all the FST data from this jurisdiction, but only the data for two of the three tests which comprise the FST, saying "Only those cases for which the combined 2 test score (sic) indicated there should be an arrest were included in this data set" (p. 9, third ¶). Since data for the full FST were not presented, the full FST cannot be evaluated based on this report.

d. No statistical tests were conducted.

The authors draw conclusions based on inspection of data, but do not conduct statistical tests to support their observations (p. 9, last ¶). That no statistical tests were used is highly unusual for this type of study, and makes the conclusions suspect.

e. The FST was not administered in a standard fashion.

The administration of the FST was not standardized. The police officers in the field decided which and how many of the three parts of the standard FST to give (p. 7, Table 1). The authors provide no reason for this non-standard administration of the FST. The authors report a new system for scoring the tests that has two types of cutoffs: a cutoff on each test "if it was the only one used" (p. 4, third ¶), and a cutoff based on specific scores on the WAT, and HGN tests combined (p. 4, Figure 1). The cutoffs reported for the WAT are not the same when used alone and with the HGN test. In the narrative for the

WAT test, the authors say "If the test score is greater than 1, classify the subject as having a BAC of above 0.10%" (p. 4, next to last ¶). In contrast, Figure 1 on the same page shows that people with WAT scores of 2, 3, 4, or 5 should pass if the HGN score is low enough. Because of the non-standard test administration and scoring, the results of the study cannot be definitely attributed to the full FST or to any of its component tests.

f. Two different devices were used to measure BAC.

The authors report using two different devices for measuring BAC, one more precise than the other (p. 7, ¶ 2). They also report that the more accurate measure was available only for people arrested, and that most of the measurements were made using the less precise device (p. 7, ¶ 2 and Table 1, last column). To the extent that the BAC measurement device was giving scores that were generally too high or too low, the evaluation of the FST accuracy is similarly flawed.

g. The authors suggest extreme caution in analyzing the data.

The authors say "Two major reasons make it necessary to be extremely cautious in analyzing the data collected in this study" (p. 9, second ¶). The first, lack of random assignment of officers to conditions, means that officers chose to give or not give the FST. It may be that officers who chose not to give the FST will not do so as faithfully or well as those officers who volunteered to give the FST, especially since officer motivation was identified in earlier reports as an important, relevant variable. Further, on p. 8 the authors say "the accuracy figures in Table 2 cannot be considered as applying to the entire population of drivers expected to be stopped by the police on suspicion of DWI" (p. 8, ¶ 2). I accept the authors' statements that the analysis of the data and the conclusions drawn are limited by these matters.

9. Evaluation of Burns and Anderson, A Colorado Validation Study (1995)

This report describes a study based on information drawn from impaired driving arrests in seven Colorado law enforcement agencies. This report is too incomplete to form the basis of an opinion regarding test validity. Specific flaws include:

a. Sections IV and V are missing, which appear to include the methodology, results and data analysis. Without these sections it is impossible to evaluate the quality of the study or rely on its conclusions.

b. Data was provided by volunteer officers (p. 2, column 2, first ¶). The use of volunteer officers raise a serious question of bias since officer motivation was identified in earlier reports as an important, relevant variable.

c. No checks on the data reporting methodology were described. Police merely reported results. Officers may well have provided data only from those FSTs for which they had high confidence, particularly since there was no check on whether breath test results were also available.

d. Results were unclear. The authors report that "officers' decisions to arrest and release were 86% correct," without defining "correct decision" (p. 5, column 1, third ¶). This lack of clarity is compounded by the use of two standards for arrest: between .05 - .10, driving while impaired; and greater than or equal to .10, driving under the influence (p. 2, column 1, first ¶).

10. Evaluation of Burns and Dioquino, A Florida Validation Study (undated)

Like the 1995 report, this report is too incomplete to allow for meaningful evaluation. Specific flaws include:

a. Complete sections – III and IV, including the methodology – are missing. Methodology was not described at all in the report as provided to me.

b. The data is incompletely described. The authors refer, variously, to "379 records," the "BACs of 256 drivers," and "313 cases" without explaining why the number changed (p. 4, second ¶; p. 5, first ¶).

11. Evaluation of all five studies.

Although all five reports concern FSTs, the procedures for administering the tests, the scoring of the tests, and the criteria change from study to study, sometimes in important ways. The five studies thus cannot be taken together to validate any particular version of the FST.

The scoring procedures changed over studies. The 1977 study used a single cutoff of 28 points (1977, p. 28, last ¶). The 1983 study used a scoring approach which had cutoffs on each of the three tests, as well as cutoffs based on specific combinations of the HGN and WAT tests (1983, p. 4). The BAC of interest also changed. The 1995 study describes two limits: .05% and .10%. Earlier, the test had been validated only for .10% (1977, p. 28, last ¶).

These changes are meaningful. What may be true for one set of test administration instructions, or for one scoring procedure, or for one criterion, may not be true for another. Thus the studies give only a general indication of the level of potential validity of the tests as described in the NHTSA manual: "DWI Detection and Standardized Field Sobriety Testing." Rather than the five studies supporting each other, they evaluate somewhat different combinations of test content and test scoring. The differences are large enough to change the validity and accuracy of the tests. The older studies are probably less germane, due to the changes in test content and scoring over time. The reports for the newer studies are grossly inadequate. Given this, and in

light of the specific critiques above (which are not exhaustive) I can only conclude that the field sobriety tests do not meet reasonable professional and scientific standards.

I declare under penalty of perjury that the foregoing is true and correct to the best of my knowledge.

Executed on: October 31, 2001 _____

Joel P. Wiesen
Director
Applied Personnel Research
27 Judith Road
Newton, MA 02459
(617) 244-8859

Affidavit of Harold P. Brull in the case of United State v. Horn

Case No. 00-946PWG

October 30, 2001MY BACKGROUND AND EXPERIENCE

My name is Harold P. Brull. My position is Senior Vice President, Public Sector Services for Personnel Decisions International (PDI). PDI is one of the world's largest industrial/organizational psychology consulting organizations with 18 U.S. offices and 19 international operations, and a staff of almost 1,000. Industrial/organizational psychology involves the definition and measurement of human attributes, particularly in employment settings.

I have been employed at PDI since 1978. In my professional capacity, I have designed and evaluated results from thousands of tests and procedures designed to measure varying quantities of specific attributes in individuals. I have worked with over 1,000 law enforcement agencies ranging in size from among the nation's largest to extremely small jurisdictions. I have taught at a variety of university settings, including Cornell University, the University of Minnesota, St. Olaf College, and the Southern Police Institute.

My educational background includes a bachelor's degree in biochemistry from Cornell University, a master's in educational psychology from the State University of New York at Cortland, and my current status as a Ph.D. candidate in educational psychology at the University of Minnesota. I am a licensed psychologist in the state of Minnesota since 1981. I am also president-elect of the International Personnel Management Association Assessment Council (IPMAAC), an organization of assessment experts operating in local, state, and national governmental settings.

Overview

For the purpose of this engagement, I was asked to review several pieces of literature that formed the basis for the use of field sobriety tests (FSTs). These tests purport to identify whether an individual has consumed alcohol, and in sufficient quantity, to exceed a threshold of impairment.

Prior to this engagement, I have had no experience, directly or indirectly, with FSTs. Rather, I viewed the evidence supplied as I would any scientific foundation for a measure which attempts to assess a human physiological, psychological, or behavioral characteristic.

Research Question

Based upon the material supplied, I have been asked to render an expert opinion as to the following questions:

> Do the procedures described accurately measure the condition in question? [An ingestion of alcohol in sufficient quantity to elevate an individual's blood alcohol concentration (BAC) to a level exceeding legal limits.]

Has the research upon which these results are based been conducted in accordance with generally accepted scientific principles?

Do the publications that I reviewed support the following legal criteria?

Is the evidence susceptible to testing?

Does it have a known error rate?

Has it been subject to peer review?

Is it generally accepted by the relevant scientific community?

The remainder of this affidavit attempts to answer these questions.

Definitions

Prior to a discussion of individual studies, several important terms and concepts must be discussed. This is particularly salient because the legal system, common word usage, and even the scientific community often use terms with little regard to their precise meaning. For example:

Validity - Validity refers to the accuracy of inferences drawn from a particular test or procedure. Thus, validity is not an inherent property of the instrument itself, but of how it is used. In lay terms, the question becomes, "What conclusions can we accurately draw from the data?" Thus, in the instance of field sobriety tests, the question, "Has the subject consumed alcohol?" is a very different question than, "Has the subject consumed sufficient alcohol to sustain an arrest and conviction?" It may be the case that field sobriety tests are valid in determining probable cause, but not in demonstrating unequivocally that a person is impaired by alcohol.

Reliability - Reliability is the property of a measurement to remain stable under different conditions. Reliability is a necessary, but not sufficient, ingredient for validity. Thus, a bathroom scale which gave a dramatically different reading each time it was stepped upon by the same person would be said to be unreliable. As such, it could not give a valid (accurate) reading of a person's weight. Reliability places an upper limit on validity.

Reliability by itself, however, does not guarantee validity. A bathroom scale which consistently gives a reading of 147 pounds when stepped on repeatedly, may still be inaccurate. Reliability estimates may take a number of different forms. For field sobriety tests, the two most salient are as follows:

Test/Re-test reliability - This refers to achievement of the same test result with the same individual under the same conditions at different points in time. It would be considered unreliable and unacceptable if the same individual with the same blood alcohol concentration produced different field sobriety test scores.

Inter-rater Reliability - For those measurements involving human judgement, inter-rater reliability refers to the likelihood that different test administrators would arrive at the same conclusion. This is of particular interest for the current inquiry, since the population of law enforcement officers administering FSTs is quite large.

Criterion - also known as dependent variable. This refers to the state or condition which is to be

predicted. Although different states use different criteria, for scientific inquiry, the criterion is generally a specific blood alcohol concentration (BAC).

Predictor - In this instance, the predictor is a single component of the field sobriety test battery, or the battery as a whole. The scientific question becomes, "To what degree do changes in the predictor correlate with (predict) changes in the criteria?"

Error Variance - This refers to differences in the predictor which are unrelated to differences in the criterion. As error variance increases, the certainty with which one can state inferences decreases. This is represented by the following diagram:

Field Sobriety Test (Predictor)

(Criterion)Not "Impaired""Pass""Fail"**Correct negative**False positive"Impaired"False negative**Correct positive**

Of the four possibilities represented by the diagram, two, the false positive and false negative, represent error variance. Both are of interest. A false negative (passing the field sobriety test but being impaired) potentially leaves dangerous individuals on the highway. A false positive renders an incorrect judgement about an individual being impaired which may then have inappropriate negative consequences for that person.

For the purposes of this issue, there are three sources of error variance:

> The test itself - What confidence can be placed, even under ideal conditions, in test results?
> The test administrator (officer) - To what extent do actions of the test administrator produce FST results unrelated to BAC?
> Environmental conditions - To what extent do these produce differences in FST results not accountable to BAC?
> The subject (arrestee) - To what extent do attributes of the subject, other than ingestion of alcohol, impact test results?

The literature supplied will now be examined to answer these questions.

Literature Reviewed

I reviewed the following documents for the purpose of rendering my opinion:

> Psychophysical Tests for DWI Arrest, U.S. Department of Transportation, contract no. DOT-HS-5-01242, June 1997, final report.
> Development and Field Test of Psychophysical Tests for DWI Arrest, Tharp, Burns, and Moskowitz, Southern California Research Institute, March 1981, final report for U.S.

Department of Transportation, contract no. DOT-HS-8-01970.

Field Evaluation of a Behavioral Test Battery for DWI, September 1983, Office of Driver and Pedestrian Research, Problem-Behavior Research Division, U.S. Department of Transportation, NHTSA Technical Note, DOT-HS-806-475.

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

A Colorado Validation Study of the Standardized Field Sobriety Test (SFST) Battery, Burns and Anderson, final report submitted to Colorado Department of Transportation, November 1995.

A Florida Validation Study of the Standardized Field Sobriety Test (S.F.S.T.) Battery, Burns and Dioquino (undated).

DWI Detection and Standardized Field Sobriety Testing, student manual, U.S. Department of Transportation, National Highway Traffic Safety Administration (undated).

Letter from Yale Caplan to Sasha Natapoff, dated 15 February 2001, and accompanying curriculum vita.

GENERAL CONCLUSIONS

The Science of FSTs

There is absolutely no question that the use of FSTs to predict impairment or blood alcohol concentrations is a scientific question. Neither the fact that the tests are behavioral or, in some cases, do not require mechanical devices, obviates this fact. The measurement of pulse by one's fingers applied to an artery is no less a scientific test than the measurement of body temperature via a thermometer. The behaviors required of a field sobriety test are not analogous to those of driving a car. One must make an inference from the former to the latter. This is comparable to an instrument reading from which one makes an inference regarding aspects of an individual's health (e.g., elevated body temperature as an indication of infection).

Sufficiency of Research Evidence

Based upon the documents reviewed, it is a reasonable question to ask whether field sobriety tests rest on a solid foundation of scientific inquiry. This foundation might reasonably include the questions raised in the legal community by the Daubert principles.

Susceptibility to testing
Known error rate
Peer review status
General acceptance by the scientific community

Each of these are discussed briefly below and in greater detail later in the report.

As for the susceptibility for testing, the predictive equation lends itself well to scientific testing. Whether in the laboratory or in the field, field sobriety test scores can be compared to a known criterion, namely blood alcohol concentration. Given that the issue is susceptible to testing, the question then becomes whether there has been sufficient research conducted to establish a known error rate.

The question of known error rate relates to the question of testing adequacy. Have sufficient tests been conducted so that the known error rate of a particular predictor may be, with any degree of certainty, stated? The answer, based on the documents I have reviewed, is an unequivocal negative.

It is of concern that the initial laboratory results have never been replicated by any other researchers or conditions lending themselves to peer review. Both the 1977 and 1981 studies were conducted by the same research organization and apparently, the same principal investigators. To establish a known laboratory error rate, one would wish to see comparable results by independent observation. However, a far more critical flaw is the complete absence, based on the documents available to me, of any evidence which would allow one to predict a known error rate in the field.

The statement by the authors of the Florida validation study (Page 2) quoting the Colorado study, "The obtained data demonstrated that more than 90% of the officers' decisions to arrest drivers were confirmed by analysis of breath and blood specimens," is simply an erroneous, misleading, and exaggerated statement regarding accuracy. The factual basis for this assertion is that over 90% of drivers arrested in the Colorado study had BAC levels above 0.05%. The average driver across the country arrested for DWI has a BAC of 0.17%. (1981, Page 19.) The combination of low BAC threshold (0.05% vs. 0.10%) and likelihood of severely intoxicated individuals being stopped makes this finding a vastly inflated estimate of predictive accuracy. Neither the Florida or Colorado studies, nor any other documents available to my review, gave any meaningful data to predict known error rate under actual field conditions.

This issue of accuracy is directly applicable to the question of peer review. One simply has more faith in results which are independently reviewed by professional colleagues. Neither of the original laboratory results or the Florida and Colorado field results meet this criteria. In fact, a single principal author, Marcelline Burns, is a principal in all results. Given that the studies all appear to be funded by federal or state traffic agencies, lack of peer review is particularly troublesome. The author's statements might lead one to believe that FSTs' error rate is less than 10%. However, this is not the case; the actual error rate must be higher by some unknown amount. Such an assertion would unlikely be permitted in a peer-reviewed article.

While the initial laboratory studies establish a baseline error rate, the field studies which I reviewed do not allow for comparable estimation of error rate in the field.

Since field sobriety tests, by their nature, are conducted in the field, this question is of paramount importance. Field studies are more difficult to control than laboratory studies. The unwanted influence of extraneous factors (error variance) almost always weakens the certainty of the experimental results.

Only one of the studies I reviewed is subject to peer review. In the scientific community, this

generally means publication in a "refereed journal;" i.e., a publication where content is judged of sufficient scientific value by professionals in the field. This study, by Cole and Mowaczyk, published in Perceptual and Motor Skills is highly critical of field sobriety tests as predictor of intoxication.

The remainder of studies, while potentially well-designed and conducted, are contract works by federal and state government agencies. As such, they may be considered as payment for delivery of a "product" to the contracting agency. They therefore represent a potential bias toward proving that field sobriety tests "work."

Regarding the question of general acceptance by the scientific community, the documents I reviewed lead me to quite different conclusions, depending upon which study is examined. The original laboratory studies, although conducted under National Highway Traffic Safety Administration (NHTSA) auspices, appears to represent solid scientific inquiry and rigorous methodology. The same, however, cannot be said regarding field studies. The initial field study in the 1981 NHTSA report was inconclusive. The documents at my disposal regarding subsequent field studies simply do not contain sufficient detail or rigor to support any hypothesis that field sobriety studies, as conducted by police officers in the field, are valid and reliable.

This last finding is particularly problematic because many of the potential sources of error in the field are simply unknowable at a later point. That is, factors which may introduce error and impact test results are simply not reproducible or subject to documentation at a later point. These might include psychological conditions on the part of the subject, interpretive skill on the part of the officer, or the impact of environmental conditions upon test results. Thus, an FST finding, presented in court, might be given erroneous deference which cannot be countered by knowable, presentable evidence which might refute it.

SPECIFIC FINDINGS FROM DOCUMENT REVIEW

Laboratory Studies

Preliminary Comments

Virtually all of the information regarding field sobriety tests rests on a foundation of laboratory studies conducted in 1977 and 1981 by the Southern California Research Institute under the auspices of the National Highway Traffic Safety Administration.

Based on the information supplied to me, I find no other laboratory studies which confirm the original findings. Nor do I find any peer-reviewed research which would support or corroborate the NHTSA studies. Nevertheless, I can state that the study design, methodology, and reporting appear to meet requirements for scientific inquiry and have been conducted with care and credibility.

The relationship of laboratory studies to actual use in the field must also be explored. I agree only partially with Marcelline Burns (co-author of the original laboratory studies) and Ellen Anderson in their introduction to the Colorado validation study (Page 1) when they state, "…it should be recognized that the laboratory data are only indirectly enlightening about current roadside use of the

tests." Since laboratory data represents measurement under "ideal" conditions, limitations in the technique which are apparent in the laboratory can only be exacerbated by the uncontrollable variables which occur in the "real world." To this, the Colorado study authors agree: "In particular, note that controlled laboratory conditions are less variable and, therefore, may be less challenging than the highly varied conditions which officers routinely encounter in the field" (Page 1).

With this foundation, let's examine the laboratory data to assess with what degree of confidence, FST results, under the most ideal conditions, can be viewed as reliable and valid predictors of blood alcohol concentration.

Reliability

As stated, this is the index of stability in a test score. Without sufficient reliability, validity is impossible because different inferences are likely to be drawn under what should be the same conditions. In other words, any differences are the result of error variance, rather than valid variance. Reliability establishes an upper limit for validity.

Even under controlled laboratory conditions, the use of field sobriety tests does not appear to meet generally accepted scientific standards. The inter-rater reliability regarding arrest/no arrest decisions is .59. This estimate of reliability is even lower than that of the FST results themselves. This makes sense in that the raters are obviously incorporating additional, non-standardized information into their decisions. Thus, test score alone is not accounting for arrest/no arrest decisions. Even raters chosen for the laboratory studies are making decisions using data outside of FST results. This use of additional, non-standardized or tested data is likely even more pronounced by the wider range of officers in actual field conditions. These officers are thus more likely to present FST results as "proof" of their arrest decisions, even though they are basing their decision on other factors.

The same difficulties with reliability are demonstrated with test/re-test reliability estimates. In this case, the same subject who has consumed the same amount of alcohol is tested again. These differences directly translate into roadside situations where factors other than BAC impact the individual's ability to perform on field sobriety tests. The researchers measured test/re-test reliability under two conditions: having the same officer make the evaluation on the person at a different point in time, and having two different officers (1981, Page 35). The test/re-test reliability with the same officer making the decision for the same individual is .77. This reliability estimate, obtained under laboratory conditions, probably represents an optimistic estimate. As such, it certainly does not support any definitive statement regarding an individual's BAC. The results by different officers are even more disturbing. The total FST score achieved by the same subject with the same BAC measured by different officers (.57) is simply not high enough to warrant any precise estimate of an individual's BAC. The authors appear to agree: "Tests/re-test reliabilities for psychomotor tests are typically on the order of 0.7." (Guilford and Fruchter, 1978; 1981, Page 34.)

Review of the 1981 studies indicates that the reliability for arrest decisions (Page 35) is substantially higher for different officers observing the same subject under the same BAC. Thus, an arresting officer's contention that an individual's BAC is over the legal limit is clearly incorporating other information. Based upon the laboratory data, it is likely that the basis upon which the officer is making such a claim lies well beyond FST results and is thus not subject to scientific inquiry or

proof. This has tremendous implication for the actual administration of FSTs in the field. It suggests that different officers administering the same tests are likely to achieve quite different outcomes, depending upon other, non-testable factors.

Validity

Reliability is a necessary, but not sufficient, condition for validity. The question remains as to the accuracy of field sobriety tests. This represents an error rate of nearly 50%, comparable to deciding whether a person should be arrested by flipping a coin. The 1977 study shows 47 of 101 arrest scores to be inaccurate based upon the criterion of BAC equal to or greater than 0.10% (Page 25).

A large proportion of these "false alarms" (incorrect arrests) occurred in the 0.08% - 0.10% category. However, mistaken arrests range from .054% to .096% (Page 36).

The authors minimize these findings by explaining that, in the field, officers more typically arrest drivers with higher BACs. While this data appears to be supported by nationwide demographic research, "the average BAC of those arrested for DWI across the United States is 0.17%" (1981, Page 19), this may be irrelevant in any particular case. What can be deduced from this finding is that individuals whose blood alcohol count is near the legal limit, but not exceeding it, are most likely to be misclassified as failing the FST. Again, giving any deference to the finding that a failed FST means a BAC above legal limits is simply not warranted by this data. In fact, the 1977 laboratory results indicate six people who would have been arrested even though they consumed no alcohol at all (Page 26).

The 1977 authors admit (Page 41), "Again, it should be pointed out that all the evidence from these data suggests it is unrealistic to attempt to use behavioral tests to discriminate BACs in the plus or minus .02% margin around a given level." They further state (1977, Page 27) that "decision errors occur most often with middle-range levels of intoxication."

Results were somewhat better in the 1981 study, probably resulting from an optimized set of decision rules for the FST. However, results still are not strong enough to support definitive statements of impairment based on FST score. For example, 1981 results are as follows (Page 22):

> Eleven percent of subjects with placebo doses (no alcohol) would be arrested
> Twenty-two percent of subjects having BACs at 0.05% would be arrested

Thus, as BAC approaches, but does not reach, legally-defined limits, the probability of an officer's arrest decision increases dramatically. The number of false positives (incorrect arrest decisions) becomes quite large at BAC levels well below 0.10%.

The issue of validity (accuracy) also can be examined by looking more closely at individual officer performance. This relates directly to the issue of validity by introducing potential unreliability on the part of the officer. If one looks at the 1981 officer group, it varied considerably:

> Experience, 1-19 years
> DWI stops, 5-10,000

The following interesting results emerge. The most accurate officer in terms of correctly arresting people who had BACs equal to or above 0.10% was an officer with 3,500 stops. The least accurate officer was one with 5,000 stops. Thus, street experience alone does not seem to account for accuracy among officers.

Summary

The 1977 and 1981 studies show that even under laboratory conditions, individuals with the same BAC produce different FST results when measured at different times by different officers. Even under these optimal conditions, the error rates for decisions based upon FST results are higher than one would expect or require for a reasonable measure of scientific certainty.

Field Evaluation

Introduction

The situation becomes even more problematic when one attempts to move the inquiry into the field. Unfortunately, the 1981 study's attempt to extend its research to the field did not allow any definitive results. "As a result, trends are reported, but the data are not appropriate for significance testing; the assumption of underlying statistics which would be of interest are not met by the data." (1981, Page 54.)

What is of interest is that the degree of predictive error in the field appeared to be substantially larger than in the laboratory. "For eleven officers for whom we have some data, the average BAC estimate was off by 0.077% before training, and the average BAC estimate was off by 0.0537% after training." (1981, Page 63.) Compare this to the error rate of BAC estimate by the officers in the laboratory study (1981, Page 21). Here, the difference between officer estimate and actual BAC ranged from .0230% to .0344%, averaging about 0.03%. Even after training, officers in the field were far less accurate than officers in the laboratory.

While training clearly brought about improvement, it does not compare favorably to the laboratory condition and is a margin of error substantially higher than one would find acceptable for predicting with any degree of certainty.

Reliability

One of the most disturbing findings from the 1981 field sobriety study is that training did not always appear to "take." "Unfortunately, some officers forgot or ignored most of the administrative procedures, except those associated with nystagmus, by the time of their second post-training ride-along." (1981, Page 70.)

Note that this second ride-along occurred less than one month after training.

The 1981 authors conclude under laboratory conditions, and in the hands of adequately trained personnel, the test battery is a sensitive index of BAC and of impairment (1981, Page 72). However, in answer to the question, "Were officers better able to discriminate 0.10% as a result of using the test battery?" the authors conclude definitive answers to the question cannot be offered (1981, Page 73). They continue, "Major effort is needed for a subsequent field evaluation." (1981, Page 73.)

Subsequent Field Evaluations

Among the documents offered for my review were validation studies conducted in Colorado (1995) and Florida. However, the information supplied to me is not sufficient to classify these findings as studies. They are merely summary reports, without foundation, of findings.

In addition, they suffer from a serious methodological flaw. Given the fact that many, but by no means all, actual DWI stops in the field occur with drivers who are severely impaired, any accuracy data from this research design is likely to be highly inflated. Thus, statements such as "field sobriety

tests are 90% correct" are quite meaningless. While this figure may be true for the average arrestee (BAC equals 0.17%), it may be quite erroneous in any other given situation.

The Colorado and Florida studies, co-authored by an original Southern California Research Institute author, are highly supportive of FSTs. Again, the studies, or the summaries available to me, do not represent peer-reviewed publications. They appear to be conducted under contract to agencies who clearly have a vested interest in a particular outcome. The presence of misleading statements the obtained data (from the Colorado study) demonstrated that more than 90% of the officers' decisions to arrest drivers were confirmed by analysis of breath and blood specimens fails to mention that the criteria for the Colorado study was a blood alcohol count of 0.05% (Page 2). The accuracy figure would be far lower using a criterion of 0.10%.

A 1983 NHTSA technical note evaluated the effectiveness of FSTs in the field. The result, while potentially useful, is not compelling:

> The accuracy of the combined procedure for all police agencies was 83%.
> This accuracy figure ranges from 75% to 96% depending on what agency conducted the tests.
> "Of the misclassifications, 16% involved classification of a driver's BAC as greater than or equal to 0.10% when his/her BAC was less than 0.10%."
> Only 1 percent of misclassifications involve classifying a driver's BAC as less than 0.10% when his/her BAC was greater than or equal to 0.10%.

Using figures from the 1983 study, field sobrieties improved the accuracy of officers, but still resulted in 31 false positives (incorrect arrests) of 200 individuals presented (Page 10). This figure is, however, an exaggerated estimate of FST accuracy. As the authors note, "…in the great majority of the cases, PBT data were available to the officers for a driver before he was arrested. Thus, most arrest decisions were based on PBT data, rather than just test battery data." (1983, Page 9.) Given the fact that virtually all of the misclassifications were false positives, this study demonstrates that there is some unknown probability, higher than 15%, that an FST "failure" would lead an officer to an incorrect assumption that the driver's BAC was equal to or greater than 0.10%.

The use of standardized FSTs appears to increase officers' confidence and make them more likely to arrest drivers who, using the 0.10% criteria, should not be arrested.

The final conclusion, "The results of the field evaluation indicate that the test battery appears to be about as effective as the use of PBTs in improving the BAC distribution of those arrested (e.g., a reduction of false positives)" (Page 11), clearly puts the accuracy of field sobriety tests on par with preliminary breath testing devices (PBTs). My understanding is that PBT results are notoriously unreliable and are therefore not admissible in court proceedings.

<u>Cole Article</u>

The article, "Field Sobriety Tests: Are They Designed for Failure?" by Cole and Nowaczyk represents the only peer-reviewed document available for my review. Their study was designed to "…test the hypothesis that sober individuals will find the field sobriety tests difficult to perform and,

as a result, will be judged to be impaired by officers viewing their performance." (Page 100.)

All of the subjects in the Cole and Nowaczyk study had BACs of 0.0. They were then asked to perform two of the three standard FST procedures. Unfortunately, the authors did not use the horizontal gaze nystagmus test because it did not lend itself to videotape review. This means that one cannot completely transfer findings from this study to the field situation.

The results, however, are quite startling. Out of 21 subjects, only three individuals were rated as "unimpaired" by all officers on both the field sobriety and normal-abilities tests (Page 102). "Forty-six percent of the officers' decisions were that an individual had 'too much to drink' from viewing the field sobriety tests."

These were individuals who had BACs of 0.0. Clearly, a finding of failure to perform adequately on two of the standardized field sobriety test battery with no alcohol in one's system seriously undermines the confidence in FSTs as a predictor of alcohol impairment.

The authors' conclusion, "Even without alcohol, the number of errors made by individuals performing the field sobriety tests was sufficient for officers to judge that the individuals had had too much to drink." (Page 103.) "The fact that these tests require unfamiliar and unpracticed motor sequences may put an individual at a disadvantage when performing them." (Page 103.)

Officer Confidence

There is also an issue regarding officer confidence and FST results/arrest decisions. The Florida study states, "Experience and confidence have a direct bearing on an officer's skill with roadside tests." (Page 3.) The student manual for DWI detection and standardized field sobriety testing makes repeated assertions regarding the validity of FSTs: "Your first task in Phase Three is to administer three scientifically validated psychophysical (field) sobriety tests." (Page VII-I.) "The most significant psychophysical tests are the three scientifically validated structured tests that you administer at roadside." (VII-I.) "Walk-And-Turn is a test that has been validated through extensive research sponsored by the National Highway Traffic Safety Administration (NHTSA)." All of these clearly are designed to give the arresting officer confidence that these procedures will be an accurate measure of the arrest/don't arrest decision. This confidence, however, might be compelling in a courtroom, but nonetheless is not supported by the evidence.

Finally, the Florida authors appear to have a vested interest in squelching the legal controversy which appears to plague their findings:

> "For more than a decade now, however, defense counsel in many jurisdictions has sought to prevent the admission of testimony about a defendant's performance of the three tests." (Page 3.)
> "Since it seems unlikely in the extreme that they [traffic officers] would continue to rely on tests which repeatedly lead to decision errors, it is a reasonable assumption than more often than not their roadside decisions to arrest are supported by measured BACs." (Page 3.)
> "If, on the other hand, it can be shown that officers typically making correct decisions, based on the SFSTs, perhaps the legal controversy that has centered on them for more than a

decade can be diffused and court time can be devoted to more substantive issues." (Page 5.) And finally, "There appears to be little basis for continuing legal challenge." (Page 6.)

It is understandable that the authors have a stake in putting legal controversy around the accuracy of FSTs to rest. Unfortunately, the evidence which I was able to review would clearly indicate that more research is required before any definitive statement can be made regarding FSTs' predictive accuracy.

CONCLUSION

After almost 25 years of use, the debate regarding the accuracy of FSTs continues. Based upon review of the documents available to me, I can draw the following conclusions:

The laboratory studies which form the foundation for FST use appear to be well-designed.
The accuracy of FSTs, even under laboratory conditions, is less than desired or expected for measures of this type.
The field studies available for my review were not well documented and produced unknown error rates that are likely to be unacceptable in real world situations.
The error rate of FSTs in the field as actually conducted by police officers is unknown.
The one article subject to peer review is highly critical of FST accuracy.
The issue of general acceptance by the scientific community is unanswerable given the information provided to me. The refereed article and the letter by Dr. Yale Caplan would appear to indicate that at least these members of the scientific community do not give FST results the weight of scientific proof.

In conclusion, it would appear that FSTs represent a useful tool in a traffic officer's armamentarium. They would serve as a helpful preliminary indicator that further inquiry is required to ascertain driver impairment due to alcohol. They were neither designed nor seem to support, without other stronger data, the contention that an individual is legally impaired.

I declare under penalty of perjury that the foregoing is true and correct to the best of my knowledge.

Executed on: November 7, 2001

Harold P. Brull
Sr. Vice President
Personnel Decisions International
45 S. 7th St., Suite 2000
Minneapolis, Minnesota 55402
612/337-8233