Field Sobriety Tests:

Two Statistical Tricks Let NHTSA Contractors Prove Any Roadside Sobriety Test Is "Extremely Accurate"

By Greg Kane

As we saw earlier, Field Sobriety Tests (FSTs) do not work.[1] Yet over and over NHTSA contractors report FST accuracy statistics close to 90%. Over and over the same result, always good. How do they do that? Here's how: two statistical tricks let contractors "validate" any roadside test as "extremely accurate."

We saw that roadside sobriety tests have been studied, restudied and studied again, always turning up data proving the same thing - roadside sobriety tests do not work. So, why do people think FSTs work? NHTSA (National Highway Transportation Safety Administration) validation contractors make FSTs appear to work by calculating technical mathematical statistics they call "accuracies." These accuracies make FSTs look good. But we discovered these statistics don't mean what you think they mean. Worse, the "accuracy" numbers go up and down as contractors vary the proportion of innocent and impaired drivers they use to calculate the statistic. Simply by adjusting the balance of impaired and innocent drivers in the study group, NHTSA researchers can dial in the accuracy their study "discovers." The so called "accuracies" are subject to manipulation.

This article shows you how the manipulation works.[2] Two tricks fix validation studies so any FST can come off "extremely accurate." Trick One: Move the FST's inaccuracy over onto the innocent drivers. Trick Two: Hide the innocent drivers.

Why read about an "accuracy" statistic you already know is useless? Well for one thing, it's fun to see how the guy in the glitter suit saws the lady in half. More importantly, unraveling the two tricks uncovers a question you otherwise wouldn't see: *Are the fundamental FST accuracies measured in various validation studies the same or different?* The answer to the question—*different*—proves once again FSTs do not work.

If you spend much time reading FST validation studies, you'll notice a curious anomaly. Over and over the accuracy of roadside sobriety tests on innocent drivers is very different from the accuracy on impaired drivers. For example, the 1998 validation study found an Impaired Driver Accuracy 98%, but an Innocent Driver Accuracy of 71%.[3]

71% and 98%. Why the big difference?

Trick One: Move the Inaccuracy over onto the Innocent Drivers

To see why, you need to know FSTs are not really Yes or No tests. Drivers get scores - points. Points for swaying when they stand on one leg. Points for using their arms to balance. Like that. The more points they get, the worse they did. An FST turns a point score into "Yes" or "No," "impaired" or "innocent," by having a cutoff level. If the driver got more points than the cutoff, the test says *Impaired*. Fewer points, the test says *Innocent*.

So here's a question. What's the correct cutoff level?

The answer is: whatever you want. I mean, NHTSA contractors have picked a cutoff number, but other numbers would work too.

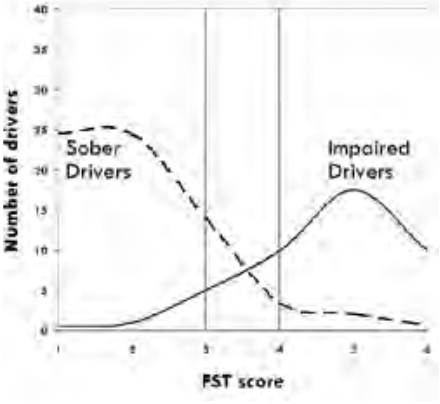
Look at Figure 1. It's a graph of the FST scores of innocent and impaired drivers. Find an FST score along the bottom axis, read the number of drivers, innocent or impaired, who got that score on the vertical Number of Drivers axis at the left.[4]

The vertical lines at FST score 3 and 4 are two possible cutoff lines. If a driver's FST score is above the cutoff, she's arrested. If it's below the cutoff, she goes home.

Say you set the cutoff at 4. Every driver above 4 would be arrested; impaired drivers above 4 would be arrested, innocent drivers above 4 would be arrested. Every driver below 4, innocent or impaired, would go home. Say you set the cutoff at 3. Everyone above 3 would be arrested; everyone below 3

Figure 1.

Sober Driver and Impaired Driver accuracies compete

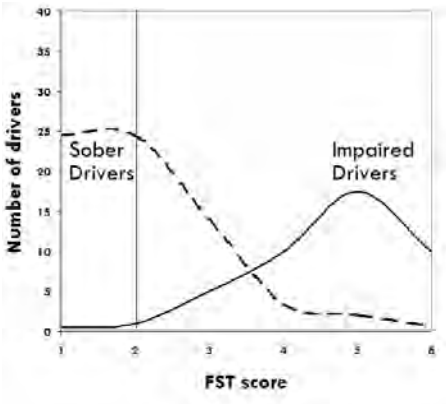


Changing the FST cutoff changes the fundamental accuracies of the field sobriety test.

Compared with FST 4, FST 3 is more accurate for impaired drivers, less accurate for sober drivers.

Figure 2.

How NHTSA contractors set the cutoff

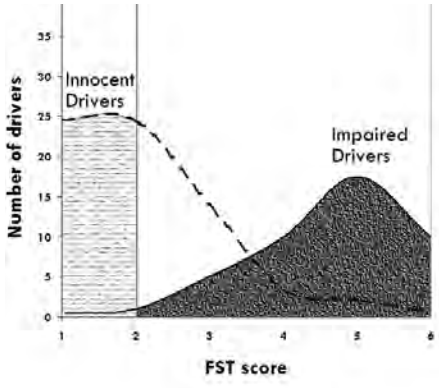


Trick One. Move the inaccuracy onto the Innocent Drivers

Set the cutoff score very low. This makes the Impaired Driver Accuracy very high—by moving the FST's inaccuracy over onto the innocent drivers.[6]

Figure 3a.

"Overall accuracy"

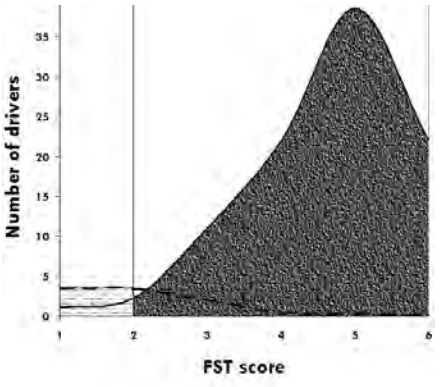


Drag down: the FTS cutoff is set so low that officers arrest 29% of the *innocent* drivers they assess

Trick One creates a validation credibility problem. The Overall Accuracy of the test is the percentage of the-area-under-both-curves that is shaded. Look at the Innocent Driver curve. See the big unshaded area under the dashed line? The unshaded area represents FST errors, innocent people whose mistaken arrests drag the FST's Overall Accuracy down.[8]

Figure 3b.

How to get your FST to "validate" every time



Trick Two: Hide the innocent drivers

Trick Two picks a study group with lots of impaired drivers, and not so many innocent drivers. Hiding the low Innocent Driver Accuracy keeps it from pulling the weighted "Overall Accuracy" down.[9]

would go home. The point of all this is for you to see how changing the cutoff level changes the accuracies of the test, even the single-percentage, group-independent[5] Innocent Driver Accuracy and Impaired Driver Accuracy.

The accuracy of the roadside test on impaired drivers is the percent of the impaired driver lump to the right of the cutoff. Take a minute to be sure that makes sense to you. Everything from here on depends on you getting that.

Compared with the FST with a cutoff of 4, an FST with a cutoff of 3 has more of the impaired driver lump on the right hand side of the line. The FST-3 is more accurate, on impaired drivers, than FST-4. Scientists say, "It is more sensitive."

All right, fine, we'll maximize the Impaired Driver Accuracy by moving the cutoff way down. Yes, but if you do, look what happens to the Innocent Driver Accuracy. That accuracy works the other way. The Innocent Driver Accuracy is the percent of the innocent driver lump to the left of the cutoff. Take a minute. At FST-4, most of the innocent drivers go home. At FST-3, a large percentage of innocent drivers go to jail.

So, Innocent Driver Accuracy and Impaired Driver Accuracy are in tension. You move the cutoff; one accuracy goes up, the other accuracy goes down. Think of it this way: the FST itself contains some level of inaccuracy. Sliding the cutoff score up and down doesn't change the basic accuracy of the FST (roughly speaking); it just moves the *in*accuracy back and forth between the innocent and impaired drivers.

Look at Figure 2. Now that you understand how changing the cutoff moves the FST's *in*accuracy back and forth between innocent and impaired drivers, you can unravel what NHTSA contractors did when they invented the standardized FST. Do you remember those accuracies from the 1998 validation study? Impaired driver accuracy

98%. Innocent driver accuracy 71%. The NHTSA contractors who invented the standardized Field Sobriety Test *did* set the cutoff level way low. That way their test catches 98% of the impaired drivers. Hurray! But here's the thing, moving the cutoff way down doesn't make the *FST* 98% accurate, it makes the *Impaired Driver part of the FST* 98% accurate. It does that by forcing the test's *in*accuracy over onto the other part of the FST, over onto the innocent drivers. Now the Innocent Driver Accuracy is only 71%. Which means, according the NHTSA, 29% of innocent drivers assessed by police using standardized FSTs are incorrectly arrested for DUI.[7]

Setting the cutoff score very low moves the FST's inaccuracy over onto the innocent drivers. Setting the cutoff score very low is Trick One.

Trick Two: Hide the Innocent Drivers

Trick One has the make-FSTs-look-good advantage of jimmying the roadside test's Impaired Driver Accuracy up to 98%. That pushes the FST's inaccuracy over onto the innocent drivers. For NHTSA validation contractors that Innocent Driver *In*accuracy problem is also a validation credibility problem. The so called "overall accuracy" statistic contractors use to validate roadside sobriety tests is calculated as a combination (a weighted average) of the innocent and impaired accuracies. When the Innocent Driver Accuracy is low, a straight up "overall accuracy" calculation makes roadside sobriety tests look bad—the 71% drags the 98% down.

Figure 3a shows you why drag down happens. Trick One has set the cutoff so low, the solid impaired-driver lump is 98% accurate—that makes FSTs look good. But setting the cutoff this low makes the dashed innocent-driver lump only 71% accurate. That makes FSTs look bad. Not just *look* bad. *Be* bad. NHTSA data indicates officers using standardized FSTs arrest 29% of the *innocent* drivers they assess.

Trick Two hides the Innocent Driver *In*accuracy.

Figure 3b shows how it's done. To get the NHTSA validation contractors' solid + dashed *weighted* average close

Table 1. Reversing the two tricks with data from FST validation studies.

		These studies *did* use two tricks			These studies *did not* use two tricks		
	Study	**1998**[12]	**Florida**[13]	**Colorado**[14] **BAC >0.05**	**1977**[15]	**1981**[16]	**Colorado BAC > 0.10**
1	Impaired Driver % Accuracy	98	96	89	84	64	96
2	Innocent Driver % Accuracy	71	82	76	73	88	56
3	# impaired drivers	214	206	184	101	125	139
4	# innocent drivers	83	50	50	137	316	95
5	In their "accuracy" calculations, NHTSA contractors counted impaired drivers → times as much as innocent drivers.	**2.6**	**4.1**	**3.7**	**0.4**	**0.4**	**1.5**
6	NHTSA's % "Overall Accuracy"	**91**	**93**	**86**	**76**	**81**	**79**
		↓	↓	↓	↓	↓	↓
7*	NHTSA's % "Overall Accuracy" **with Trick Two reversed**	**79**	**85**	**78**	**82**	**71**	**72**
8	NHTSA's % **Officer Arrest "Accuracy"**	**90**	**96**	**93**	**53**	**68**	**76**
		↓	↓	↓	↓	↓	↓
9*	NHTSA's % Officer Arrest "Accuracy" **with Trick Two reversed**	**57**	**56**	**48**	**90**	**93**	**60**
10	% Accuracy of a coin toss	50	50	50	50	50	50

Lines 1 – 5 are the raw data reported in FST validation studies

Lines 6 and 8 are the calculated so called "accuracies" NHTSA validation contractors report their studies discovered

Lines 7* and 9* are these same so called "accuracies," recalculated by leaving the FSTs two fundamental accuracies (Impaired Driver Accuracy and Innocent Driver Accuracy) unchanged, and simply by reversing the ratio of impaired/innocent drivers in each study group.

Here are important facts revealed in Table 1.

1. FST validation studies use the two tricks. (lines 1, 2, 5) The first three studies did use the two tricks: their Impaired Driver Accuracies are high, their Innocent driver accuracies are low; and in each case the validation contractors chose to study many more impaired drivers than innocent drivers.

   The 1977 and the 1981 studies use neither trick.[17] The Colorado BAC > 0.10 results use Trick One but not Trick Two.
2. Only studies using the two tricks discover high "accuracies." Look at line 6, and especially line 8.
3. Removing Trick Two removes the high "accuracies." Look at line 6 → 7, and especially line 8 → 9.

to the solid lump's 98%, you just make the solid lump big and the dashed lump small.[10] You make the solid lump big by adding lots of impaired drivers to your study group. You make the dashed lump small by keeping innocent drivers out of your study group. Over and over, that's just what NHTSA validation contractors do.

Please realize that all Trick Two does is *hide* the Innocent Driver *In*accuracy. It hides the inaccuracy by hiding innocent drivers from the "overall accuracy" calculation. Trick Two does not *fix* the Innocent Driver *In*accuracy. According to the NHTSA, officers using standardized FSTs still arrest 29% of the innocent drivers they assess.

Validation Studies Use the Two Tricks

If you think this analysis is too pat and fantastical to be true, I don't blame you. I'm not asking you to believe me. But please, believe the NHTSA's own data. Table 1 gives you the raw facts, taken directly from NHTSA reports. It's a busy table, but important, so take your time.

NHTSA reports show some studies use the two tricks, some don't. But remember, not all "validation" studies discover FSTs are "accurate." What Table 1 shows - look at line 8 - is that the studies discovering a high FST accuracy were exactly the studies using the two tricks. And the studies that did not use the two tricks failed to discover a high FST accuracy.

So, the two tricks and the high FST accuracies line up exactly. Fine. But is the perfect correlation a coincidence, or do the two tricks cause the high "accuracies"? Here's how to find out. Reverse the two tricks. More precisely, reverse Trick Two. Unhide the innocent drivers, and then recalculate the NHTSA contractors' so-called accuracy statistics. To do that, you just add drivers to the innocent driver lump (keeping the above-the-cutoff and below-the-cutoff proportions the same). This simple arithmetic preserves the accuracy of the FST itself; the cutoff doesn't change, the Innocent and Impaired driver accuracies don't change. The arithmetic just scales the size of the lumps—the number of drivers tested in the study—so you can identify what so called accuracy NHTSA contractors would have discovered if they'd chosen to study a different group of drivers, a group with the innocent to impaired driver ratio reversed.[11]

The results of reversing Trick Two are in line 9. As soon as you reverse Trick Two, in study after study, the "overall accuracy" of the roadside test plummets to a number close to the accuracy of a coin toss. The two statistical tricks are the direct cause of the high accuracies "discovered" in NHTSA validation studies. But for the two tricks, FST validation studies fail to validate.

Two tricks fix validation studies so any FST can come off "extremely accurate."

1. Move the FST's inaccuracy over onto the innocent drivers.
2. Hide the innocent drivers.

Very clever. But it won't fool you anymore. And it brings us to a very big deal…

Removing the Two Tricks Uncovers a Shocking Secret

The height of a building doesn't change. Because building heights don't change, measurements of building heights are reproducible. Every time you measure the height of a building, you get the same answer, something like, say, 60'; 60' 1"; 59' 11". The building is sixty feet high, give or take an inch. If you were to measure a building three times and get three different answers: 47', 63' and 82', you'd know your measurement test was giving answers that were wrong, answers you couldn't trust. You'd know your test didn't work.

Real scientific tests, whether they measure the height of buildings or the alcohol level of drivers, give answers that are *reproducible*. You measure the thing over and over; you get answers close to each other. Or you know the test can not be trusted. It does not work.

Over and over NHTSA validation contractors have studied the accuracy of FSTs, over and over reporting percent "accuracies" close to the low nineties. Those similar accuracies make people think roadside sobriety tests are reproducible. The reproducibility makes people think FSTs work. How else do all the studies reproduce the same answer? Well, now you know how they all reproduce the same answer. They use the two tricks. The low nineties "accuracies" are phony. Which has got to make you wonder, do validation studies really find the same answer?

No, they don't.

Look at Table 2. It's a list of roadside sobriety tests' fundamental accuracies, the single-percentage, group independent, Impaired Driver Accuracy, and Innocent Driver Accuracy, discovered by various validation studies.

Notice the Impaired Driver Accuracies are all similar and all high; that's Trick One.[19] Now look at the Innocent Driver Accuracy. It bounces up and down. The NHTSA asks you to believe the Innocent Driver Accuracy of their roadside sobriety test is 56%, 63%, 71%, 76%, 79% and 82%. It asks you to believe the impossible.

The Acme Building Height Company is pushing a new scientific standardized Building Height Test. Company validation contractors have "validated" the BHT with several pseudo-scientific studies manipulating the so called "accuracy" of the BHT, but a glance at the fine print shows you the BHT measured the height of the building as 56 feet, and 63 feet, and 82 feet.

Do you believe Acme's validation contractors? Is the standardized BHT "extremely accurate"? Of course not.

Each time the BHT "measures" the height of a building, it gives a different answer. The BHT's answers are not reproducible. They can not be trusted. The BHT does not work.

Roadside sobriety tests work the same way. Sure, the several alcohol levels of a group of drivers is more abstract than the height of one building, but to be a meaningful scientific test, to give answers that can be trusted, a roadside test must have the simple property of reproducibility. A roadside test must have a reproducible Innocent Driver Accuracy.[20] One accuracy. Not 63% and 82%.[21] The opposite is also true: if the FST is not reproducible—if NHTSA contractors measure it over and over, and repeatedly measure different accuracies, then the FST does not give meaningful, trustworthy, answers about driver impairment; it does not work.

Over and over NHTSA validation contractors have measured the FST. Over and over the accuracies they measure are different. The FST is not reproducible. The FST's answers can not be trusted. Roadside sobriety tests do not work.

Table 2. The FST's fundamental accuracies, measured in various validation studies.[18]

Study	Impaired Driver	Innocent Driver
1998 Walk and turn	92	47
Colorado BAC > 0.10	96	56
1998 One Leg Stand	92	59
1998 H. G. Nystagmus	98	63
Colorado BAC > 0.08	94	63
1988 Officer Decision	98	71
1977 FST	84	73
Colorado BAC > 0.05	89	76
Colorado BAC > 0.04	86	78
1977 Officer Decision > 0.08	76	79
Florida BAC > 0.08	96	82

The Impaired Driver Accuracies are all high – that's Trick One, but the Innocent Driver Accuracies go up and down. Is the accuracy 63%, or is it 82%?

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

End Notes

1 Greg Kane, M.D., Field Sobriety Tests: How Basic Science Proves They Have Little Power to Tell Impaired From Sober, TRIAL TALK (October/November 2006).

2 Nothing in this article is a statement about the knowledge or intentions of the NHTSA or its contractors. This article is about the mathematics of the two *mathematical* tricks.

3 Jack Stuster and Marcelline Burns, *Validation of the Standardized Field Sobriety Test Battery At BACs Below 0.10 Percent*, 1998, at 18, fig. 4.

4 The graphs in this article are conceptual, not exact. NHTSA validation contractors don't release these results in validation studies.

5 These important terms are explained Kane, *supra*, n. 1.

6 These curves are conceptual. NHTSA contractors keep the actual data out of their validation study reports.

7 Stuster & Burns, *supra*, n. 3.

8 These curves are conceptual. NHTSA contractors keep the actual data out of their validation study reports.

9 These curves are conceptual. NHTSA contractors keep the actual data out of their validation study reports.

10 Scaling the size of a lump does not change the accuracy of the test on the group of drivers in that lump. In that group, the proportion of drivers on each side of the cutoff stays the same.

11 The principle is explained in, Kane, Greg, *Field Sobriety Tests: Percentages of Percentages, Why "Validation" Studies Fail to Validate* , TRIAL TALK, August/September 2006.

12 Stuster & Burns, *supra*, n.3.

13 Marcelline Burns, *A Fla. Validation Study of the Standardized Field Sobriety Test (S.F.S.T.) Battery* NHTSA, U.S. Department of Transportation, 1997, data from Figure 5.

14 Marcelline Burns and Ellen W. Anderson, *A Colo. Validation Study of the Standardized Field Sobriety Test (SFST) Battery,* Final Report Submitted to Colo. Dep't of Transp. & NHTSA, 1995, data from Appendix IV.

15 Marcelline Burns and Herbert Moskowitz, *Psychophysical Tests for DWI Arrests*., DOT-HS-802-424, NHTSA 1977, data at 32, unlabeled table.

16 V. Tharp et al., *Devel. & Field Test of Psych. Tests for DWI Arrests*, DOT-HS-805-864, NHTSA,1981, data from Table 8.

[17] Actually, both the 1977 and the 1981 studies inadvertently used the two tricks in reverse. You might like to puzzle through the implications and explain why line 8's 68% → 93% in line 9.

[18] Burns and Moskowitz, *supra*, n. 15; Burns and Anderson, supra n. 14; Burns, supra n. 13; Stuster & Burns, *supra*, n. 3.

[19] Notice the low Impaired Driver Accuracy in the 1977 study. This accuracy is reproducible only if contractors use Trick One. In 1977 contractors didn't use the two tricks, so their study failed to "validate" FSTs.

[20] Impaired Driver Accuracies are generally close to each other; that "reproducibility" is caused by Trick One.

[21] The accuracies might theoretically be different at different BACs, i.e. impairment levels. No NHTSA report I've found supports this hypothesis in the relevant BAC range; in fact NHTSA validation contractors are anxious to refute it - "revalidating" FSTs at ever lower BACs. What's more, as Table 2 proves, even within a single BAC level, the Innocent Driver Accuracies still vary widely. The different-BAC-different-accuracy hypothesis fails to fix the not-reproducible problem.

I haven't mentioned "confidence intervals," which you probably know as "margin of error." A 79% accuracy and an 80% accuracy are within the margin of error. Is a 63% accuracy within the margin of error of an 82% accuracy? The analysis is beyond the scope of this article.

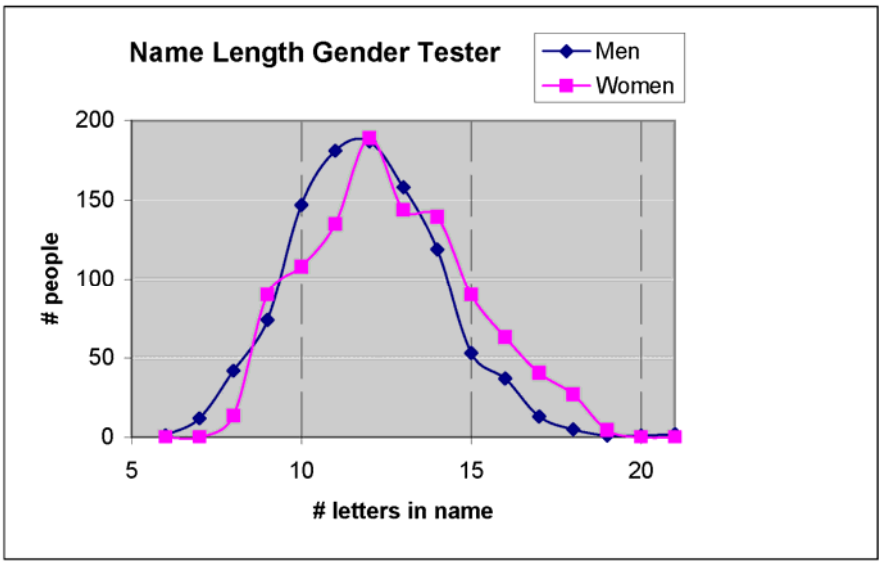
Using the Two Tricks to develop a *standardized Field Gender Test*.

Let's use real world data to see how the NHTSA's Two Tricks can make a silly test look "extremely accurate."

The NTHSA had decided it needs a scientific roadside test that will let police officers determine the gender of drivers they stop. The agency has paid you to develop the test, and then to "validate" the test you developed.

You decide to base your test on the number of letters in the driver's first and last names. Using your scientific test, officers will determine a driver's gender simply by counting the number of letters in the driver's name. Of course to "validate" your test, you'll need to do an NHTSA sanctioned scientific research study.

Your "validation" study gathers data on 1,033 men and 1,044 women. (The data here are real.) Your company scientist graphs the results as follows:



men	1033
women	1044
ratio:	1.0

Test cutoff	Test accuracy
6	**50%**
10	49%
15	45%
20	50%

FIG B1
To read this graph, select the number of letters in the name on the X axis. The two curves show you the number of men and women in the "validation" study group who had that number of letters in their name. Read that number off the Y axis.

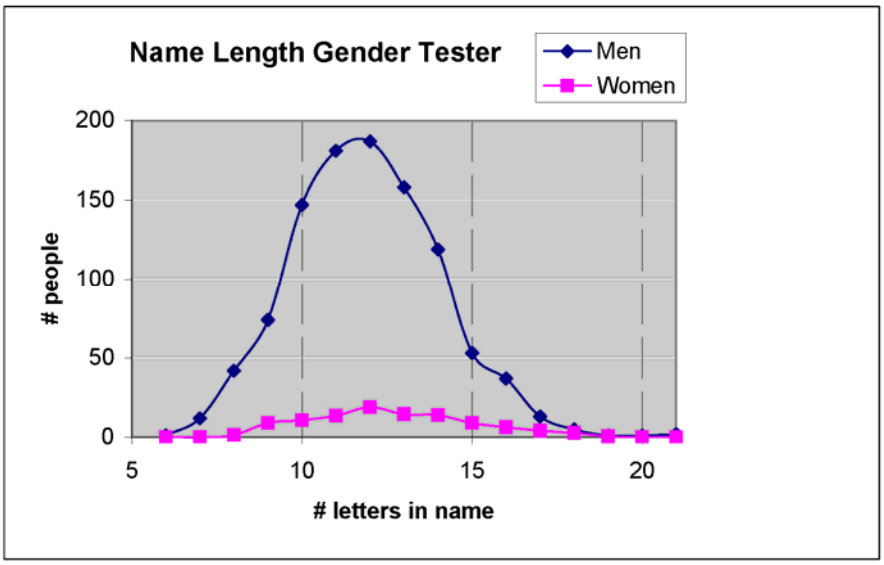
At the right, in blue and pink, are the number of *men* and *women* in the FGT "validation" study group. The *ratio* is the ratio of men to women your scientist chose to include in the study group.

The accuracy table shows the results of your scientific research study so far. If the cutoff is set at 6 letters, the test "accuracy" is 50%. If the cutoff is 15, the "accuracy" is 45%. Etc.

The FGT works as follows: your scientist picks a cutoff score—any score. Drivers with that many or more letters in their name are *men*—according to the FGT test. Drivers with fewer letters in their name are *women*—according to the FGT.

As things stand now, your study reveals a 50% accuracy that is just not good enough for government work. Your scientist rethinks her study protocol and performs a second scientific study...

In the second phase of your "validation" study, your scientist carefully selects a study group with 1,033 men and 109 women [Trick Two]. The results are as follows:



men	1033
women	104
ratio:	9.9

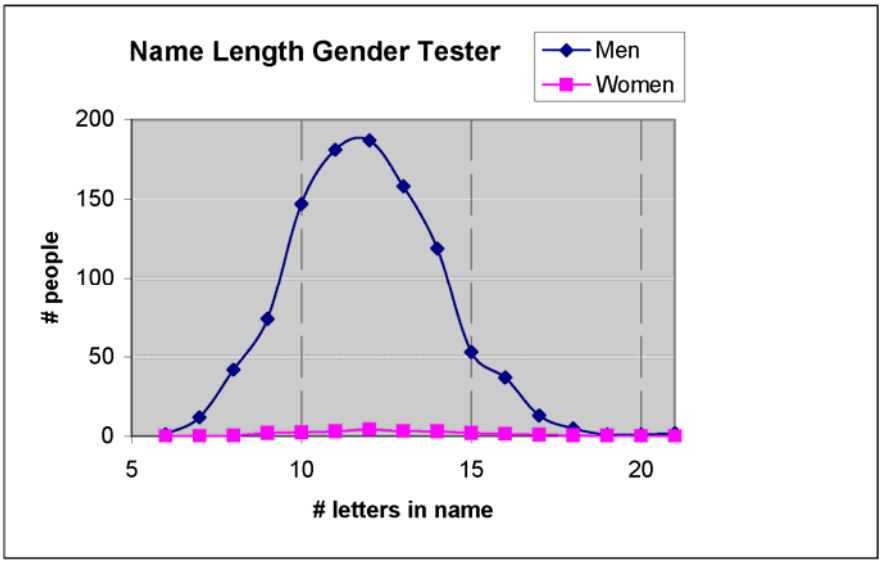
Test cutoff	Test accuracy
6	**91%**
10	80%
15	17%
20	9%

FIG B2

You again select a cutoff score of 6 letters [Trick One]. Now look at the accuracy table. Congratulations, your scientific "validation" study has proven that using only the number of letters in a driver's name, the NHTSA's new scientific standardized FGT identifies the gender of drivers with 91% accuracy!

Your success depends on exactly the same Two Tricks real world NHTSA "validation" contractors use to "validate" FSTs.
Trick one: set the cutoff way low.
Trick Two: study many more target subjects (men / impaired drivers) than non-target subjects (women / innocent drivers).

Using these tricks you can dial in the "accuracy" you'd like your study to "discover." You can even prove the FGT is "*highly* extremely accurate:"

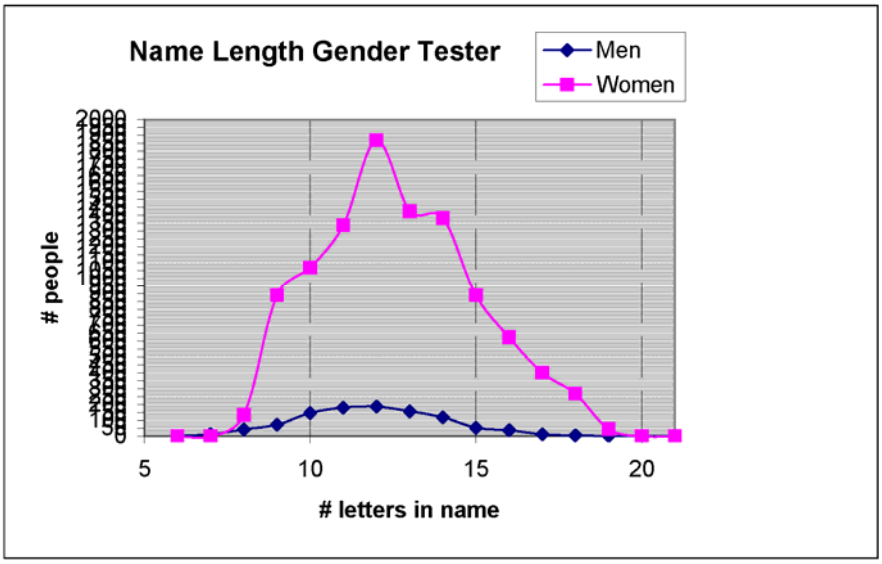


men	1033
women	23
ratio:	44.5

Test cutoff	Test accuracy
6	**98%**
10	86%
15	12%
20	2%

FIG B3

Or, if the NHTSA decides to pay to prove the FGT does *not* work, you can do a scientific study certain to prove that result:



men	1033
women	10324
ratio:	0.1

Test cutoff	Test accuracy
6	**9%**
10	17%
15	72%
20	91%

FIG B4

The so called "accuracy" of the standardized FST is not science. It is an amateur parlor trick.
Now you know how the trick is done.