

Koehler, J. J. (1997). One in Millions, Billions and Trillions: Lessons from People v. Collins (1968) for People v. Simpson (1995). *Journal of Legal Education*, 47, 214-223.

**ONE IN MILLIONS, BILLIONS AND TRILLIONS:
LESSONS FROM PEOPLE V. COLLINS (1968) FOR
PEOPLE V. SIMPSON (1995)**

Jonathan J. Koehler¹

April, 1996

¹ Jonathan J. Koehler (Ph.D. 1989, University of Chicago, Department of Behavioral Sciences) is Associate Professor of Behavioral Decision Making in the Graduate School of Business at the University of Texas at Austin (CBA 5.202, Austin, TX 78712, e-mail: koehler@mail.utexas.edu). He also teaches "Probability and science in the courtroom" at the University of Texas Law School. This paper is based on a presentation given at the American Association of Law Schools Annual Meeting (Section on Evidence) in San Antonio, TX in January, 1996.

On June 18, 1994, newspaper headlines across the world trumpeted the arrest of sports superstar O.J. Simpson in Los Angeles for the murder of Nicole Brown Simpson and Ronald Goldman. After a contentious nine month trial, the jury took only three and a half hours to find Mr. Simpson not guilty of both murders.²

For years to come, students will enter law school with knowledge of the basic facts of the Simpson case, and opinions about the incriminating value of the DNA evidence that linked Mr. Simpson to the crime scene. But few among them will know the significance of the events that took place just a few miles from the Simpson crime scene exactly thirty years earlier.

People v. Collins (1968)

On June 18, 1964, Juanita Brooks was pushed to ground as she walked down an alley in the San Pedro area of Los Angeles. According to Mrs. Brooks, a blonde-haired woman dressed in dark clothing grabbed her purse and ran away. John Bass, who lived at the end of the alley, heard the commotion. He too saw a blonde-haired woman wearing dark clothing run from the scene. He also noticed that the woman had a ponytail, and that she entered a yellow car that was driven by a black man who had a beard and a mustache.

Armed with this description, the police identified Janet and Malcolm Collins as suspects and, eventually, charged them with the robbery. At trial, the prosecution had some difficulty establishing that Janet and Malcolm Collins were the people who committed this crime. The victim, Mrs. Brooks, could not identify either defendant. The witness, Mr. Bass, identified Malcolm Collins as the driver of the get-away car, but his testimony was weakened by his

² Linda Deutsch, O. J. Jury Reaches Decision in Less Than 4 Hours, AUSTIN AMER. STATES. A1 (October 3, 1995). The jury spent two hours and twenty minutes in morning deliberations, and concluded after listening to a one hour and ten minute re-reading of testimony from limousine driver Allan Park.

admission at a preliminary hearing that he was uncertain of his identification of Mr. Collins in a police lineup when the defendant was beardless.³

In an attempt to bolster the identifications, the prosecutor called a local college math instructor to the stand.⁴ This witness testified about the "product rule" in probability theory. According to this rule, the probability that a series of independent events will occur is given by the product of the probabilities for each of the individual events. Thus, the probability that a random draw from a deck of cards will yield the six of spades (i.e., a six and a spade) is given by multiplying the probability of obtaining a six by the probability of obtaining a spade (i.e., $1/13 \times 1/4 = 1/52$).

In argument to the jury, the prosecutor assigned a series of individual probabilities to the characteristics of the perpetrators in a chart similar to Table 1 below.⁵

Insert Table 1 About Here

The prosecutor applied the product rule to the individual probabilities and claimed that the probability that a couple would have all six characteristics is 1 in 12,000,000. He said that there was therefore only 1 chance in 12,000,000 that the defendants did not commit this crime. He further suggested that the probabilities assigned to the characteristics were "conservative estimates" and that the true probability that the defendants were innocent is "something like one in a billion."⁶

³ People v. Collins, 68 Cal 2d 319, 66 Cal Rptr. 497, 438 P.2d 33, 36 A.L.R.3rd 1176 (1968) at 34.

⁴ Collins at 36.

⁵ Table 1 appears Collins at 37, fn 10.

⁶ Collins at 37.

The jury convicted and the defense appealed (in part) on grounds that introduction of the probability evidence was prejudicial error. The California Supreme Court agreed, and reversed Collins on four grounds.

First, the individual probabilities listed in the table lacked adequate evidentiary foundation. Prosecutors may not invent probabilities for various components of evidence, even if they concede that these probabilities are merely estimates.

Second, there was insufficient proof of statistical independence among the events. Indeed, as the Court pointed out, beards and mustaches are not independent events since "most Negro men also have mustaches ... in a hirsute continuum."⁷

Third, the computation implicitly assumed that the six reported characteristics were true and accurately reported. It made no allowance for the possibility that the perpetrators were disguised (e.g., false beard, dyed hair, etc.), or that one or more of the characteristics were incorrectly reported.

Fourth, the prosecutor erred when he equated the 1 in 12,000,000 probability estimate with "mathematical proof of guilt." In other words, even if the first three problems did not exist, it was fallacious to equate the probability of observing these characteristics in a random couple with the likelihood that the Collinses were innocent of this crime.

So what does this famous 1960s case have to do with the even more famous 1990s case of People of California v. O. J. Simpson? I would argue that the same problems that arose in Collins arose in Simpson, and will continue to arise in other cases that include testimony about extremely small DNA frequency statistics.

People of California v. Simpson (1995)

⁷ Collins at 39, fn 15.

In Simpson, DNA blood evidence that matched the defendant and each of the two victims was offered into evidence. In most instances, numerical estimates were provided for the frequency with which the matching DNA characteristics would be found in various populations.

Like Collins, Simpson included presentation of a variety of numerical frequencies. But unlike Collins, jurors in Simpson were inundated with hundreds--if not thousands--of frequencies to describe the evidence. Consider, for example, the data listed in Table 2. Table 2 shows the frequencies that were offered by one witness during one afternoon of cross-examination. On this afternoon, at least 76 different estimates were offered for various blood drops. These estimates ranged from 1 in 1 (i.e., a characteristic that everyone shares, such as "red blood") to 1 in 1 trillion. Because many of the frequencies were repeated multiple times (Table 2 records the first mention only), the jurors heard hundreds of frequencies on this particular afternoon.

Insert Table 2 About Here

Although each frequency listed in Table 2 corresponds to a specific item of evidence introduced by the prosecution, no one know how jurors process and weigh such a barrage of numerical data. There is some evidence from psychological studies that mock jurors attach relatively less weight to numerical evidence than it deserves,⁸ but no studies have examined how jurors treat large sets of frequencies such as those presented in Simpson.

⁸ For reviews see David H. Kaye, & Jonathan J. Koehler, Can Jurors Understand Probabilistic Evidence? 154 J. ROYAL STAT. SOC'Y A 75 (1991); Brian C. Smith, Steven D. Penrod, Amy L. Otto, & Roger C. Park, Jurors' Use of Probabilistic Evidence, 20 LAW & HUM. BEHAV. 49 (1996); William C. Thompson, Are juries Competent to Evaluate Statistical Evidence in Criminal Trials? 52 LAW & CONTEMP. PROBS. 9 (Autumn 1989).

Returning to the similarities between Collins and Simpson, all four of the concerns that were raised by the California Supreme Court in Collins were raised by the Simpson defense team. Specifically, the defense questioned (a) the reliability of the databases that were used to produce the DNA frequencies; (b) the validity of the product rule as a way to identify the probability that a series of DNA characteristics will co-occur; (c) the accuracy of the tests that produced the DNA profiles; and (d) the meaning of DNA frequency statistics.

The first two issues played a central role in several early DNA cases. For example, questions about the reliability of the FBI's DNA database and the validity of the FBI's frequency computation methods led to the exclusion of DNA evidence in state appellate and supreme court rulings.⁹ More recently, however, these concerns have abated.¹⁰ Indeed, most scientists believe that the conservative method for estimating DNA frequencies given by the 1992 National Research Council's report on DNA effectively silenced these criticisms.¹¹

The latter two issues--the accuracy of the reported DNA profiles, and the meaning of the frequency statistics--have now taken center stage in the great DNA debate. These issues

⁹ See e.g., *State v. Anderson* 115 N.M. 433, 445; 853 P.2d 135, 147 (Ct. App. 1993) (FBI's DNA evidence inadmissible because "lack of current scientific acceptance of the FBI database"); *Commonwealth v. Lanigan* 413 Mass. 154, 162-3, 596 N.E. 2d 311, 316 (1992) (FBI's computation method is not appropriate because it fails to account for the possibility of population substructure); *State v. Vandebogart*, 136 N.H. 365, 616 A.2d 483, 494 (S.Ct. 1992) (FBI's use of product rule "has not found general acceptance in the field of population genetics").

¹⁰ Eric Lander & Bruce Budowle, DNA fingerprinting dispute laid to rest, 371 NATURE 735 (1994).

¹¹ B. Devlin, Neil Risch, Kathryn Roeder, Comments on the Statistical Aspects of the NRC's Report on DNA Typing, 39 J. FOREN. SCI. 28 (1994); Richard Lempert, DNA, Science and the Law: Two Cheers for the Ceiling Principle 34 JURIMET. 41 (1993); but see Joel E. Cohen, The Ceiling Principle is Not Always Conservative in Assigning Genotype Frequencies for Forensic DNA Testing, 51 AM. J. HUM. GENET. 1165 (1992). A second NRC report on DNA evidence is expected in 1996.

are addressed below.

What Does 1 in 57,000,000,000 Mean Anyway?

The very small DNA frequency statistics that were presented in Simpson held a special fascination with the American public. Perhaps the most widely reported DNA statistic was a 1 in 57 billion frequency that was associated with a blood stain recovered from the rear gate of Mrs. Simpson's condominium that matched Mr. Simpson's blood characteristics.¹² Television and newspaper reporters and commentators relied on these statistics to suggest that the case against O. J. Simpson was a powerful one. Even some cartoonists had fun using DNA-like frequencies to mock the idea that Mr. Simpson was not involved in this crime.

Insert Cartoon Here

But what does a frequency like 1 in 12,000,000 or 1 in 57,000,000,000 really mean? These are questions that can and should be raised in evidence classes to prepare future litigators for the world of statistical and scientific evidence that awaits them.

The simple answer is that a frequency of 1 in many millions, billions or trillions tells us that--assuming the first three Collins problems do not exist--there is only 1 chance in many million, billion or trillion that a randomly selected person would share the observed characteristics. These tiny frequencies do not themselves tell us (a) the probability that a matching suspect committed the crimes, (b) the probability that someone other than the matching suspect committed the crime, or even (c) the probability that the someone other than the matching suspect is the source of the observed characteristics.

¹² People v. Simpson (1995), Transcript, vol. 171 (June 20, 1995) (testimony by Gary Sims).

To illustrate, consider the blood evidence that was introduced in a pre-trial hearing in Simpson by Los Angeles Police Department blood expert Gregory Matheson.¹³ Mr. Matheson testified that approximately 1 in every 200 people (.43%) have the blood characteristics that were found in a trail of blood drops that led away from the scene of the murders.¹⁴ Mr. Matheson also testified that the blood drops matched Mr. Simpson, but did not match either of the victims.

Two features of this evidence must be understood.¹⁵ First, it is valuable because it excludes a large proportion of people as possible contributors of the blood, but it fails to exclude Mr. Simpson. On the other hand, the evidence is limited because it fails to exclude many people who are not the source of the blood. Indeed, the non-excluded group to which Mr. Simpson belongs might well include thousands of people in Los Angeles, a few of whom might reasonably be considered potential suspects in the case.¹⁶

For this reason, one cannot hear the 1 in 200 frequency and conclude either that (a) there is therefore 1 chance in 200 (i.e., 0.5% chance) that the blood drops are not Simpson's, or (b) there are therefore 199 chances in 200 (i.e., 99.5% chance) that the blood drops are Simpson's. How can we conclude that there is a 99.5% chance that Simpson is the source of

¹³ This discussion is based on a portion of Jonathan J. Koehler, *On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios, and Error Rates*, U. COLOR. L. REV. (in press).

¹⁴ B. Drummond Ayers Jr., *The Simpson Case: The Overview; Simpson Ordered to Stand Trial in Slaying of Ex-wife and Friend*, NY TIMES, A1, (July 9, 1994).

¹⁵ The evidence presented by Mr. Matheson was based on non-DNA blood tests. However, the statistical principles are identical to those that accompany DNA blood matches.

¹⁶ This point was made by the Simpson defense team. On cross-examination, Mr. Matheson agreed that approximately 40,000 people in Los Angeles County share these blood characteristics. See Ayers.

the matching blood when we know that there are hundreds or thousands of others who share this blood profile? Surely, it cannot be that each member of the non-excluded group has a 99.5% chance of being the source of the matching blood. Though surprising, it is true: forensic science analyses alone cannot identify the probability that O. J. Simpson--or any other criminal defendant--is or is not the source of recovered genetic evidence.

Here, the skeptical reader may say, "Yes, but how many of those others who might share Mr. Simpson's blood profile had cuts on their hands, were married to Nicole, or had a history of violence against Nicole?" The point is well-taken, and reinforces my own point that one cannot identify the probability that someone is or is not the source of genetic evidence through genetic testing alone. Nongenetic considerations, such as those listed above, must be factored into any equation that purports to identify the chance that someone is the source of genetic samples.

But have I cheated here? Have I made it too easy for Mr. Simpson? Weren't the frequencies offered in the Simpson case on the order of 1 in many millions and billions? If this doesn't prove conclusively that he committed the crime, doesn't this at least prove that it's his blood? After all, in the face of a frequency estimate of 1 in 57 billion, and an estimated world population of 5 1/2 billion people, aren't the chances of there being another person anywhere who matches this genetic profile pretty much nil? I have two responses. One general, and one specific.

In general, we should be skeptical of odds that live in the stratosphere. What are the odds that Mrs. Evelyn Adams would win the New Jersey state lottery two years in a row? According to statisticians at Rutgers University, the odds of this occurring were 1 in 17.3 trillion.¹⁷ Yet, it happened to her in 1985 and 1986. Probably never to happen to anyone

¹⁷ Robert D. McFadden, Odds-Defying Jersey Woman Hits Lottery Jackpot 2d Time, NY TIMES, A1 (Feb. 14, 1986).

again.

But then in 1987 employees of the Shuttle Meadow Country Club won their 2nd Connecticut State lottery. Another miracle. And the following year a man won the Pennsylvania state lottery for the second time.¹⁸

And what would you say the odds are that two identical twins will bowl perfect games (score=300) on the same night. A mathematician at the University of Memphis calculated these odds at 1 in 385 billion. "This would happen once in 10 million years," he said. Yet, it happened in Plainview, NY in September, 1995.¹⁹

What is going on here? Do events that defy astronomical odds happen? Perhaps, but in the case of these tiny lottery and bowling frequencies, the assumptions that lurk behind the computations are misleading and help illustrate a phenomenon called "the selection fallacy." It is not appropriate to compute the odds associated with a particular set of events when the phenomenon of interest would be satisfied by any of a large number of events. We don't conclude that a bizarre event has occurred every time a particular person wins the lottery because the odds that someone will win are quite high. Thus, for the dual lottery win problem, statisticians should not compute the frequency that each of two tickets bought by Mrs. Adams for two different lotteries will win the grand prize; instead, they should compute the frequency that someone like Mrs. Adams, who buys multiple tickets for weekly lotteries on a regular basis, will win at least two lottery grand prizes. This computation gives much less extreme values. According to one published set of calculations, we should expect a dual lottery winner approximately every 7 years.²⁰ Indeed, it would be surprising if such dual winners did not

¹⁸ Cite.

¹⁹ Marshall Lubin and Andrew Smith, Twin Brothers Beat Long Odds: They Both Bowl Perfect Games, *NEWSDAY*, A7 (September 2, 1995).

²⁰ Stephen M Samuels, George P. McCade Jr., More Lottery Repeaters are on the Way *NY*

appear!

The point in this admittedly general response is that when we come across extreme frequencies, we must think through the assumptions that were used to create those numbers. We must then consider whether those assumptions adequately map onto the questions of interest and what interpretations are justifiable.

And now my specific response to the question of whether DNA statistics on the order of 1 in millions, billions, or trillions prove conclusively that the matchee is, in fact, the source of the recovered genetic material. On May 15, 1995, an instructive cross-examination took place in the Simpson trial between Simpson defense attorney Peter Neufeld and prosecution witness Dr. Robin Cotton. Dr. Cotton is the Director of Cellmark laboratories, which was one of the laboratories that reported DNA matches between recovered blood stains and Mr. Simpson.

During this cross-examination, Mr. Neufeld elicited testimony that Cellmark reported a frequency statistic of 1 in 1.8 billion on a blood sample in a proficiency test situation several years ago.²¹ However, Cellmark mistyped this sample.²² What, then, are we to make of the 1 in 1.8 billion frequency statistic?

Q: Well, Dr. Cotton, would you agree that if this particular proficiency test was a real case that the statistic of 1 in 1.8 billion would be irrelevant if, in fact, you knew it was a false match

TIMES, A22 (February 27, 1986).

²¹ California Association of Crime Laboratory Directors, DNA Committee Report #6, (October 1, 1988). DNA analysts participate in proficiency tests, in part, to determine the rate at which they commit false positive and false negative errors. In the typical test, analysts are provided with blood stains from known sources, some of which are from common sources, and some of which are from different sources. The analysts are asked to use DNA profiling methods to draw conclusions about which samples match and which do not.

²² Id at 5.

between 59 and 57?

*A: Yes, it would be irrelevant.*²³

This should be clear. Of course the frequency would be irrelevant if the laboratory mistyped the sample, and no ethical prosecutor would suggest otherwise. But this point becomes central when we consider that, in actual case work (as opposed to proficiency testing situations), it is not known whether a mistake was made or not.

One surprising finding that emerges from studies of DNA proficiency tests is that laboratory false positive error rates appear to be on the order of 1 in 100 or so. That is, when a laboratory reports a match, there is about one chance in one hundred that the match report is in error.²⁴ In documents obtained by the Simpson defense under discovery, Cellmark reported its own false positive error rate to be on the order of 1 in 200.²⁵

A complete discussion of errors is beyond the scope of this article. But suffice it to say that errors occur for a variety of reasons. Some errors are technical, others are human. The human errors--which appear to be more common--include such problems as contamination, mislabeling, misrecording, case mixups, interpretive errors and misrepresentations.

With this in mind, the following question may be raised: Do the tiny DNA frequencies--frequencies on the order of 1 in millions, billions and trillions--have any probative value

²³ *People v. Simpson*, Transcript, vol. 154 (May 25, 1995).

²⁴ Another way to think about false positive error rates is in terms of the proportion of nonmatches that are erroneously reported as matches. False positive error rates computed in this manner from proficiency test data are on the order of one in several hundred. See Jonathan J. Koehler, Audrey Chia, J. Sam Lindsey, *The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?* 35 JURIM. 201, 206-211 (1995).

²⁵ Linda Danielsen, Quality Assurance Manager at Cellmark Diagnostics, INTERNAL MEMORANDUM on LABORATORY RFLP AND DQ ALPHA PROFICIENCY TEST ERROR RATES (June 23, 1993) (RFLP proficiency test error rate from 1988-1992 = 2/365 = .0055).

beyond that which is given by the laboratory error rate when the error rate is many orders of magnitude greater than DNA frequency? My answer is that they do not, and I have suggested elsewhere that admission of both sets of numbers may be unduly prejudicial to a defendant.²⁶

As a matter of policy, I recommend admission of a single, easily-understood index that captures the probative value of the evidence. In most cases, this index would be controlled almost entirely by the error rate rather than the more impressive-looking DNA frequency statistic. *If an error occurs, say, 1 time in 100, it makes no difference whether the DNA frequency statistic is 1 in 1,000,000 or 1 in 57 billion. The chance that a reported DNA match is erroneous is about 1 in 100 regardless of the how small the DNA frequency statistic becomes.* Creating a statistic that is vanishingly small has nothing to do with decreasing the chances of a contamination or mislabelling a test tube. In the end, the probative value of a reported DNA match is more closely linked to scientists' ability to avoid committing errors than it is to the tiny theoretical probabilities that forensic scientists testify to in open court.²⁷

Conclusion

In both Collins and Simpson, very tiny statistical frequencies were used by prosecutors to link defendants with a crime. And in both cases, the matching characteristics that lurked behind these frequencies (eyewitness characteristics in Collins, DNA characteristics in Simpson) did have probative value.

But in both cases, the probative value of this evidence, as suggested by the frequency

²⁶ See Koehler, Chia, and Lindsey (1995).

²⁷ Simpson defense attorney Peter Neufeld attempted to develop this theme in his cross-examination of Dr. Cotton, but was thwarted by Judge Ito's belief that the attempt to combine large and small probabilities created "a Collins problem" (People v. Simpson, (1995), Transcript, vol. 146, May 15, 1995).

statistics, was severely constrained by the possibility of error. This is not to say that error was likely in either case.²⁸ But because the possibility of error was much more likely than one in 12,000,000 in Collins, and one in 57 billion in Simpson, a serious FRE 403 concern arose. That is, jurors might not only have failed to understand how to combine the possibility of error with the theoretical frequency statistics to identify the probative value of the DNA evidence, but they may have believed mistakenly that the possibility of error was either irrelevant or was already assimilated into the frequency statistics.²⁹

If jurors are confused by DNA statistics, then there is a real risk that they may mistakenly believe that frequencies like 1 in a million or 1 in a billion estimate the chance that defendants are innocent of the crimes they are accused of committing.³⁰ This translation error, which has been called the "inverse fallacy," is well-documented in the scientific literature on probabilistic reasoning.³¹

²⁸ In Simpson, error probably cannot provide a sufficient explanation for the reported DNA matches because multiple samples were tested at different laboratories. Although error rates in such situations are unknown, it is extremely unlikely that all laboratories erred when testing all samples. On the other hand, multiple testing will not ordinarily guard against the problems caused by early errors such as those that can occur in the sample collection and preparation stages. For example, an early mislabeling (or contamination) of a sample could lead each of several laboratories to report a match between recovered genetic material and a suspect even where no match existed.

²⁹ See Jonathan J. Koehler, Audrey Chia and J. Sam Lindsey at 211-216 (1995) (reporting that mock jurors appeared to be confused about how to interpret DNA error rates and frequency statistics).

³⁰ See also William C. Thompson & Edward L. Schumann, Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy, 11 LAW & HUM. BEHAV. 167 (1987) (reporting the frequency of "prosecutor's fallacy" reasoning among mock jurors).

³¹ Ward Casscells, Arno Schoenberger & Thomas B. Graboys, Interpretation by Physicians of Clinical Laboratory Results, 299 NEW ENG. J. MED. 999 (1978); Gerd Gigerenzer & Ulrich Hoffrage, How to Improve Bayesian Reasoning Without Instruction: Frequency Formats, 102 PSYCH. REV. 684 (1995); Robert M. Hamm,

In light of the problems statistical evidence and arguments can create in the courtroom, mechanisms for handling these problems should be discussed. Some have argued that statistics should be excluded from court because they are easily manipulated and difficult to understand.³² I disagree but believe that the odds are greater than one in a million that this issue, and others raised here, will spark lively debate in law school classrooms for years to come.

Explanation for Common Responses to the Blue/Green Cab Probabilistic Inference Word Problem, 72 PSYCH. REP. 219 (1993); Willem A. Wagenaar, The Proper Seat: A Bayesian Discussion of the Position of Expert Witnesses, 12 LAW & HUM. BEHAV. 499 (1988); Christopher R. Wolfe, Information Seeking on Bayesian Conditional Probability Problems: A Fuzzy-Trace Account, 8 J. BEHAV. DEC. MAK. 85 (1995).

³² Lawrence H. Tribe, Trial by Mathematics: Precision and Ritual in the Legal Process, 84 HARV. L. REV. 1329 (1971).