Article

Review of Several False Positive Error Rate Estimates for Latent Fingerprint Examination Proposed Based on the 2014 Miami-Dade Police Department Study

Madeline A. Ausdemore
Jessie H. Hendricks
Cedric Neumann

Department of Mathematics and Statistics
South Dakota State University
Brookings, SD

Abstract: In 2014, the Miami-Dade Police Department (MDPD) Forensic Services Bureau conducted research to study the false positive error rate (FPR) associated with latent fingerprint examination. They report that approximately 3.0% of latent fingerprint examinations result in a false positive conclusion. In their 2016 report, the President's Council of Advisors on Science and Technology (PCAST) advise that this estimate of the FPR be used to inform jurors that errors occur at a detectible rate in fingerprint estimation and declare that false positive conclusions may occur as often as 1 in 18 cases. In this paper, we review the MDPD study and design a simulation study to model the behavior of the participants in the MDPD study. We use our model to simulate the number of erroneous identifications that occur under any assumed FPR and compare the results to the actual number of erroneous identifications observed in the MDPD study. We conduct experiments associated with the error rates proposed by the MDPD and the Organization of Scientific Area Committees Friction Ridge subcommittee. We note that the results of these experiments indicate that none of the proposed FPRs are reasonable estimates of the true FPR associated with the MDPD study. We propose two solutions based on a Bayesian analysis of the data, each resulting in two separate FPRs. Our solutions are comparable to the estimates offered by the Noblis "black-box study".

Received August 6, 2018; accepted November 13, 2018

Introduction

During the past decade, several studies have been conducted to estimate the false positive error rate (FPR) associated with latent fingerprint examination. The so-called "Black-box study" by Ulery et al. [1] is regularly used to support the claim that the FPR in fingerprint examination is reasonably low (0.1%). Ulery et al.'s estimate of the FPR is supported by the results of the extensive study of the overall fingerprint examination process by Langenburg [2].

In 2014, the Miami-Dade Police Department (MDPD) Forensic Services Bureau conducted research to study the false positive error rate associated with latent fingerprint examination [3]. They report that approximately 3.0% of latent fingerprint examinations result in a false positive conclusion. Their estimate of the FPR becomes as high as 4.2% when inconclusive decisions are excluded from the calculation. In their 2016 report, the President's Council of Advisors on Science and Technology (PCAST) proposes that the MDPD FPR estimate be used to inform jurors that errors occur at a detectable rate in fingerprint examination; more specifically, they declare that false positives may occur as often as 1 in 18 cases [4].

The large discrepancy between the FPR estimates reported by Ulery et al. [1] and Langenburg [2] on the one hand, and the MDPD on the other hand, causes a great deal of controversy. For example, a recent Canadian case study advocates for a re-analysis of the MDPD data and of its interpretation by PCAST, before the PCAST point of view "becomes an urban myth" [5]. The Organization of Scientific Area Committees Friction Ridge subcommittee (OSAC FRS) [6] proposes an alternative estimate of the MDPD FPR. In this paper, we review the MDPD study and the various error rate calculations that have been proposed to interpret its data. To assess the appropriateness of the different proposed estimates, we develop a model that re-creates the MDPD study. This model allows us to estimate the expected number of false positive conclusions that should be obtained with any proposed FPR and compare this number to the actual number of erroneous identifications observed by MDPD.

Review of Pacheco et al. [3]

In 2014, the MDPD reported [3] the results of a study designed to evaluate the accuracy and reliability of fingerprint examination conducted using the standard ACE-V procedure. This study consisted of three phases. Phases 1 and 2 were designed to study the analysis, comparison, and evaluation stages of the ACE-V process; Phase 3 was designed to study the verification stage.

The following paragraphs summarize the design and results obtained by Pacheco et al. [3] The design is based on a set of 80 latent prints that were collected and partitioned into two disjoint subsets consisting of 40 prints each, where each subset included latent prints of varying difficulty. In addition, a full set of control prints (consisting of impressions from all 10 fingers and from both palms) was collected from 10 individuals. Control prints from the true sources were available for 56 out of the 80 test latent prints and were absent for the remaining 24. The repartition of these 56 and 24 latent prints across the two subsets of 40 is unknown.

In Phase 1 of the study [3], the participants were provided with identical packets containing the first group of 40 unknown latent prints and all 10 sets of control prints. For each of the 40 latent prints, participants were instructed to first perform an initial analysis to determine whether the latent print was suitable for comparison purposes (or, of value). If the latent print was deemed to be of value, the participants were instructed to compare the print to three prespecified sets of control prints (from the 10 sets available for the study) and make a conclusion of identification, exclusion, or inconclusive (control prints from the source of the latent print were not necessarily present in these three specified sets of prints). In Phase 2, the 109 participants who returned their Phase 1 packets were divided into two groups, A and B. The second set of 40 latent prints (that had not been used in Phase 1) was further divided into two sets of 20 latent prints each. Individuals placed in group A were provided with identical packets containing the first set of 20 latent prints and the 10 sets of control prints. Similarly, the individuals placed in group B were provided with identical packets containing the second set of 20 latent prints and the full set of control prints. The participants then followed the same process outlined for Phase 1. Finally, in Phase 3, participants were presented with sets of prints that consisted of the “identifications” made in Phase 2 (regardless of whether the “identification” was correct). The participants were then asked to verify these “identifications" by indicating whether they agreed or disagreed with the conclusions made in Phase 2. If the participant believed there was insufficient information to identify or exclude, he or she was instructed to make an inconclusive decision. To test for bias and repeatability, the packets constructed for this phase were not identical. A summary of the data collection methods is given on pages 36 through 39 of the MDPD report. We are not concerned with Phase 3 of the MDPD study and it is not considered further in the current paper.

Phase 1 packets were sent to 140 participants. A total of 109 packets with results from Phase 1 were returned, including a total of 4233 decisions out of a possible 4360 decisions (109 packets x 40 latent prints per packet). This indicates that some participants in the study returned results for some of the latent prints provided in the packets and failed to return results for others. Out of the 4233 decisions that were returned, 1023 latent prints were deemed of no value. The remaining 3210 latent prints were compared to the prespecified sets of control prints, and a conclusion about their source was reached.

Phase 2 packets were distributed to all 109 respondents from Phase 1. Results were returned by 88 participants in Phase 2, including 1730 decisions out of a possible 1760 decisions (88 packets x 20 latent prints per packet). Out of the 1730 decisions that were returned from Phase 2, 1342 latent prints were deemed of value for comparison and the remaining 388 were deemed of no value for comparison.

The design of the MDPD study involves the comparison of each latent print (if it is deemed of value for comparison) to three sets of control prints. This design is closer to casework conditions where latent print examiners have to "search" for corresponding control impressions. This is different from the study by Ulery et al. [1], where each latent print was paired with a single control print. With the Ulery et al. design, only two types of error are possible during the evaluation stage of ACE-V: erroneous identifications and erroneous exclusions (assuming that inconclusive decisions are not considered to be "errors"). With the MDPD design, a third type of error is possible: it is also possible to report the wrong finger of the correct person as the source of a latent print (when the control prints of the true donor were provided as part of the three sets of control prints).

Error Rate Estimates Based on the MDPD Data

In Table 5 of their report, the MDPD claims an FPR estimate of 3.0% [3]. This rate was obtained by relating the 42 erroneous identifications observed during the study to a total of 1398 decisions (including inconclusive conclusions). The MDPD estimate raises to 4.2% if the inconclusive conclusions are not accounted for (42 erroneous identifications out of 1398 – 403 = 995 decisions).

The original data used in this calculation can be seen in Table 1, which corresponds to a modified version of Table 11 in the MDPD report.

Source Present (Y/N)	# of Latent Prints	# of Decisions	Correct ID	Erroneous ID		Inc.	Correct Excl.	Erroneous Excl.
				Correct Person, Wrong Finger	Incorrect Person			
Yes	56	3177	2457	35	4	446	N/A	235
No	24	1359	N/A	N/A	3	403	953	N/A
Totals	80	4536	2457	35	7	849	953	235

Table 1

A modified version of Table 11 in the Miami-Dade report where the Erroneous ID column has been divided into erroneous identifications in which the correct individual was identified, but the incorrect finger was reported and erroneous identifications in which the wrong individual was identified.

MDPD reports that the 42 erroneous identifications consist of 35 cases in which an examiner correctly identifies the individual who produced the print but fails to associate the latent print with the correct finger, and 7 cases in which an examiner incorrectly identifies the individual who produced the latent print (3 of which occur when control impressions from the true source are not present in the predetermined set of control prints).

MDPD considers these 42 erroneous identifications in the numerator of all of their calculations. Although these 42 cases consist of decisions that occur when control impressions from the true source are present in the comparison process and when they are not present in the comparison process, the denominator only considers the 1359 test cases for which these impressions are not present in the comparison process and the sole 39 decisions for which the control impressions were present but errors were made (1359 + 39 = 1398).

Thus, the main point of contention with the MDPD calculation is that, although most of the erroneous identifications are observed when control prints from the true source are present in the test packets, the FPR estimate does not account for the number of decisions that are rendered in this scenario. In other words, the scenario of the numerator does not match the scenario of the denominator of the FPR estimate. This issue is reinforced by the surprisingly large discrepancy between the proportion of erroneous identifications made when the source is present (39 erroneous ID out of 3,177 decisions or 1.2%) and when it is not (4 out of 1359 decisions or 0.3%). This discrepancy is noted by the authors of the study but is not accounted for in their calculations and interpretation of the data [3].

Despite these issues and that the MDPD study was never published in a peer-reviewed journal, the PCAST report relies heavily on these results to discuss the estimated error rate associated with latent fingerprint examination [4]. Furthermore, the PCAST report considers, using the upper bound of a 95% confidence interval on the FPR estimate, that the true FPR could be as high as 1 in 18 (or approximately 5.4%) cases.

The OSAC FRS has raised the issue of the inconsistency between the numerator and the denominator of the FPR calculation in the MDPD study [6]. The OSAC FRS proposes that the calculation for the FPR considers all 42 erroneous identifications out of the total number of decisions made, excluding inconclusive decisions (4536 total decisions – 849 inconclusive decisions = 3,687). This lowers the FPR to approximately 1.1%, with a 95% confidence interval upper bound of approximately 1.5%. The OSAC FRS does not expand on its decision to exclude inconclusive decisions from the calculation. In this paper, we also consider an FPR estimate obtained by accounting for inconclusive decisions. Thus, we consider all 42 erroneous identifications, but this time, out of the total number of decisions made, including inconclusive decisions (42 erroneous identifications out of 4536 total decisions). This calculation further decreases the FPR to 0.9%, with a 95% confidence interval upper bound of approximately 1.2%.

Although disagreements are expressed on how the FPR should be calculated, the issue of the over-representation of the 35 cases in which examiners correctly identified the individual who produced the test latent prints, but failed to associate them to the fingers from which they originated, has not been raised. Although it is clear that some type of error has occurred, it is not possible to determine a posteriori what happened. For example, these 35 cases may result from clerical errors; alternatively, they may be due to some individuals having very similar patterns across their 10 fingers. We do not want to speculate on the reasons that resulted in these 35 cases; we simply consider that errors were made since the wrong conclusions were reported. However, these errors are erroneous identification to the wrong finger of the correct donor, and it is clear that, from a statistical perspective, these 35 cases need to be considered separately from erroneous identifications made to the wrong donor. The following simulations support this classification.

Simulation Study

Model of the MDPD Study

To assess the appropriateness of the error rate calculations proposed by the MDPD research team, the OSAC Friction Ridge subcommittee, and our alternative that includes inconclusive decisions, we have modeled the behavior of the participants to the MDPD study. This model allows us to simulate the decisions and conclusions that would be obtained using different values for the rates of correct, inconclusive, and erroneous conclusions when the source was present in the three sets of controls (R_{SP}), and when the source was not present (R_{SA}) (Table 2). Algorithms 1, 2, and 3 (Appendix) show the process used to simulate the Miami-Dade error rate study.

Proposed Rates	True Source Present (Yes: R_{SP}/ No:R_{SA})	FPR		FNR	TPR	TNR	Inc.
		Correct Person, Wrong Finger	Incorrect Person				
Observed frequencies	Yes	0.011	0.001	0.74	0.774	N/A[4]	0.140
	No	N/A[1]	0.002	N/A[2]	N/A[3]	0.698	0.300
Miami-Dade	Yes	0.030		0.075	0.755	N/A	0.140
	No			N/A	N/A	0.667	0.300
Miami-Dade (common inconclusive)	Yes			0.075	0.708	N/A	0.187
	No			N/A	N/A	0.783	
OSAC FRS	Yes	0.011		0.075	0.774	N/A	0.140
	No			N/A	N/A	0.689	0.300
OSAC FRS (common inconclusive)	Yes			0.075	0.727	N/A	0.187
	No			N/A	N/A	0.802	
OSAC FRS alternative	Yes	0.009		0.075	0.776	N/A	0.140
	No			N/A	N/A	0.691	0.300
OSAC FRS alternative (common inconclusive)	Yes			0.075	0.729	N/A	0.187
	No			N/A	N/A	0.804	

[1] It is not possible to identify the wrong finger of the correct person when control prints from the true sources are not present.
[2] It is not possible to erroneously exclude the correct source when control prints from the true sources are not present.
[3] It is not possible to correctly associate the test latent prints with control prints from their true sources when those are not present.
[4] The design of the MDPD study made it impossible to have a rate of correct exclusion of a source when the true sources were provided in the test packets.

Table 2

Rates considered in simulation experiments.

Our model is designed to keep the total number of test cases (4360 for Phase 1 and 1760 for Phase 2) and the number of cases where control impressions from the true sources are provided (56 out of 80 cases) fixed across all experiments. Our model is based on the four following arguments and assumptions:

(1) The MDPD report indicates that 56 (out of 80) latent prints were presented with sets of impressions from the sources that yielded them (the remaining 24 latent prints were not compared against the sets of impressions from their true sources); however, the report does not specify how many of each type of comparison were present in the Phase 1 and 2 packets. To account for this lack of information, we initiate each simulation by generating a pseudo-set, *P*, of 80 latent prints where impressions from the true sources of the first 56 prints are considered to be available, and impressions from the true sources of the remaining 24 prints are considered to be absent. We then randomly sample 40 latent prints, P_1, from the entire set of 80 latent prints, *P* (each of the 80 prints has an equal chance of being selected for inclusion in P_1). Finally, we randomly sample 20 latent prints from the remaining 40 to create set P_{2A} (each of the remaining 40 prints has an equal chance of being selected for inclusion in P_{2A}) and designate the final 20 latent prints as set P_{2B}.

(2) During Phases 1 and 2 of the study, examiners were provided with a fixed set of test cases: 109 examiners were provided with 40 cases at the beginning of Phase 1; 88 examiners were provided with 20 cases at the beginning of Phase 2. However, only 4233 decisions were returned at the end of Phase 1 (out of a total of 4360 possible decisions), and 1730 decisions were returned at the end of Phase 2 (out of 1760 possible decisions). Given that some latent prints were selected to be more challenging than others and given that some examiners may have been busier than others, it is likely that the missing responses are concentrated on a small number of latent prints or a small number of examiners. However, the MDPD report does not provide information on which examiners did not finish the study and on which latent prints were favored. Assuming that all examiners are equally likely to complete the study and that all test cases are equally

likely to be considered by the participants, we model the number of test cases processed by an examiner during both Phases, n_{ri}, by binomial distributions (with parameters $n = 40$ cases and probability of processing a case of $p = 4233/4360$[1] in Phase 1; with parameters $n = 20$ cases and probability of processing a case of $p = 1730/1760$ in Phase 2).

(3) Not all of the latent prints for which a decision was returned ended up being compared to the prespecified sets of control prints. Some latent prints were deemed of no value and no conclusions were reported on their source (3210 latent prints were deemed of value out of 4233 decisions returned at the end of Phase 1; 1342 were deemed of value out of 1730 returned at the end of Phase 2). It is likely that challenging prints were more often deemed of no value; furthermore, it is also likely that some participants were more prone to deem that test cases should not be compared to control prints and that significant variability exists between the value determination of some latent prints by the participants. The MDPD report does not provide data on these elements. Assuming that all test cases for which decisions were made were equally likely to be deemed of value, and that all examiners were equally performing, we model the number of latent prints deemed of value in a given packet during Phases 1 and 2 by binomial distributions (with parameters $n = n_{ri}$, where n_{ri} is the number of decisions returned for the i^{th} packet, determined using the binomial distribution in (2) above, and probability of deeming the print of value $p = 3210/4233$ during Phase 1; and, $n = n_{ri}$ and $p = 1342/1730$ during Phase 2).

(4) According to the design of the MDPD study, any conclusion on the source of a test case that was deemed of value for comparison must belong to one of six categories (Table 1): (a) correct identification to the true donor; (b) identification to the wrong finger of the true donor; (c) identification to the wrong person; (d) inconclusive examination; (e) correct exclusion of the donors of all three sets of control prints;

[1] Note that it is possible to refine the model by accounting for the uncertainty on the probability of success parameter of the binomial distribution. However, this is outside of the scope of this paper and the point estimates obtained from the data provided by MDPD are used instead.

(f) incorrect exclusion of the true donor. Depending on whether control prints of the true donor are provided as part of the three sets of control prints, some of the categories described above cannot be used. For example, it is not possible to categorize a decision on the source of a test case as (e) when prints of the true donor are present among the provided control prints. In such case, exclusion of all donors automatically results in the decision being categorized as (f). Conversely, category (f) cannot be used when prints of the true donor are not provided.

It is likely that some of the more challenging test latent prints resulted in more inconclusive examinations than others. It is also possible that a large number of erroneous conclusions can be associated with a small number of participants. However, as mentioned above, the MDPD report does not provide data on these elements. Assuming that all latent prints are equally likely to be correctly associated to their true source and assuming that all participants are equally performing, we model the categorization of the decision resulting from the examination of a test case for which the control prints of the true source are provided by a multinomial distribution over the six categories (a) through (f) described above with vectors of probability parameters, R_{SP}, listed in Table 2. Similarly, we model the categorization of the decision resulting from the examination of a test case for which the control prints of the true source are absent by a multinomial distribution over the six categories (a) through (f) described above with vectors of probability parameters, R_{SA}, listed in Table 2.

Our model relies on four main assumptions:

(1) All 80 latent prints are equally likely to be selected for inclusion in Phase 1 and Phase 2 packets.

(2) All examiners are equally performing.

(3) All latent prints are equally likely to be deemed of value.

(4) All latent prints are equally likely to result in an error (false positive or false negative).

Based on the MDPD report, there is no indication that assumption (1) is unreasonable. All research to date on the performance of latent print examiners shows that assumption (2) is unlikely to be correct. Furthermore, by the design of the MDPD study, assumptions (3) and (4) are equally incorrect because they have purposely selected test prints with various levels of quality. However, we are interested in estimating the average FPR for a population of examiners and for a range of latent print quality. Thus, we believe that these assumptions are reasonable enough to enable us to estimate the average FPR. The appropriateness of these assumptions is tested below by using the model to repeatedly simulate the MDPD study and comparing the results of the simulations with the data observed by the MDPD project team. Because the results presented in Table 3 are similar to those presented in Table 1, we can conclude that, although incorrect, the assumptions are fit-for-purpose in that they allow the model to replicate the MDPD study.

		Source Present	**Source Not Present**	**Totals**
# of Latent Prints		56	24	80
# of Decisions		3188.48 [2983, 3369]	1362.71 [1169, 1563]	4551.19 [4479, 4612]
Correct IDs		2466.29 [2302, 2620]	N/A	2466.29 [2302, 2620]
Erroneous IDs	Correct Person, Wrong Finger	35.38 [24, 47]	N/A	35.38 [24, 47]
	Incorrect Person	4.06 [0, 7]	2.98 [0, 6]	7.04 [1, 11]
Inconclusive Examinations		447.41 [396, 490]	404.72 [335, 468]	852.12 [800, 918]
Correct Exclusions		N/A	955.01 [806, 1097]	955.01 [806, 1097]
Erroneous Exclusions		235.34 [202, 266]	N/A	235.34 [202, 266]

Table 3

Reproduced Miami-Dade results using simulations and plug-in estimates for the rates of the various conclusions calculated using the values in Table 1. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals.

Simulations

We report below the results from a total of seven experiments involving simulations with different rate vectors, R_{SP} and R_{SA}, for the categorization of the conclusions regarding the sources of the test cases. The first simulation uses point estimates of the rates calculated from the observed frequencies reported by MDPD (Table 1). This experiment is designed to verify that our model is able to reproduce the observations made by the MDPD research team, despite the simplifications and assumptions discussed in the previous section.

The remaining six experiments are associated with the three error rates proposed by the MDPD, the OSAC FRS, and our alternative to the OSAC FRS estimate that includes inconclusive decisions. We do not consider the 4.2% FPR proposed by MDPD, or any of the upper bounds on the MDPD FPR estimates proposed by PCAST [4], for reasons that will become clear later. We also do not formally test the estimates proposed by Ulery et al. [1], because they are close to the estimates proposed by OSAC FRS, and considering both sets of results would be redundant. We note that, aside from the first set of rates (based on the observed frequencies from the MDPD study), a single FPR is considered for all simulated comparisons, regardless of whether (a) the true source was present during the comparison process, or (b) the examiner identified the correct person, but the wrong finger. In our simulations, the chosen FPR corresponds to the error rate estimates proposed by MDPD, the OSAC FRS, and our alternative to the OSAC FRS estimate. The use of a single FPR relies on the assumption that every latent print has the same chance to result in an erroneous identification[2].

For each of the three proposed error rates, we conduct two experiments: in the first experiment, we consider the proposed FPRs described in section III and two inconclusive rates based on the frequencies observed in the MDPD study (446/3177 and 403/1359 in Table 1); in the second experiment, we consider the same FPRs and a common inconclusive rate based on the pooled frequencies observed in the MDPD study (849/4536 in Table 1).

Finally, the same false negative rate (FNR) is used throughout all six experiments. We used the one reported by MDPD. The true positive rate (TPR) is calculated by subtracting the sum of all other "source present" rates from 1 [e.g., for the Miami-Dade proposed rates, the TPR is calculated by 1 – (0.030 + 0.075 + 0.140) = 0.755]. Likewise, the true negative rate (TNR) is calculated by subtracting the sum of all other "source not present" rates from 1 [e.g., for the Miami-Dade proposed rates, the TNR is calculated by 1 – (0.030 + 0.30) = 0.0667]. The various rates associated with our experiments are reported in Table 2.

In each experiment, we simulate 1,000 iterations of the MDPD study to estimate expected counts for each category of decision as well as a 95% credible interval (Table 3).

[2] This assumption is not reasonable: lower quality latent prints have been shown to result in higher error rates [7]. However, the key question raised by the MDPD, as we will discuss below, is whether the rate of erroneous identifications is the same when control impressions are provided as when they are not.

Results of the Experiment Using Point Estimates from MDPD Report

To assess the reasonableness of the assumptions underlying our model, we begin by defining the set R : = $\{R_{SP}, R_{SA}\}$ of classification rates according to the rates specified in the Observed Frequencies row of Table 2, such that R_{SP} : = {0.011, 0.001, 0.074, 0.774, 0, 0.140} and R_{SA} : = {0, 0.002, 0, 0, 0.698, 0.300}. It is not possible to observe all of the categories presented in Table 2 (e.g., it is not possible to have a correct exclusion when the source is present in the comparison process, a correct identification when the source is not present in the comparison process, an erroneous exclusion when the source is not present in the comparison process), and so the unobservable categories are not applicable and are indicated in the tables throughout the paper by "N/A" and are assigned a zero probability in the set R.

When comparing the range of frequencies resulting from our experiment (Table 3) to the actual frequencies observed by MDPD (Table 1), we can see that our algorithm produces a reasonable model of the MDPD study, despite its assumptions and approximations. Thus, we proceed by using it to investigate the reasonableness of the proposed error rates.

Results of the Experiments Using the MDPD Proposed FPR of 3.0%

To test the MDPD proposed FPR of 3.0%, rate vectors R_{SP}, R_{SA} are defined according to the rates specified in the Miami-Dade and Miami-Dade common inclusive rows of Table 2. In the first experiment, we use the explicit rates proposed by Pacheco et al. [3] In the second experiment, we consider a common inconclusive rate. This enables us to test the consistency of the inconclusive rate depending on whether control prints from the true sources are provided.

When using the 3.0% FPR suggested by MDPD, the average number of erroneous identifications that are produced by the model (Tables 4, 5) is remarkably larger than the number of actual erroneous identifications observed by the MDPD team (Table 1): Table 1 reports a total of 42 erroneous identifications; Table 4 indicates that the number of erroneous identifications should have been between 109 and 157 with 0.95 probability, if the FPR was truly 3.0%; Table 5 indicates that the number of erroneous identifications should have been between 115 and 159 with 0.95 probability, if the FPR was truly 3.0%.

		Source Present	Source Not Present	Totals
# of Latent Prints		56	24	80
# of Decisions		3187.07 [3003, 3392]	1364.26 [1163, 1533]	4551.33 [4481, 4612]
Correct IDs		2405.07 [2252, 2554]	N/A	2405.07 [2252, 2554]
Erroneous IDs	Correct Person, Wrong Finger	N/A	N/A	N/A
	Incorrect Person	95.85 [76, 116]	40.74 [27, 54]	136.59 [109, 157]
Inconclusive Examinations		447.54 [402, 494]	403.73 [339, 468]	851.28 [785, 902]
Correct Exclusions		N/A	919.79 [791, 1051]	919.79 [791, 1051]
Erroneous Exclusions		238.61 [209, 270]	N/A	238.61 [209, 270]

Table 4

Average number of conclusions in each category using Miami-Dade FPR (3.0%) and FNR (7.5%)–with different rates of inconclusive examinations. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals from our 1,000 simulations.

		Source Present	Source Not Present	Totals
# of Latent Prints		56	24	80
# of Decisions		3188.44 [2986, 3369]	1364.26 [1165, 1549]	4552.70 [4490, 4616]
Correct IDs		2256.75 [2107, 2401]	N/A	2256.75 [2107, 2401]
Erroneous IDs	Correct Person, Wrong Finger	N/A	N/A	N/A
	Incorrect Person	95.50 [74, 114]	41.02 [28, 55]	136.53 [115, 159]
Inconclusive Examinations		597.33 [539, 648]	255.81 [213, 300]	853.13 [802, 904]
Correct Exclusions		N/A	1067.43 [905, 1210]	1067.43 [905, 1210]
Erroneous Exclusions		238.86 [205, 269]	N/A	238.86 [205, 269]

Table 5

Average number of conclusions in each category using Miami-Dade FPR (3.0%) and FNR (7.5%)–with common rate of inconclusive examinations. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals from our 1,000 simulations.

Furthermore, we note that reporting a single inconclusive rate is not a reasonable interpretation of the MDPD results. Although the total number of inconclusive decisions is consistent between Tables 1, 4, and 5, the number of inconclusive decisions when the source is present in the comparison process is larger in our experiment, and the number of inconclusive decisions when the source is not present in the comparison process is much smaller in our experiment.

The implications of these results are two-fold: first, it is clear that 3.0% is an overestimation of the true FPR of the participants of the MDPD study; second, examiners behave differently when impressions from the true sources are present in the comparison process. When control impressions are present, examiners make an inconclusive decision approximately 14% of the time, compared to approximately 30% of the time when control impressions are not present.

Results of the Experiments Using the OSAC FRS Proposed FPR of 1.1%

To test the FPR of 1.1% proposed by the OSAC FRS, we define *R* according to the rates specified in the OSAC FRS and OSAC FRS common inclusive rows of Table 2. Although we define the FPR according to the OSAC FRS's proposed error rate, we use the same FNR and inconclusive rates that are obtained from the MDPD study.

The results of these experiments show that the average number of erroneous identifications, obtained through our simulations (52 expected erroneous identifications in Tables 5 and 6), is slightly higher than the 42 observed by MDPD (Table 1). Because 42 is within the 95% credible intervals for both experiments (39 in Table 6, and 37 and 65 in Table 7), it may appear, at first, that an FPR of 1.1% for this study is not unreasonable. However, our results show that, while considering an FPR of 1.1% is reasonable when control prints from the true sources are presented with the test cases, this FPR is an overestimate of the true FPR when control prints from the true sources are not present: indeed, the number of erroneous identifications reported in Table 1 (3 erroneous identifications) is largely outside of the credible intervals resulting from our simulations (7 and 23 in Table 6, and 8 and 23 in Table 7). Thus, the FPR of 1.1% proposed by the OSAC FRS is not necessarily a fair estimate of the true FPR of the participants in the MDPD study, and our results suggest again that a single FPR may not be appropriate to interpret the MDPD observations.

Finally, we confirm that considering a single inconclusive rate for the MDPD study is not reasonable: the average number of inconclusive decisions when control prints from the true sources are present (596 in Table 7) is greater than the corresponding number in Table 1 (446), whereas the number of inconclusive decisions when control prints from the true sources are not present (256 in Table 7) is smaller than the corresponding number in Table 1 (403).

		Source Present	Source Not Present	Totals
# of Latent Prints		56	24	80
# of Decisions		3185.49 [3000, 3405]	1365.77 [1157, 1555]	4551.26 [4486, 4617]
Correct IDs		2462.62 [2298, 2633]	N/A	2462.62 [2298, 2633]
Erroneous IDs	Correct Person, Wrong Finger	N/A	N/A	N/A
	Incorrect Person	36.35 [24, 47]	15.49 [7, 23]	51.84 [39, 65]
Inconclusive Examinations		447.78 [402, 498]	405.75 [342, 473]	853.52 [797, 912]
Correct Exclusions		N/A	944.53 [808, 1091]	944.53 [808, 1091]
Erroneous Exclusions		238.74 [206, 272]	N/A	238.74 [206, 272]

Table 6

Average number of conclusions in each category using OSAC FRS (1.1%) and FNR (7.5%)–with different rates of inconclusive examinations. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals from our 1,000 simulations.

		Source Present	Source Not Present	Totals
# of Latent Prints		56	24	80
# of Decisions		3184.69 [2986, 3373]	1368.15 [1186, 1562]	4552.85 [4484, 4616]
Correct IDs		2313.27 [2167, 2467]	N/A	2313.27 [2167, 2467]
Erroneous IDs	Correct Person, Wrong Finger	N/A	N/A	N/A
	Incorrect Person	36.39 [25, 47]	15.66 [8, 23]	52.05 [37, 65]
Inconclusive Examinations		596.12 [540, 652]	256.50 [206, 297]	852.71 [803, 904]
Correct Exclusions		N/A	1095.99 [943, 1256]	1095.99 [943, 1256]
Erroneous Exclusions		238.83 [208, 271]	N/A	238.83 [208, 271]

Table 7

Average number of conclusions in each category using OSAC FRS (1.1%) and FNR (7.5%)–with common rate of inconclusive examinations. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals from our 1,000 simulations.

Results of the Experiments Using the Alternative FPR of 0.9%

To test the FPR of 0.9%, an alternative to the OSAC FRS estimate, which accounts for inconclusive decisions, we define *R* according to the rates specified in the OSAC FRS alternative and OSAC FRS alternative common inclusive rows of Table 2. We use the same FNR and inconclusive rates as in the previous experiments.

In Tables 8 and 9, we observe that the average number of erroneous identifications obtained from the simulations is very similar to the 42 observed by MDPD. Nevertheless, and contrary to the experiments performed using the 1.1% FPR proposed by OSAC FRS, we observe that, in this case, our simulations result in an average number of erroneous identifications when control prints from the true sources are present that is lower than the 39 observed by MDPD. Additionally, the average number of erroneous identifications when control prints from the true sources are absent is higher than the three observed by MDPD. Once again, our results suggest that a single FPR shared by the two scenarios may not be appropriate in this study and that examiners may behave differently when control prints from the true sources are presented together with the test latent prints.

Lastly, as noted in the two previous experiments, the results of this experiment lead to the conclusion that considering a single inconclusive rate is not reasonable for this study.

		Source Present	**Source Not Present**	**Totals**
# of Latent Prints		56	24	80
# of Decisions		3188.56 [3004, 3378]	1363.59 [1184, 1552]	4552.15 [4486, 4618]
Correct IDs		2472.12 [2327, 2627]	N/a	2472.12 [2327, 2627]
Erroneous IDs	Correct Person, Wrong Finger	N/A	N/A	N/A
	Incorrect Person	29.49 [19, 40]	12.62 [5, 19]	42.11 [29, 54]
Inconclusive Examinations		447.61 [399, 492]	403.35 [337, 461]	850.96 [789, 908]
Correct Exclusions		N/A	947.62 [815, 1088]	947.62 [815, 1088]
Erroneous Exclusions		239.35 [204, 269]	N/A	239.35 [204, 269]

Table 8

Average number of conclusions in each category using the FPR that accounts for inconclusive decisions (0.9%) and FNR (7.5%)–with different rates of inconclusive examinations. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals from our 1,000 simulations.

		Source Present	**Source Not Present**	**Totals**
# of Latent Prints		56	24	80
# of Decisions		3188.87 [2984, 3373]	1363.74 [1169, 1552]	4552.62 [4492, 4620]
Correct IDs		2322.80 [2174, 2479]	N/A	2322.80 [2174, 2479]
Erroneous IDs	Correct Person, Wrong Finger	N/A	N/A	N/A
	Incorrect Person	29.09 [18, 38]	12.58 [5, 19]	41.67 [28, 54]
Inconclusive Examinations		596.89 [542, 656]	255.11 [209, 299]	852.06 [798, 901]
Correct Exclusions		N/A	1096.06 [936, 1256]	1096.06 [936, 1256]
Erroneous Exclusions		240.03 [206, 269]	N/A	240.03 [206, 269]

Table 9

Average number of conclusions in each category using the FPR that accounts for inconclusive decisions (0.9%) and FNR (7.5%)–with common rate of inconclusive examinations. The numbers in [ ] represent the lower and upper bounds of maximum density 95% credible intervals from our 1,000 simulations.

Discussion and Conclusions

The interpretation of the data resulting from a study conducted by the Miami Dade Police Department (MDPD) and designed to estimate error rates in fingerprint examination [3] has generated some controversy. The core of the controversy resides in that the false positive error rate (FPR) of 3.0% proposed by the MDPD research team results from the ratio of two numbers observed under two different scenarios: (a) the total number of erroneous identifications made during the entire study; (b) the number of decisions rendered when control prints of the true sources of the test cases were not provided. In other words, the research team did not fully account for the decisions rendered when control prints of the true sources were provided.

As a result of the MDPD study, false positive error rate (FPR) estimates have been proposed by the MDPD and the Organization of Scientific Area Committees Friction Ridge subcommittee (OSAC FRS) [6]. Furthermore, an additional calculation based on the one from the OSAC FRS that includes inconclusive decisions is considered. In this paper, we model the behavior of the participants in the MDPD study in order to simulate the number of erroneous identifications that should be observed if the true FPR was close to any of the proposed estimates. Although our model relies on necessary simplifications and assumptions, the results of our first experiment show that the model can generate data that is very similar to the observations made by MDPD (Table 3), and that it can be considered to be adequate.

Overall, our different experiments show that none of the proposed FPRs are good estimates of the true FPR of the study's participants:

(1) The 3.0% FPR proposed by MDPD largely overestimates the true FPR (Tables 4, 5).

(2) The 1.1% FPR proposed by the OSAC FRS [6] seems larger than the FPR of the participants examining test cases that were not associated with control prints from the true sources of the latent prints (Tables 6, 7).

(3) The 0.9% FPR that accounts for inconclusive decisions appears to be marginally smaller than the FPR of the participants examining test cases that were associated with control prints from their true sources, but marginally larger than the FPR of the participants examining test cases that were not associated with control prints from the true sources (Tables 8, 9).

Although the OSAC FRS [6] and the Canadian case study [5] were correct to point out the mistake made by the MDPD and PCAST when calculating their estimates of the FPR, their analyses of the original MDPD error rate estimate and their attempts to resolve the calculation for the FPR are still incorrect. The FPR calculations proposed by the OSAC FRS [6] and the alternative that accounts for inconclusive decisions certainly provide better estimates; yet, our simulations show that neither of them is satisfactory.

The main difficulty with the interpretation of the MDPD data is that participants of their study seem to have two distinctly different levels of performance, depending on whether control prints from the true sources of the test latent prints are provided or not. This observation stems from the wildly different estimates of the rates at which participants deem that examinations are inconclusive (14% when the control prints from the true source are provided vs. 30% when they are not) and at which erroneous identifications occur (~1.2% when the control prints from the true source are provided vs. ~0.1% when they are not) (Tables 1, 2). From this latter observation, and from the results of our experiments, it appears that reporting a single FPR and a single rate of inconclusive examinations based on the data acquired during the MDPD study is inappropriate.

To properly interpret the data acquired by the MDPD and reported in Table 1, we propose two alternative solutions. Both solutions consider that the participants of the study have two distinct true FPRs:

(1) First, we can consider that examiners do make more false identifications when they are unknowingly provided with control impressions from the true sources than when they are not. This solution is counterintuitive but explains the data. In this case, a Bayesian analysis of the false identification rates (using a flat Beta prior distribution on the FPR parameter and a binomial likelihood) assigns an upper bound for the 95% credible interval of the FPR of 1.7% when impressions of the true source are provided, and of 0.6% when impressions of the true source are not provided. The maximum a posteriori (MAP) estimates for these two intervals are 1.2% and 0.2%, respectively.

(2) Second, we may consider that there are two types of false identifications, consisting in (a) false identifications made to the correct person, but to the incorrect finger and (b) false identifications made to the incorrect person regardless of whether the examiner is presented with impressions from the true sources[3]. This distinction between false identifications parallels the argument made by Koehler, that "not all false positive errors are equal" [8]. A Bayesian analysis of these two scenarios results in an upper bound for the 95% credible interval of 1.1% chance to erroneously identify the wrong finger of the correct source and of 0.3% chance to erroneously identify the wrong person. The MAP estimates for these two intervals are 0.7% and 0.2%, respectively.

Both of our solutions result in FPR estimates that are much more comparable to the ones proposed by Ulery et al. [1] We note that the worst-case estimate resulting from our interpretation of the MDPD data (1.7%) is still much lower than the PCAST 5.4% upper bound or the MDPD estimate of 4.2%.

Our analysis of the MDPD data raises questions related to the dependability of examiners' conclusions. The design of the MDPD study certainly results in a more challenging analysis of the data and in some controversy regarding the estimation of the FPR; nonetheless, it allows for highlighting an important issue that did not appear in the Ulery et al. [1] study, namely, the large number of erroneous identifications made to the wrong finger of the correct person. We are deeply concerned with the very large number of cases (35 out of 3177 decisions) where the test latent prints were associated with the wrong finger of the correct source, in particular when this number is compared to the number of erroneous identifications to the wrong source (7 out of 4536 decisions). As we mentioned previously, we do not want to speculate on the reason(s) behind such a large discrepancy. We also realize that the experiment did not account for the verification stage, which would, we would hope, catch clerical errors. However, the MDPD data show that the number of this type of error far outweigh the number of erroneous identifications to the incorrect person. If these errors are true clerical errors, we believe that they could easily be avoided in the first place. If they are the manifestation of a more complex issue,

[3] It is trivial enough to show that there is no statistically significant differences between the two rates of erroneous identification to the incorrect person based on the estimates of 4/3177 and 3/1359.

adequate training and procedures should be designed to remedy the problem. Similarly, we are concerned with the large discrepancy between the two distinct rates of inconclusive examinations observed during the MDPD study. This indicates that the exclusion process is different than the identification process and that it is not necessarily well understood and implemented in practice. Overall, we believe that it is critical to investigate and address the reasons behind these different behaviors of the fingerprint examiners and limit their impacts on the fingerprint examination process.

For further information, please contact:

Dr. Cedric Neumann, Associate Professor of Statistics
Department of Mathematics and Statistics
South Dakota State University
Box 2225
57007 Brookings, SD
Cedric.Neumann@me.com

This project was supported in part by Award No. 2014-IJ-CX-K088 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect those of the Department of Justice.

References

1. Ulery, B. T.; Hicklin, R. A.; Buscaglia, J.; Roberts, M. A. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. *Proc. Nat. Acad. Sci.* **2011**, *108* (19), 7733–7738.
2. Langenburg, G. M. A Critical Analysis and Study of the ACE-V Process. Ph.D. Thesis, University of Lausanne, Switzerland, 2012.
3. Pacheco, I.; Cerchiai, B.; Stoiloff, S. *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations*. NCJRS document No. 248534, 2014.
4. President's Council of Advisors on Science and Technology (PCAST). *Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*; Executive Office of the President's Council of Advisors on Science and Technology: Washington, D.C., 2016.

5. Wilkinson, D.; Richard, D.; Hockey, D. Expert Fingerprint Testimony Post-PCAST - A Canadian Case Study. *J. For. Ident*. **2018**, *68* (3), 299–331.
6. Organization of Scientific Area Committees (OSAC) Friction Ridge Subcomittee. Response to the President's Council of Advisors on Science and Technology's (PCAST) Request for Additional References: Washington, D.C., December 2016.
7. Ulery, B. T.; Hicklin, R. A.; Roberts, M. A.; Buscaglia, J. Measuring What Latent Fingerprint Examiners Consider Sufficient Information for Individualization Determinations. *PLoS ONE* **2014**, *9* (11), e110179.
8. Koehler, J. J. Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences. *AZ St. Law J.* **2017**, *49*, 1369–1416.

Appendix

```
Algorithm 1: Simulation to reproduce phase 1 of the Miami-Dade study
Define: $R_{SP}$, a set of point estimates for the classification rates when control impressions from the true source are present in the comparison process; $R_{SA}$, a set of point estimates for the classification rates when control impressions from the true source are not provided (absent) in the comparison process; $P$, a set of 80 latent prints (to correspond to the MDPD study (p. 28) we consider, without loss of generality, that control impressions from the true source are provided for prints 1 through 56, and are not provided for prints 57 through 80);
for Phase 1 do
    Sample 40 latent prints, $P_1$, from $P$;
    for $i \in 1 : 109$ packets do
        Sample $n_{r_i} \sim \text{Binomial}\big(40, \frac{4233}{4360}\big)$, the number of decisions that will be returned in packet $i$;
        Sample $n_{v_i} \sim \text{Binomial}\big(n_{r_i}, \frac{3210}{4233}\big)$, the number of decisions that are determined to be of value in packet $i$;
        Sample $n_{v_i}$ latent prints, $P_{n_{v_i}}$, from $P_1$ to determine which prints are considered in packet $i$;
        Sample $d_{SP_i} \sim \text{Multinomial}\big(1, \sum_{P_1} I\{P_{n_{v_i}} \in P_{SP_i}\}, R_{SP}\big)$, the decisions made for $P_{SP_i}$, the set of latent prints in packet $i$ whose source is provided;
        Sample $d_{SA_i} \sim \text{Multinomial}\big(1, \sum_{P_1} I\{P_{n_{v_i}} \in P_{SA_i}\}, R_{SA}\big)$, the decisions made for $P_{SA_i}$, the set of latent prints in packet $i$ whose source is not provided;
    end
    Calculate $n_{SP}^{(1)} = \sum_{i=1}^{109} d_{SP_i}$, the total count for each decision classification when the source is present;
    Calculate $n_{SA}^{(1)} = \sum_{i=1}^{109} d_{SA_i}$, the total count for each decision classification when the source is not present;
    Return: $n_{SP}^{(1)}$ and $n_{SA}^{(1)}$.
end
```

```
Algorithm 2: Simulation to reproduce phase 2 of the Miami-Dade study
Define: $R_{SP}$, a set of point estimates for the classification rates when control impressions from the true source are present in the comparison process; $R_{SA}$, a set of point estimates for the classification rates when control impressions from the true source are not present (absent) in the comparison process; $P_2 := P \backslash P_1$;
for Phase 2 do
    Sample 20 latent prints, $P_{2A}$, from $P_2$;
    Define $P_{2B}$, the remaining set of 20 latent prints;
    for group $k \in \{A, B\}$ do
        for $i \in 1 : 44$ packets do
            Sample $n_{r_{ik}} \sim \text{Binomial}\big(20, \frac{1730}{1760}\big)$, the number of decisions that will be returned in packet $i$ for group $k$;
            Sample $n_{v_{ik}} \sim \text{Binomial}\big(n_{r_{ik}}, \frac{1342}{1730}\big)$, the number of decisions that are determined to be of value in packet $i$ for group $k$;
            Sample $n_{v_{ik}}$ latent prints, $P_{n_{v_{ik}}}$, from $P_{2k}$ to determine which prints are considered in packet $i$ for group $k$;
            Sample $d_{SP_{ik}} \sim \text{Multinomial}\Big(1, \sum_{P_{2k}} I\{P_{n_{v_{ik}}} \in P_{SP_{ik}}\}, R_{SP}\Big)$, the decisions made for $P_{SP_{ik}}$, the set of latent prints in packet $i$ for group $k$ whose source is provided;
            Sample $d_{SA_{ik}} \sim \text{Multinomial}\Big(1, \sum_{p_1} I\{P_{n_{v_{ik}}} \in P_{SA_{ik}}\}, R_{SA}\Big)$, the decisions made for $p_{SA_{ik}}$, the set of latent prints in packet $i$ for group $k$ whose source is not provided;
        end
    end
    Calculate $n_{SP}^{(2)} = \sum_{k \in \{A,B\}} \sum_{i=1}^{44} d_{SP_{ik}}$, the total count for each decision classification when the source is present;
    Calculate $n_{SA}^{(2)} = \sum_{k \in \{A,B\}} \sum_{i=1}^{44} d_{SA_{ik}}$, the total count for each decision classification when the source is not present;
    Return: $n_{SP}^{(2)}$ and $n_{SA}^{(2)}$.
end
```

```
Algorithm 3: Simulation to reproduce Miami-Dade study
Define: $R_{SP}$, a set of point estimates for the classification rates when control impressions from the true source are present in the comparison process; $R_{SA}$, a set of point estimates for the classification rates when when control impressions from the true source are not present (absent) in the comparison process; $p$, a set of 80 latent prints; $N$, the number of simulations (typically $N \geq 1000$);
for $i \in 1{:}N$ do
    Simulate phase 1 according to algorithm 1, with $R_{SP},\ R_{SA},\ P$ and $N$ as defined above;
    Simulate phase 2 according to algorithm 2,with $R_{SP},\ R_{SA},\ P$ and $N$ as defined above;
end
Calculate $\bar{n}_{SP} = \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N} n_{SP_j}^{(i)}$, the mean count of each decision classification when the source is present;
Calculate $\bar{n}_{SA} = \frac{1}{N} \sum_{i=1}^{2} \sum_{j=1}^{N} n_{SA_j}^{(i)}$, the mean count of each decision classification when the source is not present;
Return: $\bar{n}_{SP}$ and $\bar{n}_{SA}$.
```

Special Feature*

Letters Regarding:

Review of Several False Positive Error Rate Estimates for Latent Fingerprint Examination Proposed Based on the 2014 Miami-Dade Police Department Study

Letter from Glenn Langenburg
Elite Forensic Services, LLC
St. Paul, MN

I am grateful for the authors' (hereafter "AHN" [1]) re-evaluation of the MDPD data [2] and initiation of this important discussion on the topic of error rates. While there exists a body of work discussing error rates and their value [3-5], to date, there has been little exploration of the actual nuances and difficulties of computing error rates that arise from study designs that attempt to mimic casework scenarios. Furthermore, as AHN point out, PCAST has strong recommendations for how error rates must be communicated to triers of fact [6]. I agree with PCAST that error rates can be valuable tools to assess foundational validity of forensic tests; however, I believe PCAST to be perhaps a bit misguided (and maybe even biased in their views) when offering their overly authoritative requirements on communicating error rates. I wish to raise four issues tangential to the AHN paper, but should be discussed by the community of experts when addressing the computations and communications of error rates.

* Editor's Note: After the review process for the paper "Review of Several False Positive Error Rate Estimates for Latent Fingerprint Examination Proposed Based on the 2014 Miami-Dade Police Department Study" was completed and the paper was accepted for publication (included in this issue of the *JFI*), the editor and authors invited various individuals to provide their comments regarding their reactions to the paper. These reactions were encouraged to address any aspect of the paper or to address any aspect of the subject in general. This was done to encourage the discussion regarding the use of statistics in establishing error rates for latent fingerprint examinations. Five responses were received and are included in an appendix at the end of this paper. These letters are being published as received without review or editing.

An Accurate Denominator Is Challenging to Compute When the Study Design Is Complex

As AHN, MDPD, OSAC-FRS, Wilkinson, et al. 2017 [1,2,7,8] all point out, one can choose different denominators to compute error rates when the study design departs from a 1:1 (single latent print to single control fingerprint) format. When researchers, such as MDPD, attempt to design a study that is ecologically valid and more closely mimics scenarios in casework, this introduces an important question: what is the proper denominator? This is not an easy question to answer. In fact, AHN offer an innovative solution by reporting two error rates: one FPR when the source was present (~1%) and one FPR when the source was not present (~0.1%).

If we consider a study design such as MDPD's (and also Langenburg, 2009 [9], and Gutowski, 2006 [10]), where, for example, 3 fingerprint cards, each containing 10 separate fingers to compare is provided to a participant, then there are up to 30 comparisons that must be performed. A participant however may compare and exclude 20 fingers before finding the true source in the 21st finger. For the same trial, a different participant may start by chance, with the 21st finger and perform only 1 comparison. Tracking the actual number of comparisons and decisions is exceptionally difficult and likely can only be done with the aid of technology (which would depart from a more "case-like design").

Furthermore, a critical question is whether we calculate the error rate on a "per source" basis or "per finger" basis. Previous literature seems to use a "per finger" approach for FPR and "per source" approach for FNR. This can significantly under- or over-estimate an error rate. PCAST in fact makes this error in one of their calculations, severely overestimating the FPR in Langenburg 2009 [see PCAST p. 92 and p. 98].

To use the MDPD study design as an example for a "per finger" approach, there are 29 incorrect answers and 1 correct answer in same source trials, and 30 incorrect answers in different source trials. If a participant reports the wrong finger (1 of 29), even if they have the correct subject (1 of 3), the participant is scored an error. They have the correct subject, but the wrong finger. As AHN rightly point out, this may be due to a "clerical" error (also referred to as transposition transcription errors in Gutowski and Wertheim, et al. [10,11]) but also could be an erroneous individualization due to the similarity of an adjacent finger.

If we accept this calculation approach (per finger), then we can see that it will significantly impact the FPR. When we calculate the true negatives, if we maintain a “per finger” approach then every exclusion of a subject is equivalent to 10 fingers excluded. If 10 fingers and 2 palms were provided for each subject (as in MDPD, Langenburg, Gutowski [2,9,10]), then each exclusion of a subject is now equivalent to 12. See Table 1 below. We could further subdivide the exemplars into 10 finger joints (as provided in Langenburg [9]), and even specific regions of the palm. In other words, a single exclusion decision of a subject, could theoretically result in dozens of true negative decisions.

On the other hand, we could approach the computations from a “per source” approach, and therefore an exclusion decision to a subject is equivalent to 1 true negative, regardless of the number of fingers/palms/joints, etc. compared. However, if we take that approach then the MDPD study does not have 42 (“per finger”) false positives, but rather only has 7 false positive errors (as addressed by OSAC FRS and AHN).

	Ground Truth State		**Ground Truth State**	
Test Result	**Source Present**	**Source Not Present**	**Source Present**	**Source Not Present**
“Identification”	100	1	100	1
“Exclusion”	10	100	10	1200
Totals	110	101	110	1201
	In these data, an exclusion is recorded per source		In these data, an exclusion is recorded after excluding all 10 fingers and 2 palm prints per source	

Table 1

This data set for a theoretical error rate computation could result in drastically different FPRs depending on how the true negatives are scored. From a per source calculation, the FPR is 1 / 101 = ~1% FPR. From a per finger/palm calculation, the FPR is 1 / 1201 = 0.08% FPR. The per finger calculation better approximates the actual number of comparisons performed; the per source calculation gives an indication of the rate of erroneously associated subjects.

What is curious is that the literature on error rates seems to use a “per finger” approach to false positive errors and the “per source” approach to true negatives. This, in my view, is artificially inflating the error rate. If each of these 30 comparisons had been presented to the participant one after another on screen, the participant would be scored for 30 trials. Therefore it makes sense to compute error rates on a “per finger” basis, which tests the skill of the examiner. However, PCAST conflates this issue by intimating that error rates tell the trier of fact “something” about the likelihood the defendant (a “per source” approach) has been wrongly accused because he has been incorrectly associated with evidence in the case. So the real question is: what are we concerned with— the accuracy of examiners in performance tests, or the likelihood an individual will be incorrectly associated with evidence in a criminal trial? From the perspective of a researcher and author of these studies, I am interested in the former. The latter is out of my purview and provides too many variables for which I cannot control.

Discovery Rates

Another factor that affects the denominator in error rate calculations is the inclusion or exclusion of "inconclusive" decisions. My view is that 1) inconclusive decisions are viable, useful, and part of the ACE-V process and therefore should be included in the computations; 2) most of the time their inclusion/exclusion does not significantly affect the FPR in the available, published error rate studies, but rarely is it discussed that it does tend to impact the FNR[1].

PCAST spends significant effort discussing the inclusion or exclusion of inconclusive decisions. The report goes so far as to recalculate, albeit in the footnotes, all the FPRs with inconclusive decisions included. It is surprising to me that they did not promote and educate the reader to discovery rates. They reference discovery rates as a "Bayesian approach" (PCAST, p. 153-154), requiring a prior odds, but discovery rates can be computed from the available error rate studies.

Discovery rates are conditional probabilities that are conditioned on the result of the test (i.e. identification, exclusion); this is different from error rates which are conditional probabilities that are conditioned on the ground truth state of the test (i.e. same source or different source). By reporting discovery rates, the entire debate about the impact of inconclusive decisions can be avoided. In this scenario, a false positive discovery rate is the probability of a false positive given that "identification" was reported. Because the proportion of false positives are conditioned on the "identification" test result (i.e. the rows in Table 1), it is inconsequential how many inconclusive decisions were made, since they are not factored into this calculation. However, when reporting error rates, which are conditioned on the ground truth state (i.e. the columns in Table 1), then the number of inconclusive trials is part of the sum of total "same source" or "different source" trials.

[1] AHN provided insight here: there is a disproportionate number of inconclusive decisions when the images are coming from same source. This is likely indicating a different decision process and decision threshold for exclusion decisions compared to identification decisions.

It is also surprising that PCAST did not promote the value of discovery rates since they represent an answer to perhaps the most relevant question that PCAST was exploring: “When a fingerprint examiner reports an “identification” decision, what is the probability he is wrong?” (a false positive discovery rate). This is different (it is the transposed conditional) from an error rate which can be described as “When fingerprints are coming from two different people, what is the probability that the examiner will incorrectly report an “identification”. This posterior discovery rate tells us something about the performance of the expert, whereas the error rate tells us something about the efficiency of the test.

Two Sided Confidence Intervals

I have no issues with PCAST suggesting the reporting of confidence intervals. Confidence intervals are useful when assessing the result from any particular study. When there is a single point estimate, confidence intervals tell us something about the uncertainty associated with that result. Without knowing a single detail of a survey, if I said the results of a poll were 50% +/- 2%, versus a second survey where the results of the poll were 50% +/- 25% (both at a 95% level of confidence), we would look at these results, and probably use them, differently if incorporating them into some sort of decision that we had to make. It is easy to see how triers of fact could benefit from an error rate presented with a confidence interval. If the trier of fact learns that an error rate from a study was 1% +/- 0.1% (at 95% level of confidence) they may infer this to be a more precise estimate of the true error rate than a second study with a much wider confidence interval.

The issue that I have with PCAST is their recommendation that only the 1-sided confidence interval should be presented. Instead of a +/- approach giving a range, they prefer that testifying experts only say “may be as high as”. Their reasoning is even more dubious, in that laypeople may be too confused by confidence intervals and thus will anchor to only the lower bound of the interval and ignore the upper bound. I do not support this view.

Firstly, if we are so worried that triers of fact will have trouble understanding confidence intervals in the first place, this is a good argument for avoiding them in routine testimony. Confidence intervals can be included in reports, or more detailed admissibility hearings, where lawyers may be more prepared to discuss confidence intervals in a more meaningful manner. Secondly, I bristle at the idea that I should only present half of the interval (the most favorable half to the defendant) rather than present the entire interval (which most fairly represents the results of a particular study). Again, I refer to the previous point. If we are so worried that lay people will be confused and misunderstand these statistics, then perhaps we should avoid them altogether instead of withholding information from triers of fact because we do not trust them.

How can error rates be introduced through testimony in an adversarial system?

PCAST requires that if experts continue to provide conclusions of "identification" to triers of fact, then the fingerprint expert must also introduce error rates and confidence intervals. PCAST provides examples of this sort of testimony. I find this to be a very naive suggestion on the part of PCAST.

Unfortunately, I have little control over the questions that are asked of me on the witness stand. When this is coupled with two attorneys, who are non-scientists, and who do not want to confuse themselves or triers of fact by discussing statistics, it becomes very unlikely that they will want to introduce error rates and confidence intervals.

PCAST states:

> Conclusions of a proposed [fingerprint] identification may be scientifically valid, provided that they are accompanied by accurate information about limitations on the reliability of the conclusion—specifically, that (1) only two properly designed studies of the foundational validity and accuracy of latent fingerprint analysis have been conducted, (2) these studies found false positive rates that could be as high as 1 error in 306 cases in one study and 1 error in 18 cases in the other, and (3) because the examiners were aware they were being tested, the actual false positive rate in

> casework may be higher[2]. At present, claims of higher accuracy are not warranted or scientifically justified. Additional black-box studies are needed to clarify the reliability of the method [PCAST, p. 101-102].

It is amusing to me that if asked "And Dr. Langenburg, what was your conclusion in this case"? that somehow I will be able to opine all of the above in my answer. This simply is unrealistic and shows an unfamiliarity with the courtroom milieu in the U.S. To share the information above—which I do believe error rates are important to share—I need the cooperation of the attorney. In my experience, they tend to choose not to raise the issue and often express the view "We will just see if the other side brings it up on cross-examination". As a result, I do not have an opportunity to discuss error rates. My role in the court is clear: "answer only the questions asked of me, clearly, simply, and as briefly as needed".

Information regarding error rates and confidence intervals is included in my reports, but often my report will not be introduced as an exhibit and it is the testimony elicited from me that will form the record. It is now the responsibility of the prosecutor to ask about error rates (he of course won't) and then it is left to the defender to ask about them (he may be unprepared and unwilling to venture into statistics). The questions are not asked.

If the reader takes issue with this and thinks it IS my responsibility to raise the issue, then I suggest the reader is not one who testifies in the courtroom. This reader has likely never drawn objections and evoked the ire of an attorney or judge for not answering the question specifically asked. The attorney is in control of the presentation of the case—not me. If the reader disagrees, then please take issue with the U.S. adversarial system. The adversarial system is a terrible venue for scientists to attempt to explain difficult scientific issues and statistics. I am not able to lecture to the jury and I am not in control of where the testimony goes. Two adversarial lawyers are holding the reins. I will refer the reader to a wonderfully insightful paper by Prof. Pierre Margot, wherein he notes that "the first great bias" is the adversarial system [12]. PCAST seems to miss this point. PCAST's requirement should be directed towards attorneys to ask these questions. If asked, I am happy to discuss error rates.

[2] PCAST neglects to consider the possibility that the error rate in casework may be lower. Why not say "but these are simulated experiments and the actual error rate in casework is unknown at this time, but it is being inferred from these experiments."

However, I must be asked. In my experience, attorneys (for both sides) do not seem to want to venture into these waters.

Conclusion

I am grateful for AHN to open this discussion on error rates. Even though I have leveled some criticisms of the PCAST report, I wish to be clear that in fact, I do support many of their views. In the absence of statistical models, if fingerprint experts continue to report categorical conclusions (e.g. "identification", "exclusion"), error rates are valuable for considering the overall reliability of a "black box" methodology. I support PCAST here, but I take issue with a few of their specifics, namely:

1) I support reporting the two-sided confidence interval.
2) I support AHN's analysis of the MDPD and believe that the two different error rates (when the source was present and when the source was not present) is a novel way of viewing these statistics.
3) I support the communication of discovery rates over error rates (thus circumventing the argument whether to include or exclude "inconclusive decisions").
4) I support the introduction of error rates to triers of fact, but the responsibility lies with the attorneys. I have done my part by including them in my report, but it is their responsibility to raise these issues since they control the line of questioning during testimony.

References

1. Ausdemore, M. A.; Hendricks, J. H.; Neumann, C. Review of Several False Positive Error Rate Estimates for Latent Fingerprint Examination Proposed Based on the 2014 Miami Dade Police Department Study. *J. For. Ident.* **2019**, *69* (1), 59–81**.**

2, Pacheco, I.; Cerchiai, B.; Stoiloff, S. *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations*. NCJRS document No. 248534, 2014.

3. Koehler, J. J. Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter. *Hasting Law J.* **2008**, *59*, 1077–1098.

4. Haber, L.; Haber, R. N. Error Rates for Human Latent Fingerprint Examiners. In *Automatic Fingerprint Recognition Systems*; Ratha, N., Bolle, R., Eds.; Springer: NY, 2004; pp 339–360.

5. Expert Working Group on Human Factors in Latent Print Analysis. *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*; U.S. Department of Commerce, National Institute of Standards and Technology. 2012.
6. President's Council of Advisors on Science and Technology (PCAST). *Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*; Executive Office of the President's Council of Advisors on Science and Technology: Washington, D.C., 2016.
7. Organization of Scientific Area Committees (OSAC) Friction Ridge Subcomittee. Response to the President's Council of Advisors on Science and Technology's (PCAST) Request for Additional References Regarding: President's Council of Advisors on Science and Technology Report to the President. Dec 2016.
8. Wilkinson, D.; Richard, D.; Hockey, D. Expert Fingerprint Testimony Post-PCAST - A Canadian Case Study. *J. For. Ident*. **2018**, *68* (3) 299–331.
9. Langenburg, G. A Performance Study of the ACE-V Process: A Pilot Study to Measure the Accuracy, Precision, Reproducibility, Repeatability, and Biasability of Conclusions Resulting from the ACE-V Process. *J. For. Ident*. **2009**, *59* (2), 219–257.
10. Gutowski, S. Error Rates in Fingerprint Examination: The View in 2006. *For. Bulletin*, 2006, pp 18–19.
11. Wertheim, K.; Langenburg, G.; Moenssens, A. A Report of Latent Print Examiner Accuracy During Comparison Training Exercises. *J. For. Ident*. **2006**, *56* (1) 55–93.
12. Margot, P. Forensic Science on Trial-What is the Law of the Land? *Australian J. For. Sci*. **2011**, *43* (2–3), 89–103.

Letter from Igor Pacheco, Brian Cerchiai,
and Stephanie Stoiloff
Miami-Dade Police Department
Doral, FL

President's Council of Advisors on Science and Technology (PCAST), Miami-Dade Police Department Forensic Services Bureau (hereinafter referred to as MDPD), 3.0% False Positive Rate (FPR) & Organization of Scientific Area Committees (OSAC) Friction Ridge Subcommittee (FRS) 1.1% FPR

In 2014, the authors of the Miami-Dade Police Department (MDPD) study conducted research to determine different error rates associated with latent fingerprint examinations. In brief, the study was unique in that test trials (comparisons) were designed to simulate real case work conditions. Participants were asked to perform an Analysis, Comparison, Evaluation and Verification (ACE-V) and record their conclusions. Latents were assigned difficulty scores, test trials were mated and non-mated, and participants compared latents against multiple fingerprint/palm print standards, which required a ”searching” component to their examinations. Test trials were given under both biased and unbiased conditions, participants verified each other's answers, and in some cases they were given back their own answers to determine if they could repeat their conclusions. As many as 109 examiners from different local, state, and federal agencies participated across the United States. Participants included both International Association for Identification (IAI) latent certified and non-certified examiners with a broad range of latent print examination experience.

At the conclusion of the study, several different False Positive Rates (FPR) and False Negative Rates (FNR) were reported. Mainly, the FPR and FNR for both ACE and ACE-V was reported with and without inconclusive decisions recorded in the calculations. The reported 3.0% FPR (including inconclusives decisions in the calculation) for ACE in the MDPD study and the 1.1% FPR by OSAC (excluding inconclusives decisions in the calculation) using the data reported in the MDPD study is discussed.

The 2016 PCAST report cites several studies in reporting false positive error rates in latent print examinations. The PCAST reports on the Miami-Dade study's ACE FPR, and concludes that errors could occur, “as high as 1 in 18 cases.” We strongly disagree with PCAST's reported error rate in latent print casework for some of the same reasons pointed out by the OSAC FRS in their response to the PCAST report. Nonetheless, the PCAST's report on latent print error rates has caused great concern within the latent print community. Researchers with both the Royal Canadian Mounted Police (RCMP) and members of the OSAC FRS have raised issue with how the 3.0% FPR for ACE was calculated, as well as the failure of the PCAST to, "detect the calculation error in the false positive rate reported by Miami Dade." The researchers with the RCMP and the OSAC FRS point out that in calculating the FPR for ACE in the MDPD study, consideration should have been given to all 42 erroneous identifications from both the total number of decisions that were made when the source was and was not present. This is due to the fact that 39 of the erroneous IDs occurred when the source was present and 3 occurred when the source was not present. Normally, the FPR would be calculated as was done in the MDPD study; however, since the study was unique and involved a search to 3 sets of fingerprint/palm print standards (as opposed to a 1 to 1 side-by-side comparison), it was also possible to make an erroneous ID when the source was present in one of the three standards provided to the participants for comparison. Accordingly, the OSAC FRS estimates a 1.1% FPR for ACE by including the total number of decisions for both source present and source not present trials.

The authors of the MDPD study agree that the FPR for ACE should have considered both source present and source not present trials. Although the FPR for ACE is reduced from 3.0% to 1.1% (with and without inconclusives, respectively), these error rates must be further evaluated. Aside from reasons noted in the article, an additional point of consideration is that both of these error rates assume that each participant made only one decision per trial. That is to say, a participant compared a latent print impression to a single finger or palm print, even though 3 different donor standards containing a total of 30 fingers and 6 palm prints were provided for comparison. Due to the fact that the number of comparisons a participant conducted before making a final decision was unknown, each participant's decision was only counted once and the assumption was the most conservative. However, the possibility that participants did not make multiple comparisons before reaching a final decision is not plausible. On the contrary, it is highly likely the FPR is even lower than the original 3.0% reported in the MDPD study or the OSAC's 1.1% FPR estimate. For example, if each participant's correct exclusion decision for different source trials was counted three times instead of once (since each participant correctly excluded three standards), an alternative FPR of .75% (without inconclusives) could have been estimated. As noted in the article and by the OSAC, different FPR estimates can be calculated for a variety of conditions and we agree that a single FPR estimate for the discipline is not reasonable. Alternative calculations for FPR estimates include, but are not limited to, source present vs. source not present; distinguishing erroneous IDs to the correct donor (but wrong finger or palm) from erroneous IDs to the wrong donor; separating errors based on the quality or difficulty of the latent impression; and whether to include the total number of inconclusive conclusions in the FPR calculation.

The PCAST report also chose to only report the FPR for ACE and not ACE-V in the MDPD study, stating, "The Miami-Dade study involved a small test of verification step, involving verification of 15 of the 42 false positives. In these 15 cases, the second examiner declared 13 cases to be exclusions and 2 to be inconclusive. The sample size is too small to draw a meaningful conclusion. And, the paper does not report verification results for the other 27 false positives." Although the sample size was small, we disagree the results are not meaningful. Of the 42 false positive errors reported from Phases 1 and 2 of the MDPD study, 25 errors occurred in Phase 1 and 17 errors occurred in Phase 2. Phase 1 false positive errors were not sent for normal verification in Phase 3 and instead sent out as a *second verification* in the MDPD study to test for bias and repeatability. This was not the case for false positive errors that occurred in Phase 2 in which the participants were divided into two subgroups and each group was assigned two different sets of 20 latent prints to examine. Since the participants were divided into two subgroups, it was possible to send false positive errors for latents not yet examined from one group to the other group for normal verification and vice versa. Of the total 17 false positive errors from Phase 2 that were sent out for normal verification in Phase 3, 15 responses were returned. The sample size is small, however, the results are meaningful in that of the possible false positive errors not already examined previously by the same participant in Phase 1, none of the participants agreed with any of the false positive errors reported in Phase 2. We agree with the OSAC FRS response regarding the omission by PCAST of the verification step in the examination methodology, "The PCAST based their quoted estimates on only a subset of the examination methodology. It is common practice within the latent fingerprint community to ensure conclusions have been verified by a separate examiner prior to a conclusion being released. While the OSAC FRS recognizes that many laboratories may not perform ‘blind’ verifications, the error rates quoted by the PCAST did not consider any verification being performed. Accordingly, the error rates quoted by the PCAST do not necessarily reflect actual casework methodology.”

Additional Information & Clarification of Statements

Several statements in the article mention information that was unknown about the MDPD study. Additional information is provided below to assist with any research efforts in improving error rate estimations.

- **“Control prints from the true sources were available for 56 out of the 80 test latent prints, and were absent for the remaining 24. The repartition of these 56 and 24 latent prints across the two subsets of 40 is unknown.**”
    - The distribution of source present and source not present trials across the two phases is as follows:
        - <u>Phase 1</u>
          Source Present Trials - 28
          Source Not Present Trials - 12
        - <u>Phase 2a</u>
          Source Present Trials - 14
          Source Not Present Trials - 6
        - <u>Phase 2b</u>
          Source Present Trials - 14
          Source Not Present Trials - 6
- **“Finally, in Phase 3, participants were presented with sets of prints that consisted of the ‘identifications’ made in Phase 2 (regardless of whether the ‘identification’ was correct).”**
    - In addition to “identifications”, participants were also presented with exclusions and inconclusive results under bias and repeatability conditions in Phase 3 of the MDPD study.
- **“Despite these issues and that the MDPD study was never published in a peer-reviewed journal, the PCAST report relies heavily on these results to discuss the estimated error rate associated with latent fingerprint examination [4].”**
    - Although the MDPD study has not yet been published in a peer-reviewed journal, the report was peer reviewed by two anonymous sources designated by the

National Institute of Justice (NIJ) and edited accordingly prior to the final report being accepted by the NIJ in 2014. Additionally, we are currently drafting a report of the MDPD study for publication in a peer-reviewed journal that will include calculations of the ACE false positive rate.

- **"However, only 4,233 decisions were returned at the end of phase 1 (out of a total of 4,360 possible decisions), and 1,730 decisions were returned at the end of Phase 2 (out of 1,760 possible decisions). Given that some latent prints were selected to be more challenging than others and given that some examiners may have been busier than others, it is likely that the missing responses are concentrated on a small number of latent prints or a small number of examiners. However, the MDPD report does not provide information on which examiners did not finish the study and on which latent prints were favored."**
- **"It is likely that challenging prints were more often deemed of no value; furthermore, it is also likely that some participants were more prone to deem that test cases should not be compared to control prints and that significant variability exists between the value determination of some latent prints by the participants."**
    - To determine the number of missing responses, a distinction must be made between responses missing for sufficiency decisions vs responses missing for conclusions (identification, exclusions, or inconclusive). Furthermore, these two responses do not necessarily equate since some participants may have reported a sufficiency decision, but did not report a conclusion, some participants reported no sufficiency decision or a sufficiency decision of "no value," but still reported a conclusion, and some (if not most) participants only reported a conclusion for sufficiency decisions of "value".
    - Phase 1:
        - 76.15% of the 109 participants provided a sufficiency response (value or no value) for all 40 latent prints presented to them
        - 97.09% (4,233) of the total possible number of sufficiency responses (4,360) were reported by 109 participants
            - 75.83% Value
                - Insufficient to Difficult (3.40%)

      - Difficult to Moderate (40.37%)
      - Moderate to Easy (56.23%)
    - 24.17% No Value
      - Insufficient to Difficult (82.11%)
      - Difficult to Moderate (17.11%)
      - Moderate to Easy (.78%)
  - 72.85% (3,176) of the total possible number of conclusion decisions (4,360) were reported by 109 participants
    - There were 9 latents that were rated Insufficient to Difficult (6 Source Present and 3 Source Not Present).
      - Source present: 73 of 654 possible conclusions reported
      - Source not present: 72 of 327 possible conclusions reported
      - 3.33% total conclusions reported
    - There were 14 latents that were rated Difficult to Moderate (9 Source Present and 5 Source Not Present).
      - Source present: 800 of 981 possible conclusions reported
      - Source not present: 476 of 545 possible conclusions reported
      - 29.27% total conclusions reported
    - There were 17 latents that were rated Moderate to Easy (13 Source Present and 4 Source Not Present).
      - Source present: 1,358 of 1,417 possible conclusions reported
      - Source not present: 397 of 436 possible conclusions reported
      - 40.25% total conclusions reported
- Phase 2a
  - 80.43% of the 46 participants provided a sufficiency response (value or no value) for all 20 latent prints presented to them

- 97.93% (901) of the total possible number of sufficiency responses (920) were reported by 46 participants
    - 73.92% Value
        - Insufficient to Difficult (9.76%)
        - Difficult to Moderate (42.19%)
        - Moderate to Easy (48.05%)
    - 26.08% No Value
        - Insufficient to Difficult (86.81%)
        - Difficult to Moderate (13.19%)
        - Moderate to Easy (0%)
- 74.35% (684) of the total possible number of conclusion decisions (920) were reported by 46 participants
    - There were 6 latents that were rated Insufficient to Difficult (4 Source Present and 2 Source Not Present).
        - Source Present: 39 of 184 possible conclusions reported
        - Source Not Present: 41 of 92 possible conclusions reported
        - 8.70% total conclusions reported
    - There were 7 latents that were rated Difficult to Moderate (5 Source Present and 2 Source Not Present).
        - Source Present: 209 of 230 possible conclusions reported
        - Source Not Present: 75 of 92 possible conclusions reported
        - 30.87% total conclusions reported
    - There were 7 latents that were rated Moderate to Easy (5 Source Present and 2 Source Not Present).
        - Source present: 228 of 230 possible conclusions reported
        - Source not present: 92 of 92 possible conclusions reported

            - 34.78% total conclusions reported

- Phase 2b
    - 92.86% of the 42 participants provided a sufficiency response (value or no value) for all 20 latent prints presented to them
    - 98.69% (829) of the total possible number of sufficiency responses (840) were reported by 42 participants
        - 81.54% Value
            - Insufficient to Difficult (2.96%)
            - Difficult to Moderate (42.31%)
            - Moderate to Easy (54.73%)
        - 18.46% No Value
            - Insufficient to Difficult (68.63%)
            - Difficult to Moderate (29.41%)
            - Moderate to Easy (1.96%)
    - 80.48% (676) of the total possible number of conclusion decisions (840) were reported by 42 participants
        - There were 3 latents that were rated Insufficient to Difficult (2 Source Present and 1 Source Not Present).
            - Source Present: 9 of 84 possible conclusions reported
            - Source Not Present: 16 of 42 possible conclusions reported
            - 2.98% total conclusions reported
        - There were 8 latents that were rated Difficult to Moderate (6 Source Present and 2 Source Not Present).
            - Source Present: 217 of 252 possible conclusions reported
            - Source Not Present: 69 of 84 possible conclusions reported
            - 34.04% total conclusions reported

        - There were 9 latents that were rated Moderate to Easy (6 Source Present and 3 Source Not Present).
            - Source present: 244 of 252 possible conclusions reported
            - Source not present: 121 of 126 possible conclusions reported
            - 43.45% total conclusions reported

- **"It is likely that some of the more challenging test latent prints resulted in more inconclusive examinations than others. It is also possible that a large number of erroneous conclusions can be associated with a small number of participants."**
    - During Phase 1, Phase 2a, and Phase 2b of the MDPD study, there was a significantly higher rate of inconclusive decisions in each of the three phases for latents that were rated insufficient to difficult, regardless of whether the source was or was not present. There were also more inconclusive decisions reported in trials where the source was not present versus source present trials overall, regardless of latent difficulty. However the difference in inconclusives for source present vs source not present trials is less pronounced in latents that were rated insufficient to difficult. Latents in this category had similar rates of inconclusive decisions (in Phase 2b the rates were the same) regardless if the source was or was not present.
    - For false positive conclusions, there were 28 different participants that reported a total of 42 incorrect identifications.
    - For false negative conclusions, there were 77 different participants that reported a total of 235 incorrect exclusions.

- **"The use of a single FPR relies on the assumption that every latent print has the same chance to result in an erroneous identification[2]." Footnote [2]– "This assumption is not reasonable: lower quality latent prints have been shown to result in higher error rates [7]. However, the key question raised by the MDPD, as we will discuss below, is whether the rate of erroneous identifications is the same when control impressions are provided as when they are not.")**
    - We support the notion of calculating different FPRs versus a single FPR for latent print comparisons. In

reporting the initial results of the MDPD study, the FPR for ACE was not a central focus of the analysis or discussion. Rather, the focus of the study was on the determination of the FPR for ACE-V since it more closely resembles real casework. The impetus driving the design of the MDPD study was the evaluation of the work product that is released after applying ACE-V methodology. It is evident that further analysis and discussion of estimating error rates under various conditions is warranted, and it is important that questions regarding error rates be properly qualified.

  - The number of False Positive errors in the MDPD study is as follows:
    - Latents rated Insufficient to Difficult
      - Source Present - 0 errors
      - Source Not Present - 0 errors
    - Latents rated Difficult to Moderate
      - Source Present - 19 errors
      - Source Not Present - 0 errors
    - Latents rated Moderate to Easy
      - Source Present - 20 errors
      - Source Not Present - 3 errors
  - The number of False Negative errors in the MDPD study is as follows:
    - Latents rated Insufficient to Difficult
      - Source Present - 5 errors
      - Source Not Present - N/A
    - Latents rated Difficult to Moderate
      - Source Present - 151 errors
      - Source Not Present - N/A
    - Latents rated Moderate to Easy
      - Source Present - 79 errors
      - Source Not Present - N/A

- **"The implications of these results are two-fold: first, it is clear that 3.0% is an over-estimation of the true FPR of the participants of the MDPD study; second, examiners behave differently when impressions from**

the true sources are present in the comparison process. When control impressions are present, examiners make an inconclusive decision approximately 14% of the time, compared to approximately 30% of the time when control impressions are not present."

- With respect to reporting inconclusive decisions, the MDPD research data supports the notion that examiners behave differently when the source is present versus when the source is not present.
- Phase 1
  - Latents rated Insufficient to Difficult (9 of 40)
    - Source Present Trials (6 of 9 latents)
      - 45 of 73 conclusion decisions were inconclusive (61.64%)
    - Source Not Present Trials (3 of 9)
      - 47 of 72 conclusion decisions were inconclusive (65.28%)
  - Latents rated Difficult to Moderate (14 of 40)
    - Source Present Trials (9 of 14)
      - 184 of 800 conclusion decisions were inconclusive (23%)
    - Source Not Present Trials (5 of 14)
      - 166 of 476 conclusion decisions were inconclusive (34.88%)
  - Latents rated Moderate to Easy (17 of 40)
    - Source Present Trials (13 of 17)
      - 111 of 1,358 conclusion decisions were inconclusive (8.17%)
    - Source Not Present Trials (4 of 17)
      - 79 of 397 conclusion decisions were inconclusive (19.90%)
- Phase 2a
  - Latents rated Insufficient to Difficult (6 of 20)
    - Source Present Trials (4 of 6 latents)
      - 22 of 39 conclusion decisions were inconclusive (56.41%)

  - Source Not Present Trials (2 of 6)
    - 25 of 41 conclusion decisions were inconclusive (60.98%)
- Latents rated Difficult to Moderate (7 of 20)
  - Source Present Trials (5 of 7)
    - 27 of 209 conclusion decisions were inconclusive (12.92%)
  - Source Not Present Trials (2 of 7)
    - 21 of 75 conclusion decisions were inconclusive (28.0%)
- Latents rated Moderate to Easy (7 of 20)
  - Source Present Trials (5 of 7)
    - 7 of 228 conclusion decisions were inconclusive (3.07%)
  - Source Not Present Trials (2 of 7)
    - 4 of 92 conclusion decisions were inconclusive (4.35%)

- Phase 2b
  - Latents rated Insufficient to Difficult (3 of 20)
    - Source Present Trials (2 of 3 latents)
      - 9 of 9 conclusion decisions were inconclusive (100%)
    - Source Not Present Trials (1 of 3)
      - 16 of 16 conclusion decisions were inconclusive (100%)
  - Latents rated Difficult to Moderate (8 of 20)
    - Source Present Trials (6 of 8)
      - 35 of 217 conclusion decisions were inconclusive (16.13%)
    - Source Not Present Trials (2 of 8)
      - 31 of 69 conclusion decisions were inconclusive (44.93%)
  - Latents rated Moderate to Easy (9 of 20)
    - Source Present Trials (6 of 9)

        - 6 of 244 conclusion decisions were inconclusive (2.46%)
    - Source Not Present Trials (3 of 9)
        - 14 of 121 conclusion decisions were inconclusive (11.57%)

General Comments

The implications raised in the article are intriguing. In addition, the proposed algorithm used to simulate the error rates in the MDPD study and to test the appropriateness of the different proposed FPR estimates is of great interest to us. While we would need more time to evaluate the validity of the algorithm used to generate the results of the simulations and alternative FPR estimates in the article, we strongly support developing models to assist with estimating latent fingerprint comparison error rates.

Letter from Jonathan J. Koehler
Beatrice Kuhn Professor of Law
Northwestern Pritzker School of Law
Chicago, IL

Ausdemore, Hendricks, & Neumann (in press) do a wonderful job of digging into the statistical details of the famously unpublished Miami Dade fingerprint study (Pacheco, Cerchiai, & Stoiloff, 2014). Among other things, they explain that the 3.0% false positive error rate (FPR) for fingerprint examiners that this study reports may be misleading when considered in isolation. The 3.0% FPR arises from the fact that the study reports that examiners committed 42 false positive errors out 1,398 pairwise conclusions from samples that were produced by different sources. Ausdemore et al. note that these data could be parsed in various ways. For example, if 403 "inconclusives" are subtracted from that 1,398 figure, then the FPR is 4.2%. If the 35 false positives that occurred when examiners identified the wrong finger of the correct source are not scored as errors, then the FPR is 0.7% (7 out of 995) or 0.5% (7 out of 1,398, if the inconclusives are retained).

Ausdemore et al. argue that the 35 right-person-wrong-finger errors in the Miami Dade study should not be lumped with the 7 wrong-person errors. I agree (Koehler, 2017, Fauxrensics, p. 1412-1413). The point could be made more forcefully by suggesting that right-person-wrong-finger conclusions should not be scored as errors at all because the hypothesis of interest will almost never specify that a particular *finger* is the source of a latent print. Instead, the hypothesis of interest will usually be whether a particular *person* is the source of a latent print. If so, then right-person-wrong-finger "errors" should be treated as true positives rather than as false positives. It does seem, then, that the 3.0% FPR that the authors of the Miami Dade study report overstates the FPR as Ausdemore et al. claim.

However, it would be a short-sighted to argue that any one value represents "the" error rate of interest. As Ausdemore et al. show, the FPR in the Miami Dade study appears to depend on whether control prints from the true source of the latent prints were or were not provided to the examiners. Surprisingly, examiners committed false positive errors at a higher rate when provided with the true source prints than when they were not provided with those source prints. This FPR difference arises from the different rates at which the examiners rely on "inconclusive" determinations in the two situations.

A broader issue that Ausdemore et al. leave largely untouched is what role FPRs should play in assessing the probative value of a reported latent print identification. I have frequently suggested that false positive error rates should be estimated by methodologically rigorous proficiency tests and presented to triers of fact (Koehler, 2013, 2017, 2018; for a contrary view, see Budowle, Bottrell, Bunch, et al., 2009). The FPR places an upper bound on the probative value of a reported match in highly discriminating forensic fields such as fingerprints and DNA (Koehler, Chia, & Lindsey, 1995; Thompson, Taroni, & Aitken, 2003). In my view, this error rate should be estimated empirically under various relevant conditions and then communicated carefully to legal actors to help them understand it.

Courts have been slow to embrace the notion that error rates matter in the context of forensic science testimony. They have not required forensic scientists who offer identity conclusions (often with 100% certainty) to produce reliable data that show how accurate their conclusions are. This is a serious problem. Unless and until courts require such data, triers of fact will not get the information they need to evaluate the reliability of source conclusions.

Will the courts come around? Perhaps, but much of the forensic science and justice communities resist the common-sense notion that error rates are important to identify and reveal. Note that I am not referring to the importance of revealing errors to improve an examiner's performance, or even to improve the forensic sciences more broadly (as admirable as those purposes might be). The purpose that I have in mind for error rates is as indicators of the probative value of a reported match.

Unfortunately, the lines are drawn rather starkly between those who think error rates are a crucial part of the reliability picture, and those who think error rates are irrelevant and misleading. This is a problem for achieving consensus down the road because beliefs about error rates are a lot like beliefs about God, abortion, and Donald Trump: no amount of argument from someone who has an opposing view is likely to have a measurable impact. Without buy-in from the science and justice communities that error rates matter and should be estimated, courts are unlikely to require such data. And, as noted above, without error rate data, jurors won't have a sufficient basis for evaluating the reliability of match reports.

The Ausdemore et al. review of the Miami Dade study shows that error rate estimations and computations are not always straightforward and obvious. Complicating matters, when large fingerprint databases are searched and highly similar non-source exemplars are located, it seems likely that the FPR will be affected. Such highly similar-looking print pairs are referred to as “close non-matches” (CNMs) and a few investigators have started to take a look at the problems they pose. Indeed, data from a technical report co-authored by one of the Ausdemore et al. authors (Neumann) provide some reason to suspect that the FPRs that stem from CNM comparisons may be substantially higher than the percentages identified in the Miami Dade study (Neumann, Champod, Yoo, et al., 2013, p. 56).

Whatever one may think of the Miami Dade study and the critical review of it that Ausdemore et al. provide, it would be wonderful if these efforts inspired other scientists to design and conduct error rate studies that will provide a scientific basis for the general public to assign the appropriate weight to forensic reports in various disciplines.

References

Ausdemore, M. A.; Hendricks, J. H.; Neumann, C. Review of Several False Positive Error Rate Estimates for Latent Fingerprint Examination Proposed Based on the 2014 Miami Dade Police Department Study. *J. For. Ident.* **2019**, *69* (1), 59–81.

Budowle, B.; Bottrell, M. C.; Bunch, S. G.; Fram, R.; Harrison, D.; Meagher, S.; Oien, C. T.; Peterson, P. E.; Seiger, D. P.; Smith, M. B.; Smrz, M. A. A Perspective on Errors, Bias, and Interpretation in the Forensic Sciences and Direction for Continuing Advancement. *J. For. Sci*. **2009**, *54* (4), pp 798–809.

Koehler, J. J. Proficiency Tests to Estimate Error Rates in the Forensic Sciences. *Law, Probability & Risk*, **2013**, *12*, 89–98.

Koehler, J. J. Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences. *AZ St. Law J.* **2017**, *49*, 1369–1416.

Koehler, J. J. How Trial Judges Should Think about Forensic Science Evidence. *Judicature* **2018**, *102*, 28–38.

Koehler, J. J.; Chia, A.; Lindsey, J. S. The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial? *Jurimetrics J.* **1995**, *35*, 201–219.

Neumann, C.; Champod, C.; Yoo, M.; Genessay, T.; Langenburg, G. Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination. National Institute of Justice Report (2010-DN-BX-K267), Washington, DC, 2013.

Pacheco, I.; Cerchiai, B.; Stoiloff, S. *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations*. NCJRS document No. 248534, 2014.

Thompson, W. C.; Taroni, F.; Aitken, C. G. G. How the Probability of a False Positive Affects the Value of DNA Evidence. *J. For. Sci*. **2003**, *48* (91), 47–54.

Letter from Carey Hall
Saint Paul, MN

I am grateful for this special feature regarding the topic of error rates. Since 2011, the research on examiner performance has generated data that can be used for the purpose of general error rates. The following commentary is broken into two components: focus on exclusion decisions and specifically considerations of the specificity calculation, followed by general comments on error rates for use in court.

When research on the accuracy and reliability of the latent print discipline emerged in 2011 it became apparent that the biggest source of error was the exclusion decision [1, 2]. In every study, under various conditions, exclusions had a higher error rate, by as much as 70 times [3, 4].

If true improvement of the science and reliability of the examination is the goal, the intense focus solely on the identification decision is misplaced. In 2016, the President's Council of Advisors on Science and Technology (PCAST) published a review of feature matching disciplines, including latent print examination. Careful review of accuracy of non-association examinations (e.g. exclusions) and testimony fit squarely within PCAST's scope "what additional efforts could contribute to strengthening the forensic-science disciplines and ensuring the scientific reliability of forensic evidence used in the Nation's legal system" [5].

Unfortunately, PCAST replicated the mistakes the latent print community had made for decades, overlooking the importance of the exclusion decision and focusing exclusively on the identification decision. For a period, latent print offices were called "identification units". It was acceptable to fail to make or report affirmative exclusion decisions, instead rendering "failed to identify" conclusions [6]. This history lead not only to a perception of a close police laboratory relationship, but also to a lack of scientific rigor.

A quick review of erroneous conviction work would demonstrate that, at least in latent prints, 'non-association' testimony is rather common. In one published study of erroneous convictions 11 of the 13 latent print cases had some sort of non-associative results [7]. Although exclusion decisions also have use prior to judicial proceedings as well. Exclusion decisions in investigative and even in pre-trial proceedings can have a large impact, although not always. The importance of exclusion decisions is difficult if not impossible to determine without the context of the investigation. Without context some examiners prefer to limit the value of exclusion decisions with idioms like “absence of evidence is not evidence of absence”. However, an exclusion decision on a critical piece of evidence, or in cases with multiple defendants can have a large impact on courses of investigation or even trial results (e.g. Kenneth Waters, or Gene Bibbins) [8, 9].

We don’t have published non-associative testimony rates and we have limited data on conclusion rates within latent print agencies. From personal experience, tracking internal performance metrics for the Foresight project [10], exclusion decisions appear to rendered frequently, and occur at higher rates than identification decisions. While exclusion rates seem uninteresting and have not been published, publishing them could aid in the discussion of contextual and confirmation bias. High or low exclusion rates may add something meaningful about examiner’s priors, especially for suggested suspects. This is one of many reasons I advocate that the exclusion decision should be tracked and published.

The published examiner performance studies and subsequent reviews focus predominantly on the false positive rate and less so the false negative rate. A notable recent exception, explores and categorizes the vast number of “errors” examiners may make throughout the course of their examination [11]. Despite the error rate data generated by these performance studies PCAST and the American Academy for the Advancement of Science, (AAAS) Advisory Group, organizations which have the caliber to understand diagnostic trade-offs between sensitivity and specificity, focused intently on the identification decision, and minimally in the case of AAAS, the exclusion decision [5, 12].

Nonetheless, the importance of the exclusion decision should not be minimized because two reviewer groups failed to analyze the significance within forensic work.

Indeed, even in 2011 some practitioners in the latent print community quickly implemented policy changes to reduce false negative errors revealed as a result of examiner performance studies [6]. Implementing policy changes to address a problem identified by empirical research is a good use of scientific research. However, empirical research should demonstrate which solutions are best to address the problems uncovered by research. Otherwise, the changes may not address the problem and may make new problems [13]. Additionally, changes in policy always come with a cost; it is difficult to impossible to demonstrate costs or benefits without prior metrics.

Again, individual agencies should record performance metrics, including conclusion rates. As policies change having historical or baseline data to better understand how policy decisions impact reported results and will better assess the efficacy.

Many of the policy changes to reduce false negative errors were implemented quickly with little consideration for the importance of the exclusion decision, nor regard for how the changes would impact rates of reported exclusions.

While PCAST focused on false positive, mirroring this approach to simply consider the false negative rate fails to fully capture diagnosticity. The specificity calculation is also needed [14]. Specificity is defined as a measure of the proportion of true negatives that are correctly identified as such. In simple terms, this is a test's ability to correctly exclude. It is a component of diagnosticity. It can be calculated one of two ways:

First:

$$\frac{\text{Correct Exclusions}}{\text{Different Source Trials}}$$

OR

Second:

$$\frac{\text{Correct Exclusions}}{\text{False Positives and Correct Exclusions}}$$

In signal detection theory, where specificity calculations originated, decisions are (generally) binary and complementary. If something is not identified, it is excluded and vise versa. In commonly applied contexts, like medical diagnosis, a test report would be positive or negative for a condition.

In a latent print examination, there are two loop-holes to the binary decision. The first, is to determine the sample is insufficient to proceed to any further phase of examination, commonly called "no value". The second loop-hole is to render an inconclusive decision at the end of a comparison. A practitioner may render an inclusive decision in relation to perceived negative aspects of the exemplar or the latent print [15].

In medical contexts, it is unusual to allow for the inconclusive or indeterminate category, but not non-existent [16].

Deciding how to handle the no value and inconclusive decisions can provide different rates of specificity [17]. The important difference between these two ways to calculate specificity is the first will include all impressions regardless if they are determined to be no value. The second will only include attempts to compare - - no value determinations will not be considered.

To graphically illustrate this complexity, a hypothetical study is presented below. The usefulness of considering the specificity calculation, instead of only the false negative rate, becomes apparent.

Imagine a study where one hundred identical latent print images are provided to three different laboratories. Of the hundred latent prints, fifty are non-mated (from different sources); the remaining fifty are mated (same source). Assume that they are 100% accurate in the identification and exclusion decisions reported. In the table below are the hypothetical break-downs of each laboratory's performance for the exclusion decision.

	Laboratory 1		**Laboratory 2**		**Laboratory 3**	
	100 latent prints		100 latent prints		100 latent prints	
Pairs of Images	50 mated	50 non-mated	50 mated	50 non-mated	50 mated	50 non-mated
# of Exclusions	50 exclusion decisions		0 exclusion decisions		25 exclusion decisions	
Specificity of Calculation	50/50 = 100%		0/50=%		25/50%	

Table 1

Results from hypothetical laboratory performance test.

Although, all three of these laboratories are 100% accurate in their exclusion decisions, 0% false negative error rate, they have very different rates of specificity. Ideally, all labs would be like Laboratory 1, taking every opportunity to exclude. Laboratory 2 is entirely ineffective, this represents the laboratories that do not report exclusions or whose analysts never exclude. Laboratory 3 only excludes half of the latent prints that could have been excluded.

To illustrate complexity in the choice of the specificity calculation compare Laboratory 1 and 3 under the second formulation. Imagine Laboratory 3 has a policy that restricts value determination or does not allow exclusion decisions. In the second formulation to calculate specificity, number of false positives + number of correct exclusions, clearly benefits Laboratory 3. The specificity under this formulation for Laboratory 3 would be: 25/0 + 25 = 100%. If Laboratory 3 determined the other 25 non-mated pairs were not suitable, thus never compared them, they would not count against them in the second formulation.

Comparing the specificity of Laboratory 1 to 3; using the second formulation seems unfair to Laboratory 1, which correctly excluded all samples at every opportunity.

Unfortunately, this simplistic hypothetical does not represent the nuances of the determination of value. Latent prints are partial representations of the friction ridge skin; they are fragmentary, incomplete and can be distorted. There are times when it is not appropriate to render any decision on insufficient samples. This concept of insufficiency of a sample also exists in the medical context, but is not accounted for in the traditional sensitivity and specificity calculation. For instance, if there is not enough of a sample (e.g. blood) to perform a test the sample is rejected (i.e. no value) [18]. But a clinical scientist might ask for a new or better sample for testing. In latent print examination it is not possible to obtain a new or better sample than the original transfer.

The challenge in dealing with inconclusive decisions is not unique to forensic science nor to the specificity calculation [19]. It is present in the false positive and the false negative rates. PCAST addresses this dilemma summarily with a broad brush "SEN and FPR can thus be calculated based on the *conclusive* examinations or on *all* examinations. While both rates are of interest, from a scientific standpoint, the former rate should be used for reporting FPR to a jury. This is appropriate because evidence used against a defendant will *typically* be based on *conclusive*, rather than inconclusive, examinations" [5].

As mentioned above we currently lack metrics about what testimony is "typical" for latent print examiners. Association testimony certainly carries the most weight against a defendant and seemed to unilaterally be the focus of PCAST, regardless if it was the best *scientific* policy inquiry.

The PCAST recommendation to exclude inconclusive from error rate calculations is arbitrary and may be a construct of their narrow mission to evaluate the intersection of forensic science in the courtroom. Good quality assurance policy would recognize departures from normal performance; which would include no value, and inconclusive rates, not just for the courts, but for all users of the criminal justice system. My recommendation is laboratories should choose one specificity calculation to track, but if it doesn't include no value and inconclusive decisions those rates should be tracked separately. Because casework is not ground truth, special consideration of these rates should be made during competency and proficiency testing when ground truth is known.

Comments Regarding the Presentation of Error Rates During Trial

When presenting error rates in court, I am in favor of leaving the inconclusive in the false positive and false negative rates, *but* also presenting the sensitivity, specificity calculations with them left-in as well. Leaving them in for false positive and false negative rates, but removing them for sensitivity and specificity calculations would be like having cake and eating it too. Leaving inconclusive decisions in lowers the false positive and false negative error rates, but also the sensitivity and specificity as shown in Table 2.

	False Positive	**Sensitivity**	**False Negative**	**Specificity**
Ulery et., 2011 without Inconclusive	0.17%	89%	11%	99%
with Inconclusive	0.15%	61%	7.5%	88%
Pacheco et al, 2014 without Inconclusive	4.2%	91%	8.7%	96%
with Inconclusive	3%	78%	7.4%	68%

*ACE, Phase 1 including clerical errors.

Table 2

Comparison of False Positive Error, Sensitivity, False Negative Error, Specificity rates with and without the inclusion of inconclusive decisions.

PCAST offers false positive error rates that examiners can use while providing testimony, but some practitioners may feel it is better to avoid the whole topic altogether, recognizing it to be a calculated risk.

First, there is not a natural place to insert the error rate statement. It seems unnatural and forced to answer “did you reach a conclusion on this exhibit?” With “Yes, identification. But research studies have shown a false positive rate as reasonably high as 1 in 18.” The witness has offered more than was asked and brought in research to a narrowly focused question on *this* exhibit. The oddity of this response would not be lost on jurors. There may be a more appropriate place to bring in error rates elsewhere.

A prosecutor is unlikely to ask about error rates so unless the examiner inserts error rate testimony in response to another question asked, it is unlikely to come up on direct examination. Defense counsel *may not* ask either. If defense does ask, examiners recognize their credibility may be damaged if they cannot answer. But it is easier to wait to address something that might hurt their evidence rather than present it themselves. The fear is the jury may negatively view or completely disregard the latent print evidence when presented with such a high error rate. The wait-and-see approach is not ideal and certainly doesn't meet PCAST's goal to "strengthen" the value of forensic science.

Next, because practitioners may be unfamiliar with the technical aspects of the research or calculations, they may end up hurting their own credibility by not having enough technical knowledge to respond to difficult questions. The thought is: do not wade into the water if you are not sure if you can swim.

Lastly, the trier-of-fact is often perceived as not interested in these error rates. It has been my experience teaching practitioners, meeting with attorneys, and preparing for admissibility hearings the nuances of statistical calculations are of little interest. Any hint of complex issues such as; confidence intervals, inconclusive decisions, no value decisions, or study design validity means a fast track to la-la-land.

I don't mean to suggest error rates should not be reported, nor that they are not important. But one important question moving forward will be how to best relay error rates to lay people. Will they distinguish or conflate false positive error rates in research with the case at hand? How does the presented error rate apply to this case? PCAST contemplates this in reliability as applied. Psychology has a rich history of exploring these risk communication issues from which we can borrow [20]. But both PCAST and AAAS advocate further exploration of domain relevant error communication. In the meantime, further exploration of examiner performance with white box approaches will bring more attention to the contexts under which errors become more likely.

Black box studies inherent design makes it difficult to determine the cause of the errors. So while black box studies were an excellent starting place for baseline performance, community-wide white box studies and internal performance studies allow errors to be deconstructed. The conditions and contexts under which errors are made are critical for the trier-of-fact to properly integrate the probative value of the evidence with the risk of error.

References

1. Ulery, B. T.; Hicklin, R. A.; Buscaglia, J.; Roberts, M. A. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. *Proc. Nat. Acad. Sci.* **2011**, *108* (19), 7733–7738.
2. Langenburg, G.; Champod, C.; Genessay, T. Informing the Judgments of Fingerprint Analysts Using Quality Metric and Statistical Assessment Tools. *For. Sci. Int.* **2012**, *219* (1–3), 183–198.
3. Langenburg, G. A Performance Study of the ACE-V Process: A Pilot Study to Measure the Accuracy, Precision, Reproducibility, Repeatability, and Biasability of Conclusions Resulting from the ACE-V Process. *J. For. Ident.* **2009**, *59* (2), 219–257.
4. Pacheco, I.; Cerchiai, B.; Stoiloff, S. *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations*. NCJRS document No. 248534, 2014.
5. President's Council of Advisors on Science and Technology (PCAST). *Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*; Executive Office of the President's Council of Advisors on Science and Technology: Washington, D.C., 2016.
6. Ray, E.; Dechant, P. J. Sufficiency and Standards for Exclusion Decisions. *J. For. Ident.* **2013**, *63* (6), 675–697.
7. Garrett B. L.; Neufeld P. J. Invalid Forensic Science Testimony and Wrongful Convictions, *VA. L. Rev.* **2009**, *95* (1), 1-97.
8. The Innocence Project. DNA Exonerations in the United States. Gene Bibbins https://www.innocenceproject.org/cases/gene-bibbins/ (accessed December 2018).
9. The Innocence Project. DNA Exonerations in the United States. Kenneth Waters https://www.innocenceproject.org/cases/kenny-waters/ (accessed December 2018).
10. Houck, M. M.; Riley R. A.; Speaker P. J.; Witt T. S. FORESIGHT: A Business Approach to Improving Forensic Science Services. *For. Sci. Policy & Management: An Int. J.* **2009**, *1* (2), 85–95.
11. Cole, S.; Scheck B. Fingerprints and Miscarriages of Justice: 'Other' Types of Errors and A Post-Conviction Right to Database Searching. *Alb. L. Rev.* **2017**, *81* (3) 807– 850.
12. Thompson, W.; Black, J.; Jain, A.; Kadane, J. *Forensic Science Assessments: A Quality and Gap Analysis Latent Fingerprint Examination*. AAAS: Washington, D.C. Sept. 2017. DOI: 10.1126/srhrl.aag2874.

13. Behn, R. D. Why Measure Performance? Different Purposes Require Different Measures. *Pub. Admin. Rev.* **2003**, *63* (5) 586–606.
14. Yerushalmy, J. Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-ray Techniques. *Pub. Health Rep. (1896-1970)*, **1947**, *62* (40), 1432–1449.
15. SWGFAST. Document #10 Standard for Examining Friction Ridge Impressions and Resulting Conclusions (ver. 2.0). 03/13/13.
16. Feinstein, A. R. The Inadequacy of Binary Models for the Clinical Reality of Three-zone Diagnostic Decisions. *J. Clin. Epidemiol*. **1990**, *43* (1), 109–113.
17. Simel, D. L.; Matchar, D. B.; Feussner J. R. Diagnostic Tests Are Not Always Black or White: Or, All That Glitters Is Not [A] Gold [Standard]. *J. Clin. Epidemiol*. **1991**, *44* (9), 967–970.
18. Tube Fill Requirements and Common Causes of Specimen Rejection. https://www.gundersenhealth.org/app/files/public/150/Laboratory-Tube-fill-requirements.pdf (accessed December 2018).
19. Phillips V. L.; Saks, M. J.; Peterson, J. L. The Application of Signal Detection Theory to Decision- Making in Forensic Science. *J. For. Sci*. **2001**, *46* (294), 294–308.
20. Edwards, A.; Elwyn, G.; Mulley, A. Explaining Risks: Turning Numerical Data into Meaningful Pictures. *BMJ* (Clinical research ed.). **2002**, *324* (7341), 827– 830.

Letter from Brendan Max, J.D.
Cook County Public Defender Office
Cook Co., Illinois

First of all, I applaud the authors, the more attention and analysis that fingerprint error rate studies generate, the better for the discipline. I am not sufficiently familiar with simulations of the type reported here to comment on the appropriateness of the simulation, including the assumptions reported by the authors on page 11. Therefore, I cannot comment on the authors' core conclusion that *"it is clear that 3% is an over-estimation of the true FPR of the participants of the MDPD study."* Nonetheless, I offer the following comments, which touch on some claims made by the authors of the study as well as address the problems as I see them with assessing FPRs from currently-available error rate studies.

- The authors refer to the difference in FPR between the Miami Study (3%) and the Ulery Study (.1%) as a "discrepancy" that has caused a great deal of controversy in the field. I have seen no meaningful comparison of the two studies that leads me to believe that the results of the two studies are inconsistent in a way that should be characterized as a "discrepancy." First, the design of the Miami Study was more challenging than the design of the Ulery Study- participants in the Miami Study had to search across 3 fingerprint cards (30 fingers, plus 6 palms) for a potential match whereas participants in the Ulery Study had the more easy task of a 1-on-1 comparison. Second, I have seen no meaningful comparison of the complexity of the latent prints in the two studies, so we have no way to determine if, in addition to a more challenging methodology, the Miami Study involved more challenging latent prints. The 1995 CTS fingerprint proficiency test resulted in very poor performance by participants while the 2018 version resulted in near-universal passage. One would not consider the difference in passage rates a discrepancy, because the 2018 version was designed to be really easy and the 1995 version was designed to be quite challenging. Until some meaningful comparison of the Miami Study and the Ulery Study is done, I would hesitate to characterize the FPRs from the two studies as a discrepancy.
- Another reason that the results from the two studies should not be seen as irreconcilable is the effect of "inconclusive" determinations by participants in the two studies. In instances where a participant offered an inconclusive

decision in either study, there was no final decision by the participant whether the latent prints in question matched the provided suspect finger(s). Expressed in other terms, each of these cases was a potential false positive averted by shutting the analysis down. In the Ulery Study, participants took a pass and reported their analysis as "inconclusive" at about twice the rate that participants did in the Miami Study. In the Ulery Study, participants reported inconclusive decisions in 37% of comparisons of prints of value (4907 inconclusive decisions in 13,174 comparisons), while participants reported inconclusive decisions in only 19% of comparisons of value in the Miami Study (849 inconclusive decisions in 4536 comparisons). Participants in the Ulery Study were likely more cautious than participants in the Miami Study. This alone could account for the "discrepancy" in the FPR for the two studies.

- The number of inconclusive decisions in both studies (as well as the rates at which examiners claimed that latent prints in the two studies were of no value- about 23% in both studies) raises another unaddressed question when trying to assess the FPRs. The important question for the discipline is whether the FPR in the Miami-Dade Study or any similar study is a good reflection of the FPR in casework. If participants in these studies are taking a pass (by making very cautious value/no value judgments and/or by overuse of the "inconclusive" determination) in a way that does not reflect their approach to similar determinations in casework, the FPRs from both studies may dramatically underestimate the real FPR in casework. For example, why were some participants in the Miami Study claiming that prints were of "no value" for prints that were pre-determined to be easy (8 no value decisions in 1813 comparisons- .4%) or moderate (175 no value determinations in 1471 comparisons- 12%)? When debating whether 3% or 1% is a more reliable indication of the FPR in the discipline, there should be some acknowledgement that both studies may underestimate the real FPR in casework due to the overuse of "no value" and "inconclusive" decisions by participants.
- The authors state that false positives should be treated differently when the analyst got the wrong finger but the right person. It seems to me that we don't really know much about why examiners in the Miami Study selected

the right person but the wrong finger at times. Were these clericaI errors? Were examiners being fooled by close non-match fingers from the same hand/person? While there may be justification for treating this type of error differently when we know for sure that the errors are merely clerical, there seems to be little justification for doing so if it is possible that examiners are actually being fooled by close non-matches.