Field Sobriety Tests: Percentages of Percentages, Why "Validation" Studies Fail to Validate

By Greg Kane, M.D.

[Author's Note: In this issue *Trial Talk* begins a short series of articles about the scientific interpretation of field sobriety tests. I am not an attorney, so the series is not about legal strategy or analysis. I own a med mal consulting company, and I have professional experience in the math and science of interpreting imprecise physical tests. These articles are about how that science applies to the field sobriety tests used in DUI prosecutions.]

My guess is right now you are licking your finger, reaching for the next page. You do not see how that premise could be worth your time. Stick around. I will take you places you never imagined, and show you things that will amaze. The practical science you will learn about applies not only to field sobriety tests, but also to any Yes/No scientific test you will confront in your practice.

FSTs are coordination tests police officers use to check drivers for alcohol impairment. Drivers stand on one leg. They walk and turn. The officer looks to see if their eyes are jerky. The idea is that because alcohol makes people uncoordinated, testing for incoordination is a way to test for alcohol impairment. Rules vary from state to state, but courts generally admit FST results as evidence of alcohol impairment in DUI prosecutions.

Standardized FSTs were invented back in the 1970s, when the National Highway Traffic Safety Administration paid a psychologist to come up with a test to spot drunk drivers - a noble goal. Unfortunately, the first study discovered the FST's "Yes" answer was correct only sixty-nine percent of the time, and about half the "arrested" drivers were innocent. NHTSA and its psychologist tried again. And again. Eventually, after favorably adjusting the proportion of impaired drivers in the study group, after analyzing not FST results but officer arrest decisions, and after imagining, against strong evidence, that the officers' arrest accuracy was due to the FSTs they did, NHTSA validation contractors were finally able to claim FSTs work.

> Breath or blood specimens confirmed that 93% of the arrested drivers were above 0.05% BAC. . . . It is concluded that the SFSTs are valid tests. . . .[1]

Defense attorneys I have talked to complain about an incestuous relationship between the NHTSA and its scientists, and about the sloppiness of the un-peer reviewed FST validation research. Around the country, DUI defense attorneys form organizations, give seminars, write articles and share trial strategies to impeach FST evidence. Prosecutors do the same, from the other direction.

The two sides bicker about mechanics. Did the officer follow procedure exactly? Did the officer consider medical conditions that cause incoordination? What they do not do, as far as I can tell, is doubt the NHTSA contractors' analysis of what a mechanically meticulous coordination test actually implies about alcohol impairment. The driver failed the FST. No one asks, "Exactly what does that mean?"

See, physical tests are not perfect. This is particularly true of tests like FSTs, which measure one thing, incoordination, as a stand-in for another thing, alcohol impairment. You do not need me to tell you alcohol affects different people differently. Coordination level and alcohol level are related - but only loosely, not exactly. So even when it is done with laboratory precision, an FST's "Yes = uncoordinated" answer really only means "**maybe** alcohol impaired."

For science, loose-connection tests

like this are old news. So, to tell how close to certain "maybe" is, scientists have developed a well established, generally accepted technique—a simple formula—that takes as input the fundamental accuracies of a test (in our case the FST), and gives as output . . . a number. That number is a thing called the "positive predictive value," the PPV. For FSTs, the PPV number tells you the probability that a driver with a positive FST was actually impaired.

In this series, I will show you the scientific PPV formula, and I will show you the results you get when the formula is applied to field sobriety tests and DUI defendants.

The first thing you will learn is how NHTSA contractors analyze FSTs, and why their method does not work. It gives silly answers. NHTSA contractors do not use PPV science; they use a technical mathematical statistic they call "accuracy." You will see that when this so called "accuracy" statistic is applied to the NHTSA's own Colorado Validation Study data, it validates a coin-toss as 82% accurate. **A coin-toss is 82% accurate?** That is crazy! The statistic is flawed. The flaw is pernicious. The flaw opens the door for research results to be manipulated. Simply by adjusting the balance of impaired and sober drivers in their study group, NHTSA contractors can dial in the "accuracy" their research discovers.

In the next issue you will see that the standard scientific PPV formula, applied to the NHTSA's own data, proves FSTs have no meaningful power to tell impaired from sober. FSTs do not work.

"Well," you say, "if FSTs do not work, why do all those NHTSA funded studies discover they are extremely accurate?" In a later article, you will discover that the high so called "accuracies" claimed by NHTSA contractors are phony. The way studies "discover" these impressive accuracies is to use statistical tricks. I will show you exactly how it is done.

Once you know what the statistical tricks are, you can remove them, and see what validation studies actually discover. Your jaw will drop.

This series is a basic introduction, not scientific analysis. I want you to see how this stuff works, so in a few places, I have chosen beginners' simplicity over scientists' precision.

To learn more about how science uses the famous PPV formula, for dozens of handy tutorials and thousands of references in the scientific literature, web search "'positive predictive value' prevalence," or "sensitivity specificity predictive value," etc.]

The Problem

You are about to discover a critical problem with roadside sobriety tests. The so called "accuracy" of these tests is actually one of several technical mathematical statistics that are valid only for the group of drivers in the validation study. For any driver in any DUI case you will ever work on, the so called "accuracy" number is wrong. Way wrong.

Roadside Sobriety Tests

Standardized field sobriety tests were invented in the 1970s, by National Highway Traffic Safety Administration (NHTSA) contract researchers. Since then, as legal circumstances change, the NHTSA has repeatedly paid researchers to re-"validate" roadside sobriety tests. NHTSA contractors generally do not say precisely what they think "validate" means: "The purpose of the validation study was to assess the validity of the cues..."[2]

They seem to have in mind some general notion of "accuracy," or correlation, or both: "Breath or blood specimens confirmed that 93% of the arrested drivers were above 0.05% BAC. . . . It is concluded that the SFSTs are valid tests; i.e., they serve as indices of the presence of alcohol at impairing levels."[3]

What you may miss, reading NHTSA reports, is that like other scientific tests, roadside sobriety tests have several mathematical accuracies. There is the accuracy of

1. The test's "Yes" answers = Officers' arrest decisions
2. The test's "No" answer = Officers' release decisions
3. The test done on sober drivers = the Innocent Driver Accuracy
4. The test done on impaired drivers = Impaired Driver Accuracy
5. The test's "overall accuracy"

You cannot understand roadside sobriety tests unless you understand that these "accuracies" are not the accuracy of your day to day intuition. They are mathematical things. They behave in ways that are weird and non-intuitive.

For one thing, the five accuracy numbers are all different. The Colorado Validation Study did not just discover that roadside sobriety tests are 93% accurate. It discovered that roadside sobriety tests are 93% accurate **and** 89% accurate **and** 86% accurate **and** 76% accurate **and** 64% accurate.

Which accuracy did NHTSA contractors report? Ninety-three percent sounds pretty good; contractors put it at the front of the report, between big, bold "look here" lines.[4] Seventy-six percent does not sound so good. I had to calculate it myself. NHTSA contractors left it out of their report.

Here is another thing. The accuracies NHTSA validation contractors claim validate roadside tests are **group dependent**. They vary with - go up and down with - the percentage of drivers in the study group who are impaired.

How that happens is easy to see. These so called accuracies are triple percentages - percentages **applied to percentages**.

Roadside sobriety tests have two fundamental accuracy percentages - the Innocent Driver Accuracy and the Impaired Driver Accuracy. If officers

do roadside sobriety tests on **innocent** drivers, the test answers will be correct a certain percentage of the time. The same goes for **impaired** drivers; the test answers will be correct a certain, but different, percentage of the time.

NHTSA validation contractors come up with their so called "accuracies" by applying the roadside test to one particular group of drivers, the drivers in the validation study group. The contractors' so called "accuracy" formula passes the roadside test's two fundamental accuracy percentages through the impaired driver percentage of that one particular study group. The so called "accuracy" the research discovers is highly dependent on that third percentage.

Simply by adjusting the balance of impaired and sober drivers in the study group, NHTSA researchers can dial in the accuracy they want.

Let me show you how it works.

Group dependence of the NHTSA's so called "accuracy"

Let us work with the NHTSA's "overall accuracy" - a statistic FST proponents bring up as a way to claim roadside sobriety tests are "extremely accurate." Here is our exercise. Using exactly the same formula NHTSA contractors use when they "validate" FSTs, I will have you "validate" an FST for several groups of drivers. To make sure we separate how the triple-percentage statistic works from how real world FSTs work, you will "validate" an innovative new FST.

The National Highway Traffic Safety Administration pays you to develop and validate a new Time Efficient Field Sobriety Test. The test will determine quickly, yet scientifically, whether drivers are impaired. Are they impaired? Are they sober? Your Time Efficient FST will tell.

Here is your test. The driver flips a coin. If the coin comes up heads, the test says the driver is impaired. If the coin comes up tails, the test says the driver is impaired. Got that? The driver flips a coin, no matter which side comes up, the Time Efficient FST says the driver is impaired.

The Time Efficient FST has two fundamental accuracies. On impaired drivers it is 100% accurate. On innocent drivers it is 0% accurate. Is this Time Efficient FST stupid? Yes it is. But does its NHTSA "accuracy" statistic prove it is "valid"? Your study will find out.

Your plan is simple: identify a group of drivers, apply your Time Efficient FST to each driver in the group, recording the FST result and the driver's BAC (Blood Alcohol Concentration). Then calculate the mathematical "overall accuracy" statistic of the Time Efficient FST exactly the way NHTSA validation contractors calculate the so called "overall accuracy" of their roadside sobriety tests—divide the number of correct roadside sobriety tests by the number of drivers in who took the test.

Here we go.

Trial 1

You start validating your Time Efficient FST by studying your cousin's Latter Day Saints Sunday school class. LDS members do not drink. You test ten class members. They all flip either heads or tails; the test says they are all impaired.

Your test recorded zero correct answers. 0/10 = 0%. Your test is 0% accurate.

Trial 2

Next you travel to a state where they set up checkpoints to stop and test every driver coming down the road. On Monday morning one percent of drivers on the road are impaired. You test one hundred drivers. They all flip either heads or tails; your Time Efficient FST says they are all impaired.

Your test recorded one correct answer. 1/100 = 1%. Your test is 1% accurate.

Saturday night you go to the checkpoint in the parking lot of the *Bikerz Blitz* social club. Half these drivers are impaired. You test one hundred drivers.

They all flip either heads or tails - your Time Efficient FST says they are all impaired.

Your test recorded fifty correct answers. 50/100 = 50%. Your test is 50% accurate.

Trial 3

You decide to "validate" your Time Efficient FST using exactly the data reported in *A Colorado Validation Study of the Standardized Field Sobriety Test (SFST) Battery*. The study's published report gives raw data collected on each driver. This widely cited study found that officer arrest decisions were "93% accurate."

In the Colorado study NHTSA contractors analyzed field sobriety tests on 234 drivers, 191 of whom were later shown to have blood alcohol concentrations greater 0.04% (the BAC FSTs are now claimed to spot).

Conveniently, your Time Efficient FST always gives the same answer, so you simply do a virtual test on all 234 drivers. They all virtually flip either heads or tails—the test says they are all impaired.

Your test recorded 194 correct answers. 194/234 = 82%. Your test is 82% accurate.

As you put it in your validation study report, "Using only the round, flat, standardized Time Efficient Field Sobriety Test Disk – with 'Guilty!' on each side . . . officers seldom erred when they decided to arrest a driver. Breath or blood specimens confirmed that 82% of the arrested drivers were above 0.04% BAC."

Did you catch that? Using exactly the statistical formula NHTSA validation contractors use, you just "validated" a coin toss as 0% accurate **and** 1% accurate **and** 50% accurate **and** 82% accurate. You kept the fundamental accuracies of your test unchanged. Instead, simply by adjusting the balance of impaired and sober drivers in your study group, you were able to dial in the 'accuracy' you wanted.

The NHTSA contractors' mathematical "accuracy" formula is a triple percentage. It depends not just on the fundamental properties of the roadside sobriety test, but also on the percentage of impaired drivers in the study group. The result is the formula gives different answers, depending on the balance of impaired and innocent drivers picked for the study. Simply by adjusting the balance of impaired and sober drivers in the study group, NHTSA contractors can dial in the "accuracy" their research discovers.

For example, in the Colorado Validation Study NHTSA contractors chose study drivers in a way that let them discover an officer arrest accuracy

of 93%. But, keeping the fundamental accuracies exactly the same and simply reversing the percentage of impaired and innocent drivers chosen to be studied would have changed the officer arrest accuracy to 52% - a coin toss.

Intuition-wise, and implication of guilt-wise, the NHTSA contractors' mathematical "overall accuracy" and "officer arrest decision accuracy" statistics are flatly meaningless.

If you say, "Well, the accuracy of officer arrest decisions in the study - the percentage of drivers in the validation study who were arrested, who turned out to be impaired - was 93%. And my client was arrested, so my client's probability of impairment is 93%," you will be wrong. Not a little wrong. Way wrong.

Implications for DUI Prosecutions

The triple-percentage, group dependence property of the NHTSA's so called "accuracy" has several implications.

First, on the level of public policy, it means no study has ever actually validated FSTs. Every day somewhere in our favored land citizens are arrested and eventually convicted on the basis of a "scientific" test whose basic accuracy is unknown to the legal system.

Second, on the level of an individual DUI defendant, it is not just that FST accuracies are actually unknown, it is that they are wrongly taken to be quite high.

It turns out science does have standard methods for solving the group dependence problem and calculating how much power an FST/ officer arrest decision has to identify impaired drivers. The answer, which is the subject of the next article in this series, should astound you.

To learn more, and for thousands of references in the scientific literature, Google "'positive predictive value' prevalence," or "sensitivity specificity predictive value," etc.

Greg Kane, MD owns Med-mal Experts, Inc., (www.medmalExperts .com) an Englewood based consulting firm that reviews medical malpractice claims and refers attorneys across America to physician expert witnesses in all specialties. Contact: 303-741-0993 FST@medmalExperts.com

Endnotes

[1] Marcelline Burns and Ellen W. Anderson, *A Colorado Validation Study of the Standardized Field Sobriety Test (SFST) Battery*, Colo. Dep't of Transp., 1995, Technical Summary.

[2] *Jack Stuster and Marcelline Burns,* Validation of the Standardized Field Sobriety Test Battery at BACs Below 0.10*., Nat'l Highway Traffic Safety Admin., 1998 at 23.*

[3] Marcelline Burns and Ellen W. Anderson, *A Colorado Validation Study of the*

Standardized Field Sobriety Test (SFST) Battery, Colo. Dep't of Transp., 1995, technical summary.

[4] "Using only the standardized 3-test battery ... officers seldom erred when they decided to arrest a driver. Breath or blood specimens confirmed that 93% of the arrested drivers were above 0.05% BAC." Burns and Anderson, *supra* n. 2.