

Forensic Science Testing: The Forensic Filler-Control Method for Controlling Contextual Bias,
Estimating Error Rates, and Calibrating Analysts' Reports

Gary L. Wells¹

Miko M. Wilford¹

Laura Smalarz¹

¹Psychology Department

Iowa State University

West 112 Lago Hall

Ames, IA 50011

E-mail addresses:

Gary Wells glwells@iastate.edu

Miko Wilford mwilford@iastate.edu

Laura Smalarz lsmalarz@iastate.edu

Corresponding author: Gary L. Wells 515-201-0431

Accepted for publication but do not quote until final edited journal version

Abstract

Forensic science testing has been severely criticized for lacking a scientific basis, not having known error rates, overstating the certainty of conclusions, and being infested with contextual information (National Academy of Sciences, 2009). Kassin et al's (2013) review of basic social psychological science on expectancy effects, confirmation bias, and related phenomena as well as recent studies showing the influence of contextual information on forensic scientists' judgments makes a strong case for controlling contextual bias. But even if contextual bias were fully eliminated, the big questions of error rates and overstatement of conclusions will still plague forensic science tests. We develop a filler-control method that can quickly sort between reliable and unreliable forensic science as used in actual cases, quickly uncover frauds, and help calibrate forensic scientists' claims of certainty regarding their tests.

Accepted for publication but do not quote until final edited journal version

Forensic Science Testing: The Forensic Filler-Control Method for Controlling Contextual Bias,
Estimating Error Rates, and Calibrating Analysts' Reports

A large share of forensic science techniques involving the analysis of physical evidence have never been validated scientifically. The National Academy of Sciences (NAS, 2009) concluded that, with the exception of nuclear DNA analysis, no forensic method has been rigorously shown to consistently and with a high degree of certainty demonstrate a connection between evidence and a specific individual or source and have not developed evidence-based estimates of error rates. The NAS report also noted that forensic analysts¹ are subject to “contextual bias,” which occurs when the analyst is influenced by knowledge about the suspect’s background or other case information.

Kassin et al. (2013) reviewed basic social and cognitive psychology supporting the concern regarding contextual bias and makes a compelling case for the value of psychological science to demonstrate and understand contextual bias in forensic testing. Unfortunately, even if contextual bias were neutralized, it leaves untouched significant problems identified by the NAS report: Is the technique reliable? What is the error rate? Is the analyst competent to conduct the procedure? Are analysts overstating their conclusions? Forensic science testing and testimony has proliferated without answers to these questions. In effect, a large share of “forensic science” techniques, almost all of which were developed by people in law enforcement rather than scientists, end up coming down to what we call the “inter-ocular” test in which the analyst visually examines the evidence and determines whether it is a match, not a match, or is inconclusive.

¹ We use the term analyst to refer to the forensic scientists, forensic examiners, forensic technicians, and forensic analysts who perform these forensic tests and write the reports.

Here, we develop the framework and rationale for a method of forensic testing that we call the “filler-control method².” We argue that this method, which could theoretically be imposed upon forensic testing systems, could simultaneously resolve the bulk of issues addressed in the NAS report: (1) estimate error rates for both the technique and the individual analyst, (2) help calibrate forensic analysts, (3) quickly jettison fraudulent analysts and junk science, and (4) protect against contextual influences. Importantly, the filler-control method can accomplish all four of these objectives using data generated by actual cases as they unfold. Although others have alluded to the idea that a filler-control method might guard against contextual influences in forensic testing (Kassin et al., 2013; Saks et al., 2003), we argue that proper use of the filler-control method as described here can alleviate all four of the above-mentioned problems.

The filler-control method is best suited to forensic science techniques that involve match judgments. The match domain is a large one that includes fingerprints, tool marks, ballistics, bite marks, hair analysis, fibers, handwriting, shoe prints, and any other test for which an analyst is asked to determine whether a trace sample from a crime scene matches something associated with a suspect.

The Three Big Questions

Before describing the filler-control method, it is useful to focus on the three big questions that confront the courts when they encounter forensic science match evidence: (1) Is the technique itself reliable? (2) Was this analyst competent to perform the technique? And, (3) was the technique performed properly in this particular case? The filler-control method provides

² These ideas are based largely on the first author’s American Psychology-Law Society Presidential address given at the American Psychological Association meeting in 2006, which was drew heavily from the logic of fillers for eyewitness identification from lineups (Wells, et al, 1993).

answers to these three big questions with surprising statistical efficiency using the actual cases in which the technique is being used.

Unlike proficiency testing³, which uses samples of unknown representativeness, reliability estimates from the filler-control method are based on the success of the technique as applied in actual cases. Moreover, past proficiency tests cannot determine whether the technique was conducted properly (or conducted at all) in any particular case. Consider, for example, West Virginia State Police chemist Fred Zain. Zain fabricated forensic test results in hundreds of cases from 1979 until 1989 (Chan, 1994; Tilstone et al. 2007). Claiming a degree in biology (his degree was actually in English), Zain fraudulently made up serology results that always coincided with the police and prosecution theory. In later years, he also fabricated DNA results, which is how he was ultimately caught (via post-conviction DNA re-testing). The filler-control method would have “outed” the fraudulent Zain much earlier.

The Filler-Control Method

A standard practice in forensic testing is to provide the analyst with two samples—one from a suspect and one from the crime scene. Analysts are then expected to provide an answer to the question, “Does this trace evidence match this sample from the suspect?”⁴ For example, does this photo of bite marks on the victim match this denture impression from the suspect? Or, does this photo of a shoe print from the crime scene match this shoe owned by the suspect? Or, do the striations from this crime scene bullet match those made by this suspect’s gun?

³ Proficiency tests involve giving a lab certain samples to process. Often, these labs know at the time that it is a proficiency test rather than a real case.

⁴ We fully recognize that suggested guidelines for reporting exist for some techniques that involve multiple categories. For example, there is a suggested recommendation for footwear and tire tread analysis that involves six categories ranging from “identification” (definite conclusion of identity) to “elimination” (definite exclusion) and a seventh category of “unsuitable” (Scientific Working Group on Shoeprint and Tire Tread Evidence, 2006). Although our treatment here uses the match, mismatch, and indeterminate categories, a larger set of categories does not undermine the fundamental principles behind the filler-control method.

The filler-control method, in contrast, presents the analyst with a qualitatively different task. Instead of two samples, there are a minimum of three—the crime scene sample, the suspect sample, and innocent filler sample(s). The suspect sample and the innocent filler samples are intermixed, and the analyst does not know which sample is from the suspect and which are from fillers. Analysts are then confronted with a very different question, “Does the crime scene sample match any of these other samples, and if so, which one?” The analyst must perform the test and write the report without any prior knowledge of which sample was from the suspect and which were from fillers. Consider, for example, a bite-mark test. Instead of providing the analyst with a matching task in which they are limited to photos of victim bite marks and one denture impression from the suspect, the analyst would receive an additional five denture impressions thereby expanding the task (the suspect’s plus five filler sets of dentures matched for age, gender, and general oral health)⁵. Or, if matching a bullet from the crime to a particular gun, the analyst would be given the suspect’s gun along with five other (filler) guns of the same caliber and make. A necessary quality of the fillers is that they are known a-priori to not be the actual source.

Critically, unlike the traditional method, the filler-control method is a test that the analyst can conspicuously fail. For example, the analyst could incorrectly conclude that one of the filler .357 Magnum pistols was the source of the bullet from the crime scene (e.g., the police chief’s gun). Using the filler-control method, unreliable forensic techniques and analysts would be quickly discredited. In contrast, reliable analysts using reliable techniques would welcome this challenge and easily and repeatedly avoid fillers in their forensic examinations.

Note that the filler-control method manages to exert control over contextual bias as well. Even if forensic analysts have been exposed to contextual information (e.g., knows the police

⁵ For purposes of this article we use five fillers.

obtained a confession, knows that an eyewitness identified the suspect, etc.), they still do not know which samples came from the suspect and which from fillers.

Discovery of Error Rates, Unreliable Techniques, Incompetent Analysts, and Frauds

No fraud, incompetent analyst or unreliable technique can withstand the odds of exposure inherent in the filler-control method. Although a lucky guess has a $1/6^{\text{th}}$ chance of avoiding a filler on a single occasion, there is only a $1/36^{\text{th}}$ chance of avoiding a filler 2 out of 2 times, a $1/216^{\text{th}}$ chance of avoiding a filler 3 out of 3 times, and only a $1/1,296^{\text{th}}$ chance of avoiding a filler 4 out of 4 times. This means that only 1 of nearly 1300 forensic analysts could avoid making a public error for affirmative conclusions in their first four cases if their test was totally invalid.

Figure 1 displays the theoretical probability that a forensic test would produce at least one filler error as functions of the validity of the forensic test and the number of cases for which the analyst made an affirmative judgment⁶. The highest filler-error line comes from a forensic test with zero validity (i.e., accuracy is no better than chance). Notice how quickly the zero-validity test approaches near certainty of yielding at least one error; there is a 99.5% chance that an analyst would have made at least one filler error by the time a forensic analyst had tested three cases. If the validity of the test were .20, a single test has a .667 probability of a filler error but this rises to a 98.8% chance of yielding at least one filler error by the time the test is used in four cases. In contrast, a 99% valid test produces a mere 1% chance of yielding a filler error in 10 cases. The best of the forensic tests rise to the top and the unreliable ones are quickly unmasked with the filler-control method.

⁶ These analyses are based only on affirmative match judgments (as opposed to indeterminate judgments).

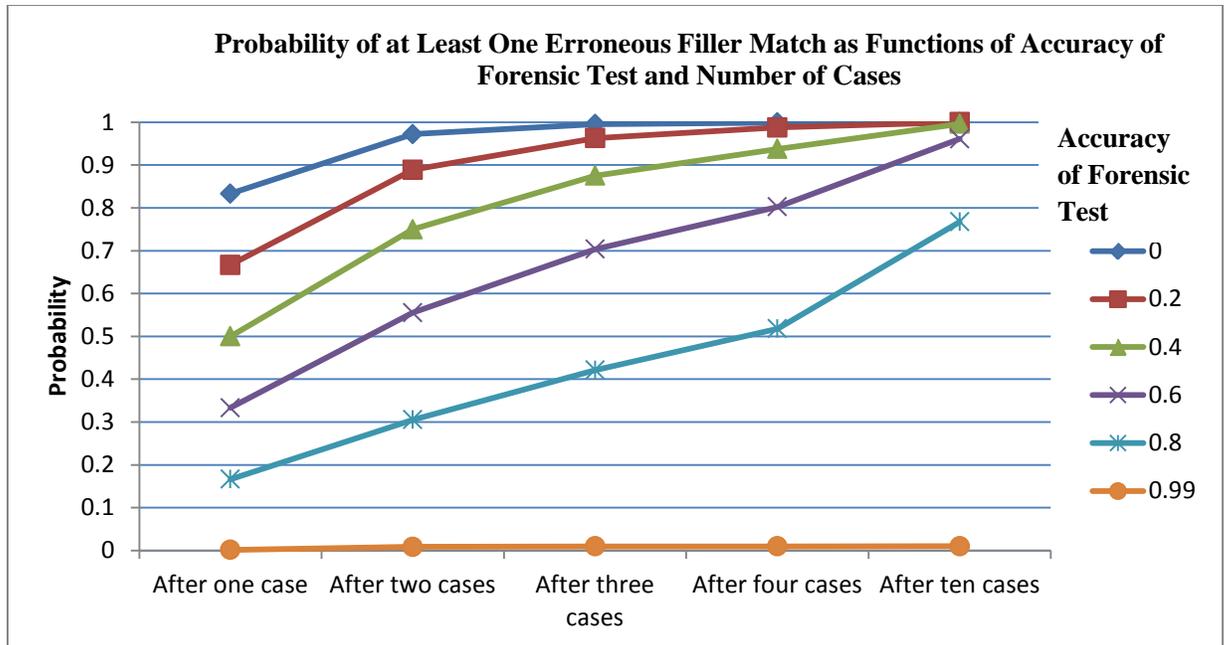


Figure 1 illustrates that the five-filler control test, although not perfectly capable of revealing known errors when used once, requires only a few uses to unmask low-validity. A Fred Zain, for example, would have almost certainly have shown failures after only three tests, thereby stopping him from erroneously claiming to have found affirmative matches in the hundreds more cases for which he remained a credible expert.

One of the great advantages of the filler-control method is that scores can be partitioned so that error-rates can be calculated separately for techniques, labs, and analysts. As the NAS report noted, we know little about error rates in forensic testing. With the filler-control method, estimation of the error rate is approximated by the filler rate, but is not exactly the filler rate, due to chance. Corrected for chance, the best estimate of the error rate is:

$$\text{Equation 1. Estimated error rate} = 1 - \left[\frac{N_{\text{fillers}}}{N_{\text{fillers}} + 1} - \text{Filler rate} \right] \frac{N_{\text{fillers}} + 1}{N_{\text{fillers}}}$$

Hence, for example, assuming a five-filler test ($N_{\text{fillers}} = 5$), if the analyst selects fillers 20% of the time (filler rate = .20) and selects the suspect sample 80% of the time, then the error rate is 24%, not 20%, not 20%. For a 40% filler rate, the error rate estimate is 48%, and so on.

Influence of the Filler-Control Method on Beliefs and Behaviors of Forensic Analysts

Forensic analysts who are reliably using highly valid techniques should be elated at the idea of using the filler-control method. These analysts and their techniques will be validated by the filler control method. Forensic analysts who believe *erroneously* that their techniques are perfectly valid, in contrast, represent a particularly interesting group. The illusion of validity (e.g., Einhorn and Hogarth, 1978) and its sister phenomenon of overconfidence are pervasive in human judgment. Although the illusion of validity can have many causes, including wishful thinking and confirmation bias, the illusion can flourish only in the absence of error feedback. The filler-control method creates a new forensic testing environment in which feedback about errors is guaranteed.

We believe that the majority of forensic analysts who use the filler-control method will discover that their technique has some validity, but also makes errors, which could surprise some of the analysts themselves. For these techniques, the filler-control method can help analysts to calibrate their reports and conclusions. For example, if an analyst using a particular technique has a record of six erroneous filler-match conclusions and 30 suspect-match conclusions, then the estimated error rate is 20% (see Equation 1) and the analyst therefore should not express confidence higher than 80% in any given match conclusion.

Final Remarks

Kassin et al. (2013) successfully articulate the problematic influence of contextual information in forensic testing. Unfortunately, eliminating contextual information fails to answer

whether the technique is reliable, the analyst competent, and whether the test was actually performed competently in this particular instance. The filler-control method manages to address all of these problems while additionally controlling for contextual bias.

Many forensic analysts who are involved in match tests will resist the idea of the filler-control method on several grounds, primarily arguing that it is quixotic. Obviously there would be practical costs (money, time) to performing the additional tests. Moreover, full realization of the filler-control method's best features (such as developing a searchable public data base of error rates across analysts and techniques) requires an oversight body, reporting requirements for every test conducted, and refined protocols for selecting appropriate fillers.

In more than one talk given by the lead author of this article, an analyst argued that because some crime scene samples are seriously degraded (e.g., smudged fingerprint, nearly-healed bite mark), the analyst would be unable to clearly match the crime scene sample to the suspect sample. Hence, the "score" for the analyst and technique would be harmed for reasons other than the analyst's competence or the technique's reliability. But this is a sophistic argument because the "indeterminate" category was created precisely for such circumstances. Only affirmative match conclusions can count for or against the technique and analyst. An indeterminate result is a non-event for purposes of scoring reliability.

The practical concerns are legitimate and deserve further consideration. Nevertheless, the bold idea of a filler-control method requirement in forensic match tests deserves a place at the table of how to control contextual bias, reduce fraud, generate meaningful estimates of error rates in actual cases, and give analysts the type of feedback that could allow them to accurately calibrate their judgments.

References

- Chan, Sau (1994). Scores of convictions reviewed as chemist faces perjury accusations. *Associated Press*, August 21, 1994.
- Einhorn, H. J., Hillel, J., & Hogarth, R.M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85, 395-416.
- Kassin, S., Dror, I., & Kukucka (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*. XXXXX
- National Academy of Sciences (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press: Washington, DC.
- Tilstone, W. J., Savage, K. A., Clark, L. A. (2006). *Forensic science: An encyclopedia of history, methods, and techniques*. ABC-CLIO, Inc., Santa Barbara, CA.
- Saks, M. J., Risinger, D. M., Rosenthal, R., & Thompson, W. C. (2003). Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States. *Science & Justice*, 43, 77-90.
- Scientific Working Group on Shoeprint and Tire Tread Evidence. (2006). *Guide for minimum qualifications and training for a forensic footwear and/or tire tread examiner*. Available at <http://www.theiai.org/guidelines/swgtread/>
- Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). On the selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, 78, 835-844.