NCJRS
National Criminal Justice Reference Service

Office of Justice Programs
BJA BJS NIJ OJJDP OVC SMART

The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

Document Title:	**Testing the Accuracy and Reliability of Palmar Friction Ridge Comparisons: A Black Box Study**
Author(s):	**Heidi Eldridge, Marco De Donno, Christophe Champod**
Document Number:	**304606**
Date Received:	**April 2022**
Award Number:	**2017-DN-BX-0170**

This resource has not been published by the U.S. Department of Justice. This resource is being made publicly available through the Office of Justice Programs' National Criminal Justice Reference Service.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

National Institute of Justice – Office of Justice Programs

Research and Development in Forensic Science for Criminal Justice Purposes

Award # 2017-DN-BX-0170

Testing the Accuracy and Reliability of Palmar Friction Ridge Comparisons: a Black Box Study

By Heidi Eldridge, Marco De Donno, Christophe Champod

Project Period: January 1, 2018 – May 31, 2020

Award Amount: $370,454

Updated Final Research Report submitted May 31, 2020

Heidi Eldridge, MS (PI)
Research Forensic Scientist
RTI International
3040 E. Cornwallis Rd.
Research Triangle Park, NC 27709
Email: heldridge@rti.org

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Summary of the Project

Major goals and objectives/Research Questions

Much has been made in recent years of the need to establish practitioner error rates for forensic science analyses. From 2009's National Research Council report [1] to the 2016 President's Council of Advisors on Science and Technology report [2], there has been a growing consensus among both critics and practitioners that the practice of forensic science requires experts to be able to report a value that reflects a discipline-wide estimate of the accuracy of their conclusions.

Prior to this research, two large-scale studies into the accuracy of latent print comparisons had been completed [3, 4]. In the first, the FBI/Noblis black box study [3], a false positive rate of 0.1% and a false negative rate of 7.5% were reported. However, this study did not address the accuracy of palm comparisons; only distal phalanx impressions (fingerprints) were presented to participants. Although the second study [4], by the Miami-Dade Police Department, did incorporate palmar comparisons into their experimental design, two limitations prevent its use as an estimate of palm comparison accuracy. The first is that they did not report the error rate obtained for palmar comparisons, instead lumping the error rates for all comparisons together. The second is that their different source trials were constructed by the expedient of providing 3 sets of randomly selected exemplars for comparison that did not come from the same donor as the mark in question. Without any deliberate attempt to find close non-match distractors, it is highly unlikely that the different source trials presented a meaningful challenge simply by coincidence.

We suggest that there are three criteria that should be met to properly establish an informative error rate for palmar impressions. These criteria are:

1. Palmar impressions should be presented and error rates constructed for them separate from those of distal phalanges;
2. Test impressions at different quality levels should be used and error rates calculated for each so that meaningful comparisons to casework images can be made; and
3. Close non-matches should be incorporated to present a realistic chance of making a false-positive error.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The goal of this study, then, was to construct a black box study modeled after the FBI/Noblis study that meets the above criteria in order to answer the main research question: what is the discipline-wide error rate estimate for the comparison of palmar impressions. Anecdotally, approximately 30% of comparison casework in forensic laboratories involves palmar impressions, and there is limited scientific foundation to support a claim that palm comparisons are analogous to fingerprint comparisons in any way that would allow for the extrapolation of the error rate from one to the other. One could reasonably argue that a higher false negative rate is expected for palmar impressions, given that there is a much larger area to search, that orientation clues are often ambiguous or missing, and that practitioners often receive less training and practice in this area compared to the comparison of fingerprints. In addition to the main research question, this research will also investigate questions of examiner consensus on decisions, appropriateness of responses that disagree with the consensus opinion, and applicability of confidence and credible intervals to court testimony.

Research design and methods

Palm marks of known sources were collected from 50 individuals at 6 partner laboratories (Arizona DPS; Columbus, MS Forensic Lab; Douglas County, NE Sheriff's Office; Durham, NC PD; Illinois State Police; and University of Lausanne). All donors read and signed an informed consent statement that had been reviewed and approved by RTI's IRB prior to donating any impressions.

The donated impressions yielded 725 known source mark/exemplar pairings (cases). Marks were made on a variety of substrates using a variety of development techniques to mimic casework. From this pool, 526 cases were selected as study samples. The study samples were drawn from the pool such that 400 cases (76%) were same source trials and 126 cases (24%) were different sources trials. Same source trials were manually compared by the principal investigator (PI) to ensure that overlapping areas containing congruent features were present in both impressions to make the comparison "fair". Exemplars for different source trials were located by searching marks in a Morpho/IDEMIA palm AFIS database of approximately 25,000 records and returning 2 candidates for each mark – the top ranked candidate by the AFIS system, and the candidate that was considered to be the closest non-matching candidate by one of the researchers. A subset of these pairings were selected for inclusion in the study.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The mark sizes used in the study were 117 small (22.2%), 284 medium (54%), and 125 large (23.8%). Small was defined as approximately 1 inch in diameter or smaller; several small, fragmented areas; or the area immediately surrounding a delta only. Large was defined as more than half of the thenar or hypothenar area, or approximately ½ the size of a palm or larger in any direction. Medium was defined as everything in between small and large.

The 526 marks used in the study were distributed among different palm regions as follows:

Palm Area	Number of Marks
Bottom half of palm	14
Carpal delta area	9
Center of palm	9
Full palm	10
Hypothenar only	84
Interdigital only	242
Thenar only	114
Interdigital and center area	1
Interdigital and hypothenar	10
Interdigital and thenar	3
Thenar web	4
Wrist bracelet area	1
Writer's palm	25

Once the cases had been selected, random draws of 75 cases were made for each participant such that each received 53 same source and 22 different source trials. The same source trials were distributed as closely as possible to present each participant with 8 no value images, 10 easy comparisons, 12 medium difficulty comparisons, 21 hard comparisons, and 2 inconclusives. These categories were assigned based on the training and experience of the PI.

"No value" designations were based on the mark alone and were the images where we expected that most or all participants would declare the mark of no value. Easy comparisons had high quantity and clarity as well as a large area of overlapping ridge detail, although they may or may not have been presented in the correct orientation (i.e. the mark may require rotation). Hard comparisons had either a very low minutiae count in the overlapping region, low clarity in either the mark or the exemplar in the overlapping region, or very little overlapping region. Medium comparisons were between easy and hard in terms of difficulty. If a comparison was of medium difficulty once the correct region was located, but it took a long time to locate, the difficulty was adjusted up to "hard" under the theory that it would be easier to miss. "Inconclusive" designations were those

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

that were ground truth same source pairs, but the information in one or both images was so degraded that the PI expected most or all participants would judge the comparison to be inconclusive. For different source trials, the difficulty was rated as CNM1 or CNM2, depending on whether the exemplar had been selected by virtue of being the top candidate returned by the AFIS system (CNM1), or by being manually selected as the closest non-match (CNM2) by one of the research team.

The 526 marks used in the study were distributed by difficulty level as follows:

Difficulty Level	Number of Marks
No Value	44
Easy	70
Medium	150
Hard	108
Very Hard	16
Inconclusive	12
CNM1 (different source trial)	60
CNM2 (different source trial)	66

The trials were listed in each participant's account in a random order but could be accessed in any order by the participant.

Participants completed all trials within a custom version of the online PiAnoS interface, developed by University of Lausanne in part using previous NIJ funds [2010-DN-BX-K267]. Within this interface, participants were presented first with the mark image alone and were asked to assess whether it was suitable for identification, suitable for exclusion only, or not suitable. Any marks that were deemed not suitable were terminated at this point, without comparison. Marks that were suitable for identification or exclusion proceeded on to the comparison phase. In the comparison phase, the mark and the putative print were presented side-by-side. Participants were asked to perform a comparison, then render a conclusion of identification, exclusion, or inconclusive. If exclusion or inconclusive were selected, participants were requested to select a reason for their decision from a drop-down list of options. Finally, participants were asked to rate the difficulty of the comparison as Very Easy/Obvious, Easy, Moderate, Difficult, or Very Difficult.

Participants were provided only limited tools for their analysis and comparison. Because we were testing accuracy of conclusions, and not skill at digital processing, we wanted all conclusions to be rendered based upon the same visual data. Therefore, no processing tools, such as brightness and contrast adjustments, or ridge and

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

furrow inversion, were provided. However, participants were provided the ability to zoom in and out, move the images around, rotate the mark, annotate minutiae, trace ridges, pair minutiae between the two images, and designate quality zones. None of these annotation tools were required to be used, but were available to suit the working style of the individual participant. Nonetheless, many participants chose to annotate, and to write notes in the free text box provided. These data were often helpful in understanding the underlying cause(s) of errors, when they occurred.

Analytical and Data Analysis Techniques

The data analysis proceeded through the following steps:

(1) Results were sorted according to the task involved (analysis and comparison) and considering results from all participants together or results for each individual.

(2) The ground truth status was set either by the known state of the case submitted or alternatively considering the majority vote by the participants as the ground truth by proxy. This allowed us to compute rates of disagreement with the majority for decisions reached in analysis and also to obtain estimates had the ground truth not been known by fact. When the ground truth associated with the comparison is known, the "inconclusive" conclusions are not counted in our main analysis, but were included in a secondary analysis for an apples-to-apples comparison with the FBI/Noblis results. When considering disagreement with the majority vote, the "inconclusives" are taken into account.

(3) Graphical representations of the results were prepared in the form of confusion matrices, overview of the results by cases respectively for all participants together and for each participant individually.

(4) The error rates were computed for ground truth comparison outcomes in terms of false positive rates (FPR), false negative rates (FNR), positive predictive value (PPV) and negative predictive value (NPV). Sensitivity and Specificity were also computed.

(5) Rates of disagreement were computed for majority vote decisions in both analysis and comparison.

(6) Participants are compared in terms of their respective FPR and FNR.

(7) A Shiny App was developed to present the results and allow access to them through the Internet. Results can be viewed at: https://cchampod.shinyapps.io/Results_BBStudy/. The application computes confusion

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

matrices, error rates, and disagreement rates and presents all results associated with cases and all participants.

(8) We have implemented a second Shiny App (https://cchampod.shinyapps.io/app_CI/) that computes two-sided confidence intervals on any of the rates (error rates or disagreement rates) obtained either in analysis or comparison. The application computes Bayesian credible intervals as well as frequentist credible intervals.

All statistical analysis and graphical representations were carried out in *R* version 3.5.3 (R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.) using *RStudio* as the integrated development environment (version 1.2.5001, RStudio, Inc., https://rstudio.com). Results were organized in *Shiny* Apps (W. Chang, J. Cheng, JJ Allaire, Y. Xie and J. McPherson (2019). shiny: Web Application Framework for R. R package version 1.3.2. https://CRAN.R-project.org/package=shiny). The development took advantage of various additional libraries among them *tidyverse* (H. Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse) and *plotly* (Carson Sievert (2018) plotly for R. https://plotly-r.com). Credible and confidence intervals are computed using the *proportion* library (M.Subbiah and V.Rajeswaran (2017). proportion: Inference on Single Binomial Proportion and Bayesian Computations. R package version 2.0.0. https://CRAN.R-project.org/package=proportion).

Expected Applicability of the research

The establishment of a discipline-wide error rate estimate for palmar comparisons will strengthen the practice by allowing examiners to testify to the foundational validity of these cases when they arise. Additionally, the court primer, while intended to assist the courts in understanding error rate testimony from latent print experts, can also assist those examiners in crafting their testimony.

Participants and other collaborating organizations

Participants were recruited through a combination of email lists, conference announcements and advertisements, a brief article in the IAI's newsletter *IDNews*, and a recruitment room set up at the 2018 IAI educational conference. Latent print examiners and trainees were invited to participate. Participants enrolled

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

voluntarily in the study and their identities were held confidential using anonymous usernames. All contact with participants was done through a confidential liaison so the researchers were never aware of participants' identities. All possible link between usernames and true identities will be destroyed at the end of the period of performance. All participants read and signed an informed consent statement that had been reviewed and approved by RTI's Institutional Review Board (IRB) prior to participation.

328 participants enrolled in the study. Of these, 134 completed all 75 trials and another 76 completed between 1 and 74 trials, for a total of 210 active participants. Only demographic information for active participants is reported. Approximately 75% of participants were female (~25% male). 10.95% of participants were aged 20-29, 42.86% were aged 30-39, 32.86% were aged 40-49, 10.95% were aged 50-59, and 2.38% were 60 or older. 27.14% of participants had been latent print examiners for less than 5 years, 26.67% for between 5 and 9 years, 35.71% for between 10 and 19 years, 6.19% for between 20 and 29 years, and 4.29% for 30 years or more. 45.71% of participants worked for US local agencies, 28.57% for US state agencies, and 3.81% each for US federal and private agencies. The remaining 18.1% worked for foreign agencies. Approximately 71% of participants worked for accredited laboratories and 29% for non-accredited laboratories. 44.29% of participants reported being IAI CLPE certified, 40.48% reported no certification, and 21.43% reported certification by an employer or some other type of certification (some participants reported more than one kind of certification, resulting in a total of more than 100%). Finally, approximately 7% of participants were trainees at the time of their participation.

Outcomes

Activities/Accomplishments

We undertook a number of activities in the completion of this project. Chief among them were:

- Collect palmar impressions of known source along with 3 sets of exemplars each from 50 donors
- Compare palm marks to their source prints to establish presence of corresponding features
- Search a sub-set of palm marks through an AFIS to locate prints for use in different sources trials
- Recruit participants to take part in the study and provide them with instructions and informed consent
- Update PiAnoS to meet the needs of the project

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

- Select images for use in the study and populate case lists for participants
- Manage data collection, including sending responses to anonymized participant queries
- Extract and clean data
- Analyze data
- Develop 2 online applications for the exploration of the data
- Prepare and deliver dissemination materials, including manuscripts for submission to peer-reviewed scientific journals, webinar delivered through Forensic Technology Center of Excellence, conference and educational meeting presentations, and court primer

Results and Findings

In all, 12,279 analysis determinations and 9,460 comparison decisions were rendered. Each case was viewed by an average of 23 examiners. We present here only a global overview of the results for all participants. Please refer to the Shiny Apps to access all the results by activity and per participant and to the manuscript we have submitted to a peer-reviewed scientific journal for more in-depth discussion of our findings.

<u>*Analysis*</u>

A high level of variability was observed for the analysis decision. Because there is no objective "ground truth" for suitability, a ground truth by proxy decision was assigned to each mark by taking the majority vote. As can be seen in Figure 1, there was never a consensus decision that mark was suitable only for exclusion. Figures 1 and 2 both illustrate that the No value and Suitable for identification decisions are not highly reliable.

Figure 1. Confusion matrix of Analysis responses

Individual decisions taken by participants	NV	VID	VEO
VEO	137	178	0
VID	599	8959	0
NV	1618	788	0

Majority vote
as voted by the majority

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Figure 2. Disagreement rates for Analysis responses

Summary statistics for all users

	False Pos Rate	False Neg Rate	Pos Pred Value	Neg Pred Value
No value	0.0794	0.3127	0.6725	0.9255
Suitable for identification	0.2545	0.0973	0.9373	0.6450

Comparison (against known Ground Truth)

Twelve false positive errors were made by 8 participants in the study, resulting in a false positive error rate of 0.7% and a PPV of 99.8%. No two false positive errors were made on the same mark-print pairing, and no false positive errors were made by trainees. 552 false negative errors were made, resulting in a false negative error rate of 9.5% and an NPV of 76.3%. Note that inconclusive decisions have been removed from the dataset entirely for these calculations so that they count neither for nor against the total error rates. To compare these results to those obtained in the FBI/Noblis study (which were reported with inconclusive decisions included), we would need to remove the inconclusive responses from their data as well. By this calculation, their false positive error rate becomes 0.2% (6/3953) and their false negative error rate becomes 14.2% (611/4314). This is interesting because their false positive error rate for fingers is actually higher than ours for palms by this reckoning. However, if the comparison is done the other way, by adding the inconclusives into our data (and not counting them as errors) rather than omitting them from the FBI/Noblis data, the resulting figures are a FPR of 0.5% for palms versus 0.1% for fingers and a FNR of 8% for palms versus 7.5% for fingers. Although the FPR is not hugely affected by whether inconclusives are included or omitted in both studies, the FNR is. This is likely because between the two studies, there are large differences in the rates of inconclusives for same source and different sources trials. This is just one

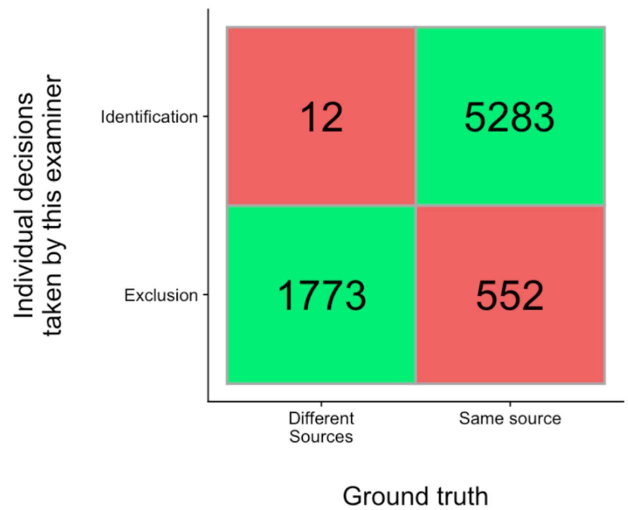


Figure 3. Confusion matrix for comparison decisions.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

of many ways in which the two studies should not be directly compared even though they drive at the same basic question. There are just too many unknown variables that differ between finger and palm comparisons and between the two studies. Figure 3 presents the confusion matrix for comparison decisions evaluated against ground truth (omitting inconclusives).

<u>Comparison (in relation to majority vote)</u>

Although establishing an error rate against known ground truth was the aim of this research, it is also instructive to consider individual examiners' responses in comparison to consensus responses to each pairing as voted by majority. There are two reasons this information is interesting: (1) sometimes, the consensus gets it wrong (e.g., the consensus decision is exclusion when the known ground truth is same source); and (2) sometimes the consensus decision can illuminate when an examiner is being too risk-averse (e.g., rendering an inconclusive when the majority reached the correct ground truth decision) or too risk-tolerant (e.g., rendering a conclusion that aligns with ground truth, but may not be amply supported as most experts would agree that an inconclusive decision should be rendered). For these reasons, we present Figures 4 and 5 to show these data.

Figure 4. Confusion matrix of comparison decisions compared to the majoritynvote. Note that although there are 45 cases where a decision of ID disagrees with the majority, only 9 of these were actual false IDs. The other 36 instances shown were cases in which the consensus incorrectly excluded on a same source trial. Also note the two boxes in the inconclusive column in which individuals reached definitive decisions against the consensus. These are the cases that may be considered overly risk-tolerant.

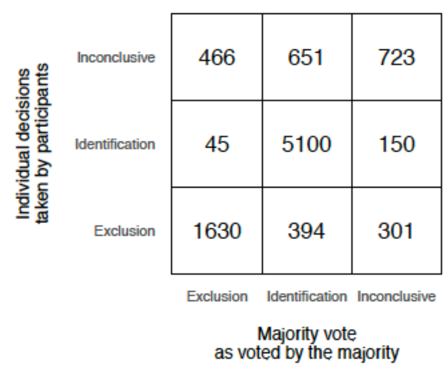


Figure 5. Screenshot of results app, showing examples of cases where the consensus got it wrong (case 423, a same source trial where the majority rendered an exclusion decision); and where individuals exhibit both risk-averse (case 468, a different source trial in which most examiners reached the ground truth decision, but some chose inconclusive) and risk-tolerant (case 316, a same source trial. Note that while a few people “correctly” made the ID, a few falsely excluded this pair, and many more either called the mark no value, or rendered an inconclusive decision. Clearly, this comparison was complex and had very little reliable information available upon which to rely) behaviors in comparison to the consensus.

Correlation between false positive errors and demographics

Eight examiners committed false positive errors in this study. Four participants made one false positive error each, while the other four made two false positive errors each. Although the sample size of people committing false positive errors is too small to perform a rigorous statistical analysis, there are some interesting commonalities in the data that are worth observing.

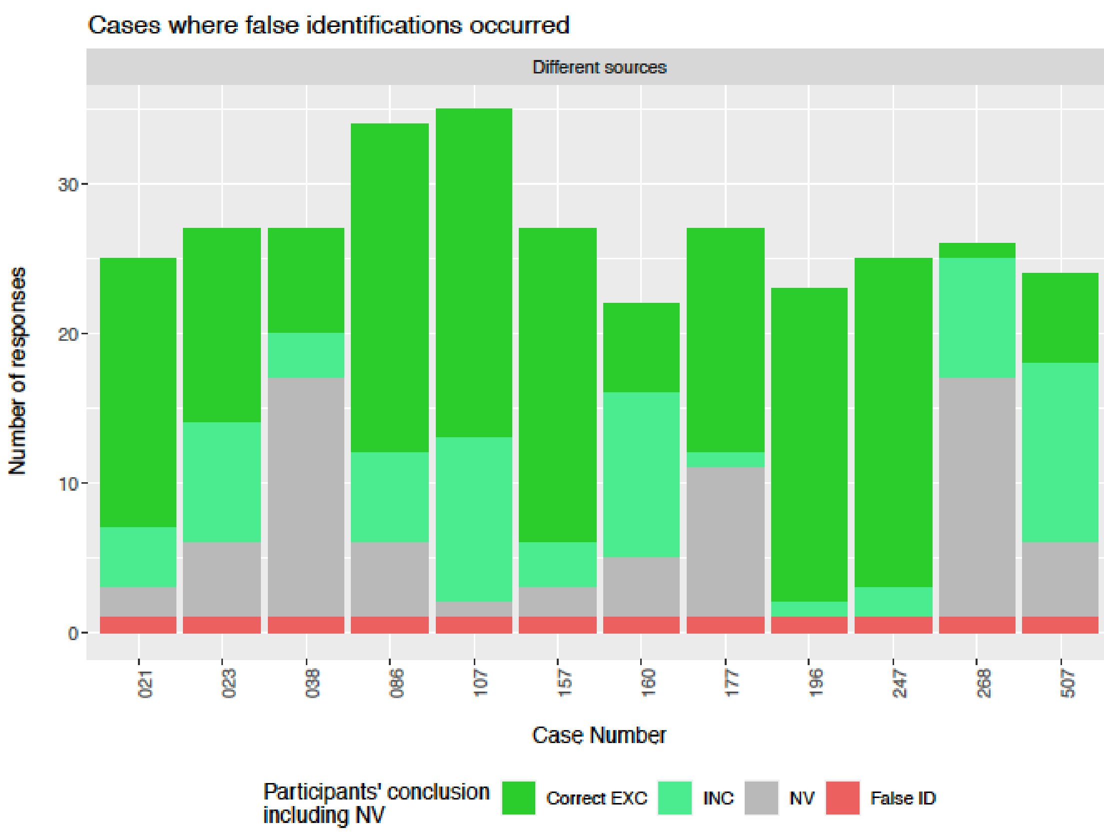
First, and possibly most importantly, although 94.76% of participants reported currently working as latent print examiners, two of the examiners who each made two of the false positive errors answered “No” to the question “Are you currently or have you previously been employed as a latent fingerprint examiner?”. Together these two examiners were responsible for 1/3 of the false positive errors made in the study, yet they were from a pool that represented only 2.86% of the study population. Another interesting pattern is that although only 18.1% of the study participants reported working for an agency

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

outside of the U.S., 6 of the 12 false positive errors (50%) were made by participants from non-U.S. agencies, a disproportionate number to their presence in the study population. Altogether, 8 of the 12 false positive errors (66.7%) were made by participants who were either non-active LPEs, non-U.S. examiners, or both. This leaves 4 of the 12 false positive errors that were committed by U.S. examiners who are currently active LPEs. For comparison, the FBI/Noblis study reported 96% participation by current LPEs, but only 1% participation from non-U.S. examiners.

Additionally, as shown in Figure 6, there were no clear patterns among the decisions of other examiners who viewed the same pairs for the 12 trials in which false positive errors were made, indicating that nothing about these pairs was universally problematic for examiners that would provide a warning.

Figure 6. Conclusions reached by participants in the trials where a false identification was reported.



False negative error rates stratified by size, difficulty, and palm area

While historically only a single false positive and false negative error rate have been reported for discipline error rate studies, an aim of this research was to examine whether different error rates were observed

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

depending on the size of the mark, the difficulty of the comparison, or the area of the palm from which the mark originated. This information could be of use to examiners testifying in court, who would be able to cite the error rates that most closely resembled the conditions in the case at hand. Unfortunately, as only 12 false positive errors were made against ground truth in this study, it would not be relevant to calculate error rates using such small numerators, once those 12 errors had been parsed into sub-categories. Thus, we only address the stratified error rates of false negative errors within this report. The “Same Source Trials” columns include only cases in which a comparison conclusion was rendered and exclude inconclusive decisions.

The summary of false negative error rates stratified by the size of the mark is as follows:

Size	False Negative Errors	Same Source Trials	False Negative Error Rate
S	103	1392	7.4%
M	272	3624	7.5%
L	177	1884	9.4%

It is apparent from these data that the size of the mark is not a determining factor in whether or not a false negative error will be made.

The summary of false negative error rates stratified by the difficulty of the comparison (as rated by the PI) is as follows:

Difficulty	False Negative Errors	Same Source Trials	False Negative Error Rate
NV	56	435	12.9%
Inconclusive	17	228	7.5%
Easy	51	1604	3.2%
Medium	123	1851	6.7%
Hard	268	2549	10.5%
Very hard	37	233	15.9%

Here an effect is observed. It is clear that as the difficulty of the comparison increases, so does the false negative error rate. These data support the idea of defining thresholds for comparison difficulty and documenting these levels in case notes.

The summary of false negative error rates stratified by the area of the palm from which the mark originated is as follows:

Palm Area	False Negative Errors	Same Source Trials	False Negative Error Rate
Bottom half	6	232	2.6%
Carpal delta only	10	181	5.5%
Center	11	78	14.1%
Full Palm	1	82	1.2%

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Hypothenar	95	1,201	7.9%
Interdigital	161	3,008	5.4%
Int/Center	1	20	5.0%
Int/Hypo	10	195	5.1%
Int/Thenar	5	53	9.4%
Thenar	220	1,522	14.5%
Thenar Web	4	38	10.5%
Writer's Palm	28	290	9.7%

Again, a distinct effect of palm area is noted. There are clearly areas of the palm that pose a greater challenge to examiners in locating marks and those that pose less challenge.

Limitations

As with all black box studies, this one has limitations. Because this is a structured testing environment, and not casework, there are inherently differences between the test and a real-world situation. How these differences may effect the accuracy of the error rate estimate is unknown. For example, the mix of same source to different sources trials may be different from casework. In addition, the difficulty of the comparisons, mix of substrates and development techniques, required comparison interface, and required conclusion labels may differ from what an examiner is used to. In this study, there was a small amount of searching introduced because the palm is a large area and marks were not necessarily presented in the correct orientation, both of which necessitated searching within the provided print. However, in casework, there is typically more searching to be done among multiple prints. Finally, the fact that examiners were aware they were being tested invokes the Hawthorne Effect, which states that people behave differently when they know they are being watched or tested. Whether that would translate into being more cautious or more relaxed is unknown, but either way, it is a confounding factor that could affect outcomes. Because of these limitations, the figures reported from this research should be taken only as an *estimate* of the error rate of the discipline at performing palmar comparisons.

Artifacts

List of products

The following products have been created as a result of this research:

- Scientific article manuscript entitled "Mind-set – How bias leads to errors in friction ridge comparisons (submitted to FSI 3/26/20)

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

- Scientific article manuscript entitled "Testing the Accuracy and Reliability of Palmar Friction Ridge Comparisons – A Black Box Study (submitted to FSI 3/27/20)
- Forensic Technology Center of Excellence webinar entitled "Results of a black box study on the accuracy and reliability of palm print comparisons," archived at https://forensiccoe.org/webinar/results-of-a-black-box-study/
- Court Primer archived at https://doi.org/10.5281/zenodo.3734560
- Online application for review of results, available at https://cchampod.shinyapps.io/Results_BBStudy/
- Online application for exploration of confidence and credible intervals associated with results, available at https://cchampod.shinyapps.io/app_CI/
- Conference presentations of results at the following conferences and educational meetings:
    - 2019 IAI Educational Conference
    - 2019 International Fingerprint Research Group (IFRG) meeting
    - 2019 Florida Division IAI Educational Conference
    - 2019 Duke University Law School symposium
    - 2019 NACDL and Cardozo Law National Forensic College
    - 2020 NIJ R&D grantees symposium
    - 2020 AAFS Educational Conference

Data sets generated

The data collected from participants in this study include demographic and policy survey data; and annotations, notes, and analysis and comparison conclusion data. All raw data have been archived by RTI and Unil. The analysis and comparison results data are available to any interested parties in the online application mentioned above.

Dissemination activities

The results of this research have been disseminated in multiple ways, which have been listed above, under "List of products." The results have been presented at numerous regional, national, and international

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

conferences; presented in a webinar that has been archived and can be viewed by anyone who missed the live presentation; and discussed thoroughly in two manuscripts that have been submitted for publication to FSI.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Bibliography

(1) National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, D.C.: The National Academies Press, 2009.

(2) President's Council of Advisors on Science and Technology. Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Executive Office of the President, Washington DC, 2016.

(3) Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. Proceedings of the National Academy of Sciences, USA 2011; 108 (19):7733-7738.

(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. National Institute of Justice, Washington DC, 2014.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.