Informing the Judgments of Fingerprint Analysts Using Quality Metric and Statistical Assessment Tools

Glenn Langenburg[1,2], Christophe Champod[2], Thibault Genessay[2], Jordan Jones[1]

ABSTRACT

The aim of this research was to evaluate how fingerprint analysts would incorporate information from newly developed tools into their decision making processes. Specifically, we assessed the impact of the following tools: 1) a "quality" (i.e. the clarity of the friction ridge features) measurement tool, 2) a tool to provide statistics that measure the strength of the evidence (i.e. a probabilistic assessment of the corresponding features in a fingerprint comparison), and 3) consensus information from a group of trained fingerprint experts. The measured variables for the effect on examiner performance were the accuracy and reproducibility of the conclusions against the ground truth (including the impact on error rates) and the analyst variation during feature selection and comparison.

The results showed that analysts using the consensus information from trained fingerprint experts were affected the most. The groups with these tools demonstrated more consistency and accuracy in minutiae selection during the Analysis phase. These groups also demonstrated higher accuracy, sensitivity, and selectivity in the decisions reported. The quality tool also impacted minutiae selection (which in turn had some apparent influence on the reported decisions); the probability tool did not impact the reported decisions.

OBJECTIVES

Our aims were to:
1) Determine how information regarding the clarity of friction ridge features (as provided in various formats to the participants) informed the judgments of fingerprint analysts.
2) Determine how information representing the strength of the corresponding friction ridge features (as provided in a likelihood ratio format) informed the judgments of fingerprint analysts.
3) Determine how information regarding the strength of the corresponding features when provided by ***other*** fingerprint experts, informed the judgments of the participating analysts. (This is typically referred to as a "consultation" among fingerprint experts).

PROCEDURES

Approximately 600 fingerprint analysts were invited to participate in this research. Potential participants represented a wide range of experience, agency size, state and local agencies, etc. A CD-ROM containing an applet was mailed to each potential participant. Each CD-ROM had a unique user name and strong password (to protect anonymity) and the applet allowed the user to log in to a secure server at the University of Lausanne (UNIL) in Switzerland. Users logged into a platform, developed by UNIL, called PiAnoS (Picture Annotation System). Over 200 users successfully[3] logged in and started the study. At

[1] Minnesota Bureau of Criminal Apprehension (BCA), St. Paul, MN
[2] Ecole des Sciences Criminelles, University of Lausanne (UNIL), Switzerland
[3] Many potential participants were unfortunately unable to participate due to technical problems stemming from overly restrictive government agency firewalls, protections, IT permissions, hardware/software compatibility issues, etc.

the close of the data collection period, 176 users completed all of the 12 experimental trials. This resulted in 176 x 12, or 2112 completed trials. Some basic information (e.g. sex, years of experience, status of expertise, etc.) was captured for each participant using a short questionnaire during the first log-in session.

The PiAnoS platform allowed for a controlled presentation of the latent print and exemplar images in a manner that follows ACE-V—namely, an analysis is first conducted on the latent print, absent an exemplar, and then the exemplar is presented for comparison, and finally a decision is reported. Analysts were provided a drop down menu to select from three reportable decisions: “Identification”, “Inconclusive”, and “Exclusion”. A comment box was provided for further explanation when “Inconclusive”[4] was chosen as a decision. During Analysis and Comparison phases, tools for selecting and annotating features were provided (Figures 1 and 2).

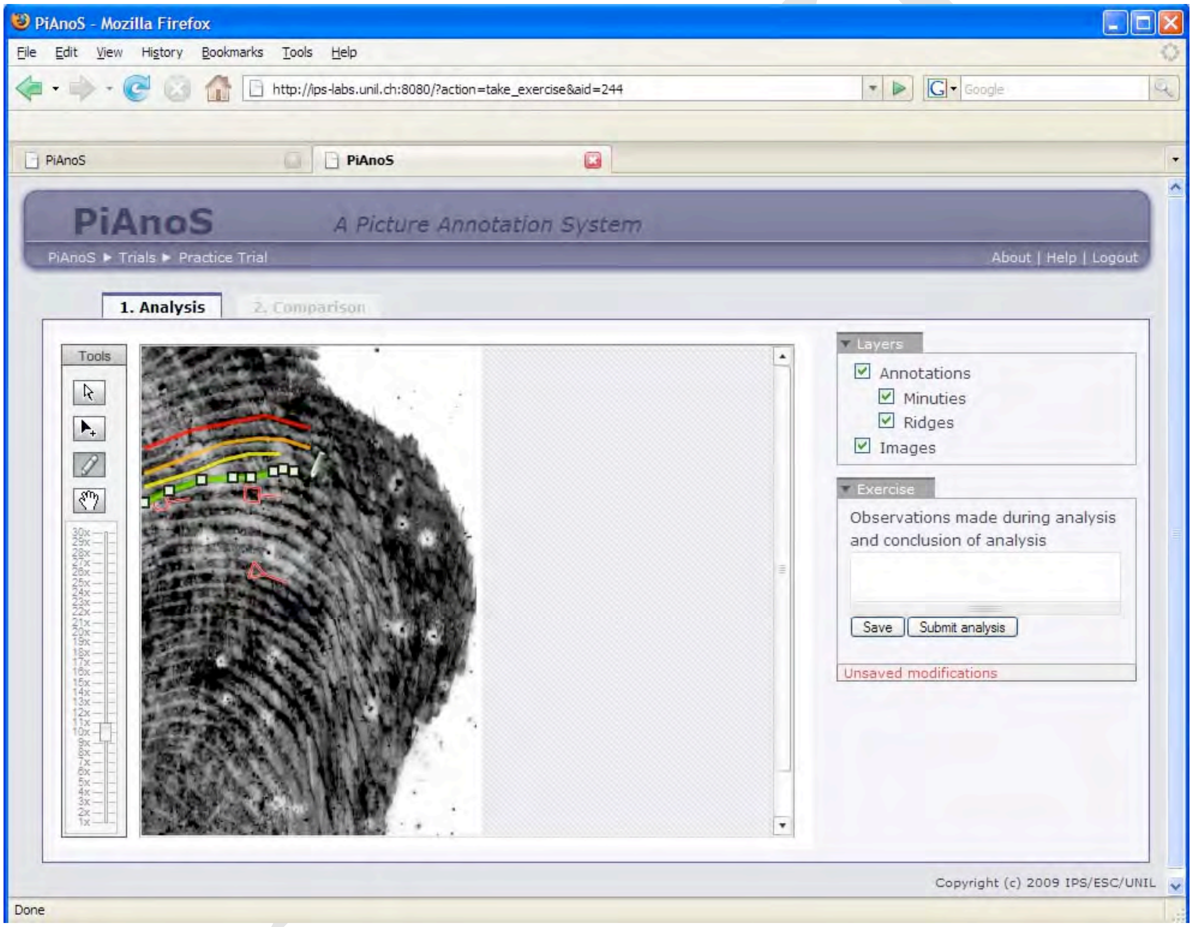
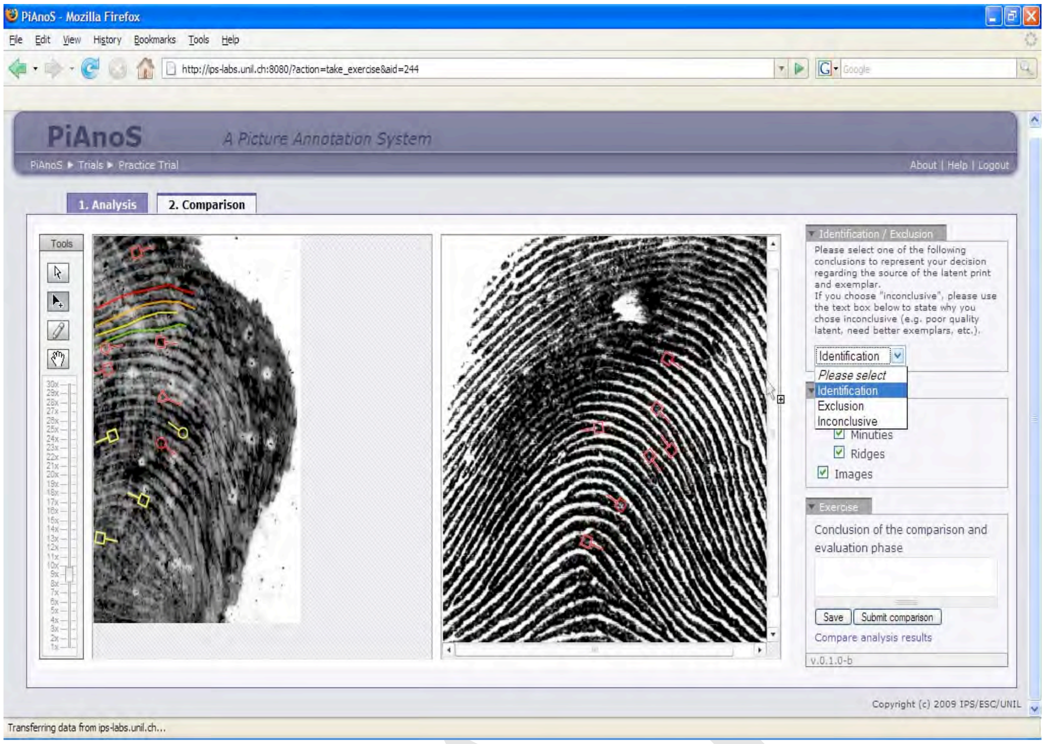

PiAnoS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ips-labs.unil.ch:8080/?action=take_exercise&aid=244
Google
PiAnoS
PiAnoS
PiAnoS
A Picture Annotation System
PiAnoS ▶ Trials ▶ Practice Trial
About | Help | Logout
1. Analysis
2. Comparison
Tools
30x 29x 28x 27x 26x 25x 24x 23x 22x 21x 20x 19x 18x 17x 16x 15x 14x 13x 12x 11x 10x 9x 8x 7x 6x 5x 4x 3x 2x 1x
Layers
☑ Annotations
☑ Minuties
☑ Ridges
☑ Images
Exercise
Observations made during analysis and conclusion of analysis
Save
Submit analysis
Unsaved modifications
Copyright (c) 2009 IPS/ESC/UNIL
Done

Figure 1: Screen shot of PiAnoS showing the latent print available for annotation of minutiae during the Analysis phase. Minutiae annotations and the ridge marking tool can be seen.

[4] For example, common reasons for “Inconclusive” were requests for additional, better quality exemplars or that the analyst felt there was insufficient, reliable features in the latent to effect a positive identification, etc.


PiAnoS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ips-labs.unil.ch:8080/?action=take_exercise&aid=244
Google
PiAnoS
A Picture Annotation System
PiAnoS ▶ Trials ▶ Practice Trial
About | Help | Logout
1. Analysis
2. Comparison
Tools
Identification / Exclusion
Please select
Identification
Exclusion
Inconclusive
Minuties
Ridges
Images
Exercise
Save
Submit comparison
Compare analysis results
v.0.1.0-b
Copyright (c) 2009 IPS/ESC/UNIL
Transferring data from ips-labs.unil.ch...

Please select one of the following conclusions to represent your decision regarding the source of the latent print and exemplar.
If you choose "inconclusive", please use the text box below to state why you chose inconclusive (e.g. poor quality latent, need better exemplars, etc.).

Identification

☑ Minuties
☑ Ridges
☑ Images

Conclusion of the comparison and evaluation phase

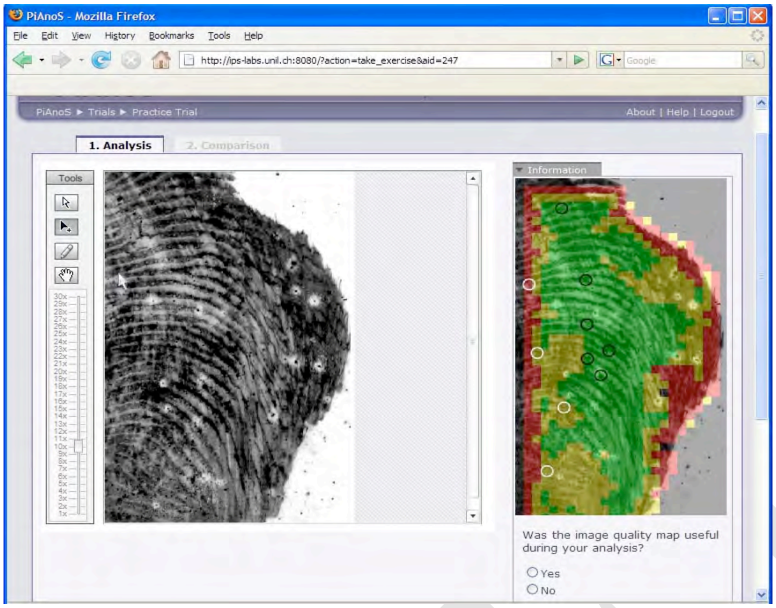
Figure 2: Screen shot of PiAnos showing the availability of the exemplar during the Comparison phase. The decision drop-down menu is shown on the right side of the screen.

Participants were given 12 latent print comparisons to perform. They could save their work and return at their leisure, working at their own pace. Analysts were randomly assigned upon their first log-in to one of six experimental groups. Each experimental group applied combinations of the tools that were to be tested. The tools that were provided and the corresponding groups were as follows:

Variables	**No LR Tool**	**Expert Consultation**	**LR Tool**
No Quality Map	**Group 1** No Quality Map No LR tool	**Group 2** Expert consensus minutiae	**Group 3** No Quality Map LR Tool
Quality Map	**Group 4** Quality Map No LR Tool	**Group 5** Expert consensus decisions	**Group 6** Quality Map LR Tool

Quality Map Tool: Participants under the quality map tool condition were provided information regarding the quality (clarity) of the ridge features during the Analysis phase. Low quality regions, containing potentially unreliable features, were distinguished from higher quality regions, according to the tool's assessment. A beta-version[5] of Noblis' Image Quality Tool was used to generate the Quality Maps. However, these images were then cleaned and modified for this experiment. Minutiae selected ***by experts*** during pilot testing were then overlaid onto the Quality Map. Thus a realistic proxy quality tool was produced, but participants were not informed that they were viewing the minutiae selected by other experts; rather, they were led to believe that the minutiae were selected by the software. Only the quality map colors were used from the Noblis software. See Figure 3.

[5] Universal Latent Workstation (ULW) Beta 5.6.0 software was used.

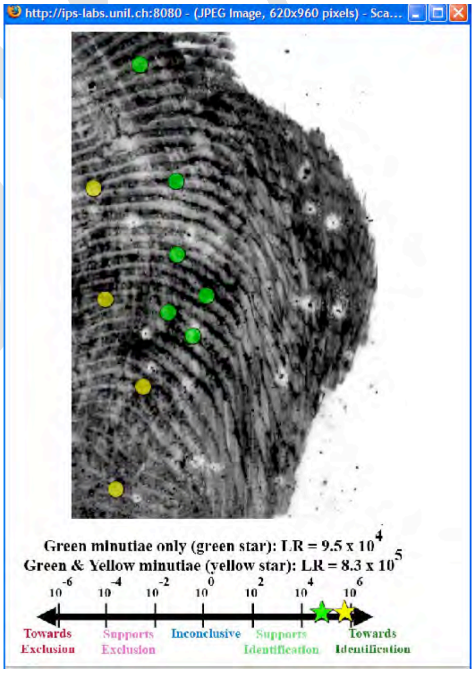

PiAnoS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ips-labs.unil.ch:8080/?action=take_exercise&aid=247
PiAnoS ▸ Trials ▸ Practice Trial
About | Help | Logout
1. Analysis
2. Comparison
Tools
Information

Was the image quality map useful during your analysis?

○ Yes
○ No

Figure 3: A screenshot of PiAnoS showing the Quality Map tool that was provided during the Analysis phase for Groups 4 and 6.

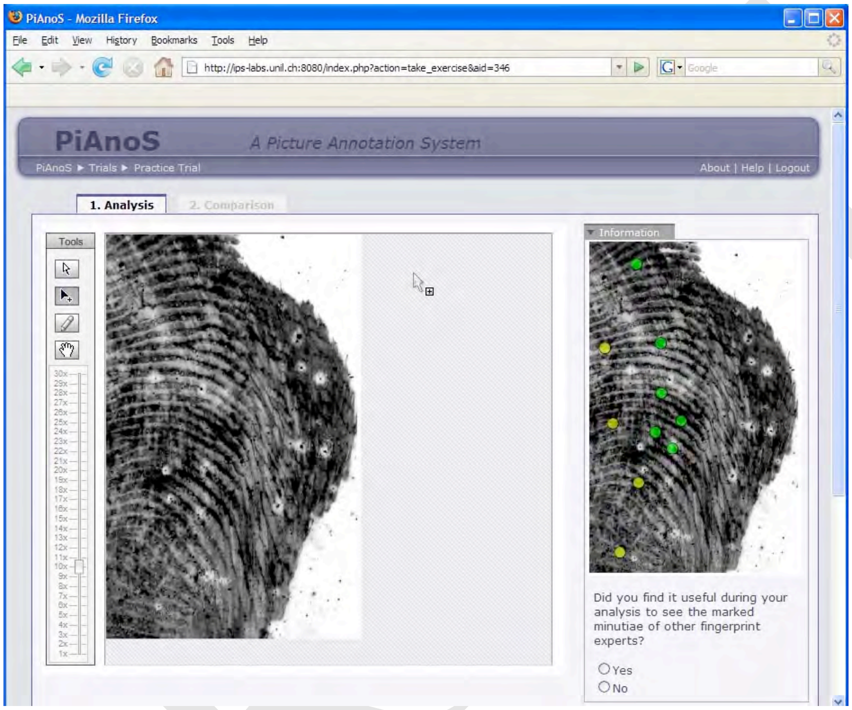
Likelihood Ratio Statistic Tool (LR Tool): Participants under the statistical tool condition were provided information during the Comparison phase. This tool provided a statistic that represented the strength of the corresponding features. The tool used a likelihood ratio to represent the weight of the corresponding minutiae. Larger LRs (much greater than “1”) indicate stronger evidence that the unknown latent print and the known exemplar share a common source. Smaller LRs (much less than “1”) indicate strong evidence the images do not share a common source. The LRs were calculated using a statistical model for fingerprints developed by UNIL. See Figure 4.



Green minutiae only (green star): LR = 9.5×10^{4}
Green & Yellow minutiae (yellow star): LR = 8.3×10^{5}

Figure 4: A screenshot of the LR Tool provided during the Comparison phase to Group 3. Participants in Group 6 received a similar image, but with a Quality Map overlaid on the image.

Expert Consensus Tools: In current practices, in the absence of objective measurement tools, analysts commonly consult each other. Participants under the consultation conditions were either provided with the consensus minutiae or the conclusions of other participating analysts. Prior to starting this study, a pilot test of the trials was performed using 25 experts, each having at least 5 years of experience. Figure 5 shows the Expert Consensus Minutiae map and Figure 6 shows the Expert Consensus Decision table that were provided to Group 2 and Group 5, respectively. Compare the minutiae selected in Figure 3 to those selected in Figure 5 (they are the same).


PiAnoS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ips-labs.unil.ch:8080/index.php?action=take_exercise&aid=346
Google
PiAnoS
A Picture Annotation System
PiAnoS ► Trials ► Practice Trial
About | Help | Logout
1. Analysis
2. Comparison
Tools
30x 29x 28x 27x 26x 25x 24x 23x 22x 21x 20x 19x 18x 17x 16x 15x 14x 13x 12x 11x 10x 9x 8x 7x 6x 5x 4x 3x 2x 1x
Information

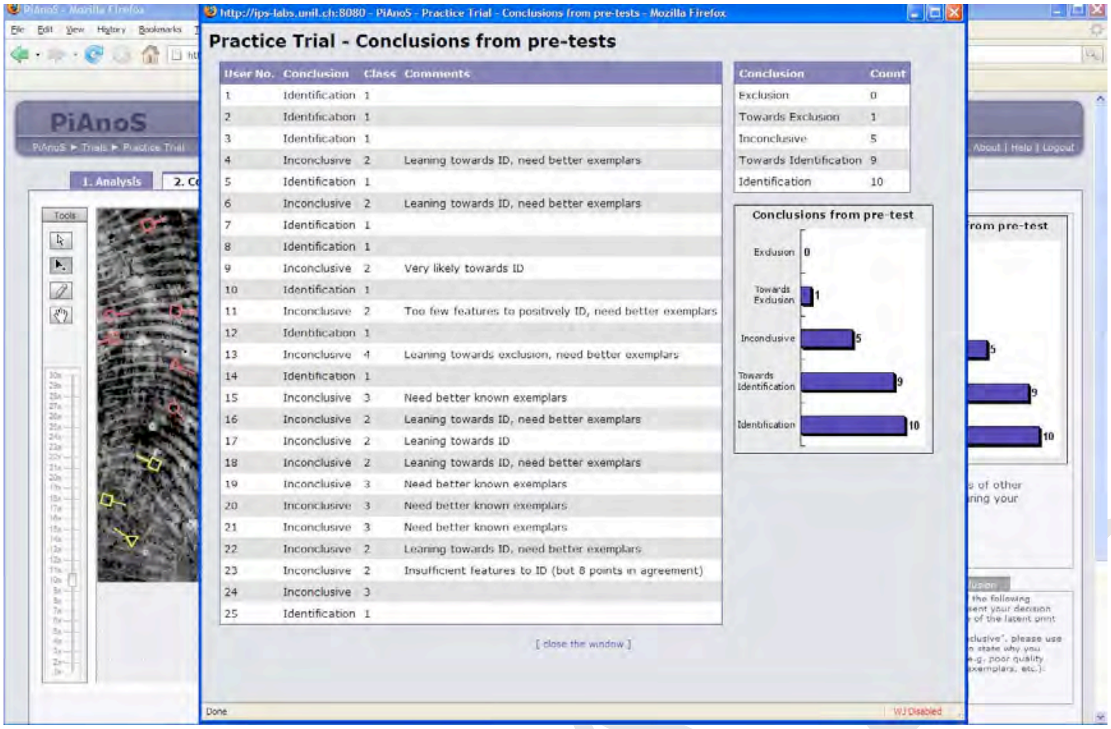
Did you find it useful during your analysis to see the marked minutiae of other fingerprint experts?

○ Yes
○ No

Figure 5: The Expert Consensus Map that was provided to participants in Group 2 during the Analysis Phase. The yellow dots represent minutiae annotated by 50% or more of the 25 experts during pilot testing. The green dots represent minutiae annotated by 75% or more of the experts.

Case Selection: Twelve cases were selected where the ground truth was known for all of the latent prints. The latent prints were produced "naturally" and not altered in any way. Participants were presented with seven trials where the latent prints originated from the same source as the exemplar. These trials (during pilot testing) were determined to be fairly challenging and possessed significant distortions. Some of these same source trials exhibited apparent differences, yet were from the same source. The remaining five trials presented latent prints against exemplars which did not originate from the same source. All five of these cases were close non-matches resulting from searches in IAFIS[6]. Searches were maximized to find the closest candidates that would produce the most challenging trials for experienced analysts. During pilot testing, most of these trials produced errors and disagreement amongst experts. This was evidence that the cases selected were sufficiently difficult to provide a challenge to experienced experts. It was important to use difficult cases to promote errors for a measurable effect on the accuracy, error rate, and variance.

[6] Integrated Automatic Fingerprint Identification System (the Federal AFIS database). At the time of the searches, IAFIS contained approximately 700 million fingerprint images in the database.



http://ips-labs.unil.ch:8080 - PiAnoS - Practice Trial - Conclusions from pre-tests - Mozilla Firefox

Practice Trial - Conclusions from pre-tests

User No.	Conclusion	Class	Comments
1	Identification	1	
2	Identification	1	
3	Identification	1	
4	Inconclusive	2	Leaning towards ID, need better exemplars
5	Identification	1	
6	Inconclusive	2	Leaning towards ID, need better exemplars
7	Identification	1	
8	Identification	1	
9	Inconclusive	2	Very likely towards ID
10	Identification	1	
11	Inconclusive	2	Too few features to positively ID, need better exemplars
12	Identification	1	
13	Inconclusive	4	Leaning towards exclusion, need better exemplars
14	Identification	1	
15	Inconclusive	3	Need better known exemplars
16	Inconclusive	2	Leaning towards ID, need better exemplars
17	Inconclusive	2	Leaning towards ID
18	Inconclusive	2	Leaning towards ID, need better exemplars
19	Inconclusive	3	Need better known exemplars
20	Inconclusive	3	Need better known exemplars
21	Inconclusive	3	Need better known exemplars
22	Inconclusive	2	Leaning towards ID, need better exemplars
23	Inconclusive	2	Insufficient features to ID (but 8 points in agreement)
24	Inconclusive	3	
25	Identification	1	

Conclusion	Count
Exclusion	0
Towards Exclusion	1
Inconclusive	5
Towards Identification	9
Identification	10

Figure 6: A screen shot of the Expert Consensus decision table shown to participants in Group 5, during the Comparison phase. The opinions presented and the comments shown are the actual decisions and results of the 25 experts tested during pilot testing.

RESULTS

The impact of the variables were measured by: 1) the accuracy and reproducibility of the participants' decisions against the ground truth; 2) the accuracy and variation during feature selection (i.e. which ridge details were relied upon during the examination).

The following hypothesis was tested: *experts would benefit from the implementation of tools that objectively assess clarity of friction ridge features and provide statistics regarding the strength of a match.* This benefit would come in the form of reduced error rates, reduced variance during feature selection, and increased reproducibility of decisions among analysts.

Error Rates

Using approaches for calculating error rates as described by Koehler[7] or Aitken and Taroni[8], a testing matrix was produced from the results of the experiment. Table 1 shows the results for all 176 participants.

[7] Koehler, J. Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter. Hastings Law Journal 59 (5), 2008, 1077-1100.

[8] Aitken, C; Taroni, F. Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed. Wiley & Sons Ltd: West Sussex, England, 2005.

As discussed in Koehler[7], ***false discovery rates*** are very appropriate measures for conveying the reliability of a method. Furthermore, there is considerable debate on how to handle the inconclusive decisions. How should these decisions be counted when calculating false positive and false negative rates, and the sensitivity and selectivity of the method? For our purposes, we have not chosen to report inconclusives as errors per se, but still have considered them in the total number of trials for calculations. A major advantage to reporting false discovery rates is that the rate is not impacted by the number of inconclusive decisions.

	Ground Truth of Latent Print Source		
Analyst Decision	Source Present (S)	Source Not Present ($\bar{S}$)	Totals
"Identification" ("ID")	840	23	863
"Inconclusive" ("Inc")	322	92	414
"Exclusion" ("Exc")	70	765	835
Totals	1232	880	2112

Table 1: Testing matrix for all 176 participants who each completed all 12 trials.

From Table 1, the following standard testing terms were calculated:

Sensitivity = P["ID"|S] = 840/1232 = 68%
Selectivity* = P["Exc"| $\bar{S}$] = 765/880 = 87% *also referred to as Specificity

False + Rate = P["ID"| $\bar{S}$] = 23/880 = 2.6%
False – Rate = P["Exc"| S] = 70/1232 = 5.7%
Not detected rate = P["Inc"] = 414/2112 = 20%

False + Discovery Rate = P[$\bar{S}$|"ID"] = 23/863 = 2.7%
False – Discovery Rate = P[S|"Exc"] = 70/835 = 8.4%

To assess the impact of the tools, error rates for each group were calculated. These are displayed in Table 2. It can be seen that Group 2 and Group 5 had decreased errors, but not necessarily a statistically significant decrease (due to relatively low error rates to begin with). Both Groups 2 and 5 were the groups exposed to Expert Consensus tools.

In Table 3, the performance statistics and error rates were stratified according to expert status. Of the 176 participants, there were experts who were currently performing case work and reporting results (N = 87), experts who were certified through the International Association for Identification (N = 71), trainees who were not producing independent case work yet (N = 13), and finally participants (labeled "Other") who provide a number of technical duties such as 10-print examination, AFIS entry, basic case screening, and comparisons to elimination exemplars (N = 5). The false positive rate (erroneous identifications) and false positive discovery rate are significantly higher for the Trainee participants compared to the experts who are trained and reporting casework. These results mirror similar findings reported elsewhere[9]. It is interesting that conversely, the rate of false negatives (erroneous exclusions) and false negative discovery rates are lower for the Trainee. This is in accordance with the increased sensitivity and decreased selectivity (i.e. they are attempting more identification decisions, but attempting fewer exclusion decisions).

		False Discovery Rate		Error Rate			
	N	**False +**	**False -**	**False +**	**False -**	**Sensitivity**	**Selectivity**
Group 1	27	4.0%	10.1%	3.7%	6.3%	63.5%	79.3%
Group 2	33	0.6%	7.3%	0.6%	5.2%	66.7%	92.1%
Group 3	28	2.1%	7.6%	2.1%	5.1%	71.4%	86.4%
Group 4	31	4.6%	9.9%	4.5%	6.5%	66.8%	81.9%
Group 5	32	0.7%	7.8%	0.6%	5.8%	66.5%	95.6%
Group 6	25	4.3%	7.9%	4.8%	5.1%	75.4%	84.0%
All Groups	**176**	**2.7%**	**8.4%**	**2.6%**	**5.7%**	**68.2%**	**86.9%**

Table 2: Error rates and other performance statistics calculated for each experimental group. As a reminder: Group 1 (control group, no tool), Group 2 (Expert Consensus Minutiae map), Group 3 (LR tool), Group 4 (Quality Map tool), Group 5 (Expert Consensus Decisions table), Group 6 (Quality Map and LR tool).

		False Discovery Rate		Error Rate			
	N	**False +**	**False -**	**False +**	**False -**	**Sensitivity**	**Selectivity**
LPE	87	2.1%	10.7%	2.0%	7.4%	64.4%	86.8%
CLPE	71	2.3%	7.0%	2.3%	4.8%	69.6%	88.7%
Trainee	13	8.0%	6.0%	9.2%	3.3%	75.8%	72.3%
Other	5	0.0%	4.0%	0.0%	2.9%	77.1%	96.0%
All Groups	**176**	**2.7%**	**8.4%**	**2.6%**	**5.7%**	**68.2%**	**86.9%**

Table 3: Error rates and performance statistics separated by expertise status

Further examining how experience may affect error rates and performance, the case working experts and trainees were separated into different categories and the experts were stratified by years of experience analyzing and comparing latent prints on a routine basis. See Table 4. A clear trend was observed. False positive (erroneous identification) rates started quite high for trainees, but then significantly dropped (to almost zero) during the first couple of years of casework. Then as experience is gained, false positive rates increased to the highest levels for the most experienced analysts. This trend was mirrored in the sensitivity (i.e. they were attempting more identification decisions with more years of experience). A possible explanation for this trend is that experts become overly confident with time and their self-confidence induces them to "push the envelope". Similar trends have been reported in the medical

[9] Wertheim, K; Langenburg, G; Moenssens, A. A Report of Latent Print Examiner Accuracy During Comparison Training Exercises. *J of Forensic Identification* **2006**, 56 (1), 55-93.

community and other expert domains[10]. Conversely, false negative (erroneous exclusion) rates are higher in the less experienced group and are reduced in the most experienced group. Incidentally selectivity is increasing. Thus we can see that experts are becoming more efficient at excluding with more experience (i.e. they are attempting more exclusion decisions while simultaneously decreasing the false negative rate).

Years of		**False Discovery Rate**		**Error Rate**			
Experience	**N**	**False +**	**False -**	**False +**	**False -**	**Sensitivity**	**Selectivity**
Trainees	13	8.0%	6.0%	9.2%	3.3%	75.8%	72.3%
Experts							
<= 2	26	0.8%	12.6%	0.8%	8.8%	64.8%	85.4%
3 - 7	48	1.8%	7.9%	1.7%	5.7%	66.4%	91.7%
8 -15	49	3.0%	8.2%	2.9%	5.2%	66.5%	82.4%
16 - 30+	35	2.8%	7.5%	2.9%	5.3%	71.4%	92.0%
All Groups	**171**	**2.7%**	**8.4%**	**2.7%**	**5.8%**	**68.2%**	**86.9%**

Table 4: Error rates and performance statistics separated by years of experience

Other factors were examined as well for the effect on error rates. Some of these factors included the sex of the participant, whether the participant had taken any courses in statistics applied to forensic science, whether the participant traced ridges with the ridge tool, and whether the participant personally found the tools useful during trials. None of these factors appeared to significantly impact error rates and performance. One very interesting factor that did impact error rates was whether the participant documented and annotated at all. A handful of participants (N = 4)[11] chose not to annotate their features or document the ACE process at all. The false positive and false negative error rates in this group were 15.0% and 7.1%, respectively, compared to 2.7% and 5.8% reported by the majority of participants that documented. This is a significant increase in error and may indicate the value of careful documentation and exercise of caution/working slowly when dealing with complex cases. Incidentally those that did not document were all case working experts, ranging from 10 to 30 years of experience.

Reproducibility of Decisions

Reproducibility is defined as the ability of a test/method that when the same sample is provided to different instruments, how consistent are the results produced. If we look at the reproducibility of the conclusions for all groups, as shown in Figure 7, it can be seen that some trials were more reproducible than others. When the reproducibility of the decisions was compared by experimental group, Group 2 and Group 5 (the Expert Consensus tools) had the highest reproducibility (*statistics forthcoming*).

Ridge and Minutiae Annotations

Recall that users could annotate minutiae in the latent print during the Analysis phase without the presence of an exemplar, adjust their annotations in the latent print once they have viewed the exemplar, and lastly, annotate minutiae in the exemplar that correspond to features in the latent print. These data were labeled, respectively, A_{min}, O_{min}, and C_{min}. Additionally, participants could use a ridge marking tool and trace ridges. Ashbaugh routinely states the importance of analyzing and comparing ridge systems, while avoiding focusing solely on minutiae configurations. While several papers discuss the highly

[10] Generally in *The Cambridge Handbook of Expertise and Expert Performance*, Ericsson, KA; Charness, N; Feltovich, PJ; Hoffman, RR, Eds. Cambridge University Press, 2006.

[11] We must use extreme caution in interpreting these results given such a small sample.

discriminating value of minutiae configurations[12], there is most likely additional discriminative value with the inclusion of ridge counts and ridge tracings.

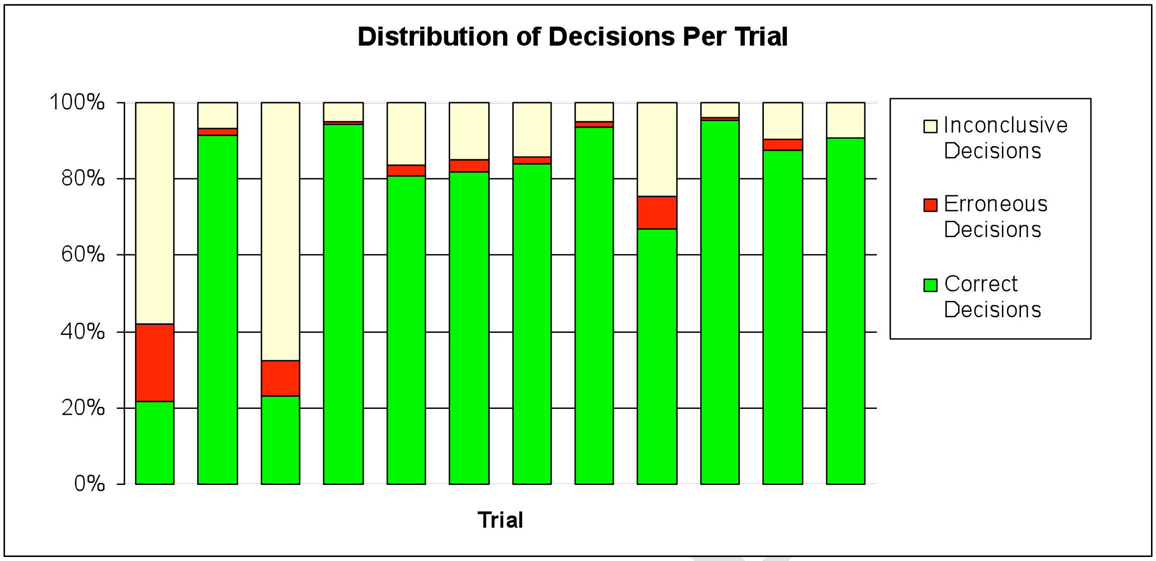


Figure 7: This graph shows the variation of reported decisions for each trial. All groups have been included. The first seven trials represent same source images, the last five trials are from different sources. This graph does not represent the order in which the trials were presented to participants.

With respect to the expert users that employed the ridge tool in the Analysis phase, in the trials where the ridge tool was used, a higher mean number of A_{min} were annotated by these users compared to the mean number of A_{min} for users that did not use the ridge tool (Mann-Whitney test, $p < 0.001$; mean (did not use tool) = 10.1 (3.4 S.D.), compared to mean (used tool) = 11.2 (3.6 S.D.). However use of this tool did not reduce the total number of erroneous decisions (those using the ridge tool had a 5% erroneous decision rate and those not using the ridge tool had a 4% erroneous decision rate). However, it must be clarified, that just because an analyst did not use the ridge tracing tool, it does not mean the analyst did not analyze and compare ridges, it simply means the analyst did not formally document it in PiAnoS.

As for the Analysis minutiae (A_{min}), there was no statistically significant differences for means, variance, or relative standard deviations (RSD) across experimental groups (Group 1 through 6), (Kruskall-Wallis test, p = 0.776; see Figure 8). ***Thus we conclude that the tools did not affect the number of minutiae selected during the Analysis phase.*** A similar trend was observed for analysis minutiae changed or added during the Comparison phase (O_{min}). Finally, for minutiae marked in the exemplar during the Comparison phase (C_{min}) there were no statistically significant differences, except for Group 5, which showed some statistically significant differences, compared to the control group (Group 1). However, it should be noted that given the instructions of the experiment, which did not instruct participants to necessarily focus on creating corresponding pairs of minutiae during the comparison phase, caution must be exercised when interpreting the C_{min}. It is quite likely, that if participants found few features in agreement, they may not have marked many minutiae in correspondence or few at all if the impressions were well below their level of sufficient ridge detail to effect a decision. Conversely, participants may have stopped annotating corresponding minutiae pairs once they achieved a threshold level to effect a

[12] Some examples include: Neumann, C; et al. Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae. *J of Forensic Sci* **2007**, 52 (1), 54-64; and Egli, N, et al. Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—Modelling within finger variability. *Forensic Sci Intl* **2007**, 167, 189-195.

decision. Had they been instructed to mark all pairs, regardless of whether the corresponding pairs fell well below or well above their decision threshold, the data may have been quite different at the extreme low and high ends.

Factors such as sex, expert status (i.e. trainee versus certified expert versus expert, etc.), having had a statistics course, and years of experience were looked as possible contributors to differences in minutiae selection. Only “years of latent print experience” showed some significant difference in means (Kruskal-Wallis, p = 0.073).

Not surprisingly, the largest driving factor that showed differences in mean number of minutiae (for A_{min}, C_{min}, and O_{min}), variance, and RSD was the latent print trial itself. In other words, as observed in other studies[13], the quantity and quality of ridge detail available in the latent print drove the minutiae selection process (Kruskal-Wallis, p << 0.001). In Figure 9, the boxplots showing the distribution of A_{min} across all trials, clearly demonstrate these significant differences.

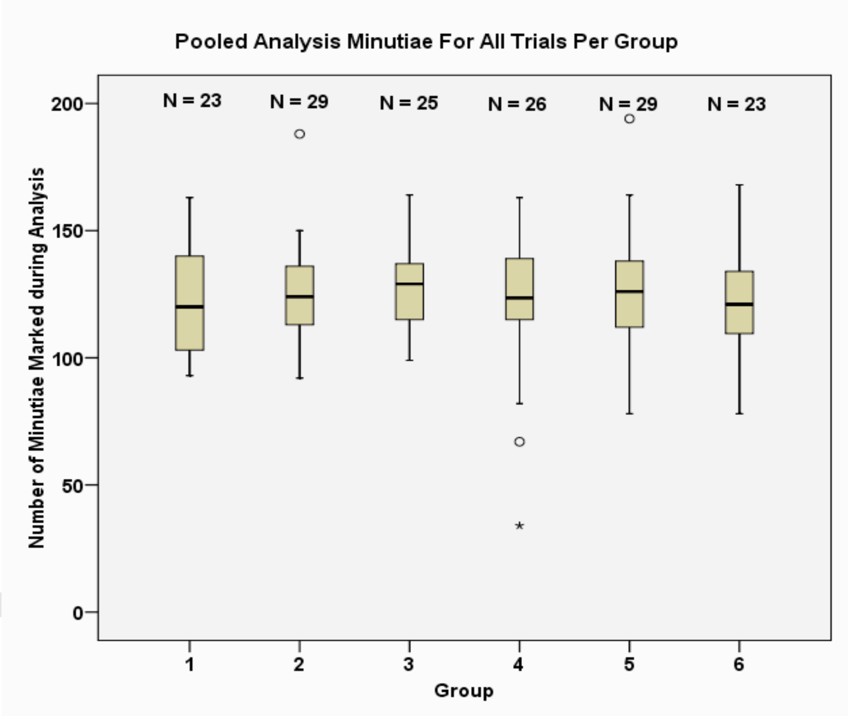


Figure 8: Bloxplots showing the total number of Analysis minutiae (A_{min}) pooled for all trials, across groups for experts only. There were no statistically significant differences for means, variance, and relative standard deviation (RSD) across groups. Note that Groups 2, 4, and 6 used minutiae maps to focus the analysis of the participants towards consensus minutiae.

Number of Minutiae and the Expert’s Decision

There was a statistically significant difference in the mean number of minutiae reported in each trial with respect to the decisions reported by the expert participants. Experts reporting a definitive conclusion (“Identification” or “Exclusion” decision) were more likely to have annotated a higher number of

[13] Dror, et. al. Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a ‘Target’ Comparison. *Submitted for publication*, 2010. Langenburg, G. Pilot Study: A Statistical Analysis of the ACE-V Methodology—Analysis Stage. *J of Forensic Identification* **2004**, 54(1), 54-79.

minutiae in the Analysis and Comparison phases (Kruskall-Wallis, $p < 0.001$). Table 5 provides the means and standard deviations for the minutiae annotated when an "Identification" decision was provided ("given an "ID" decision" or | "ID") versus when an "Inconclusive" decision was provided (| "Inc"). The data in Table 5 reflect only the trials where the images were in fact coming from the same source.

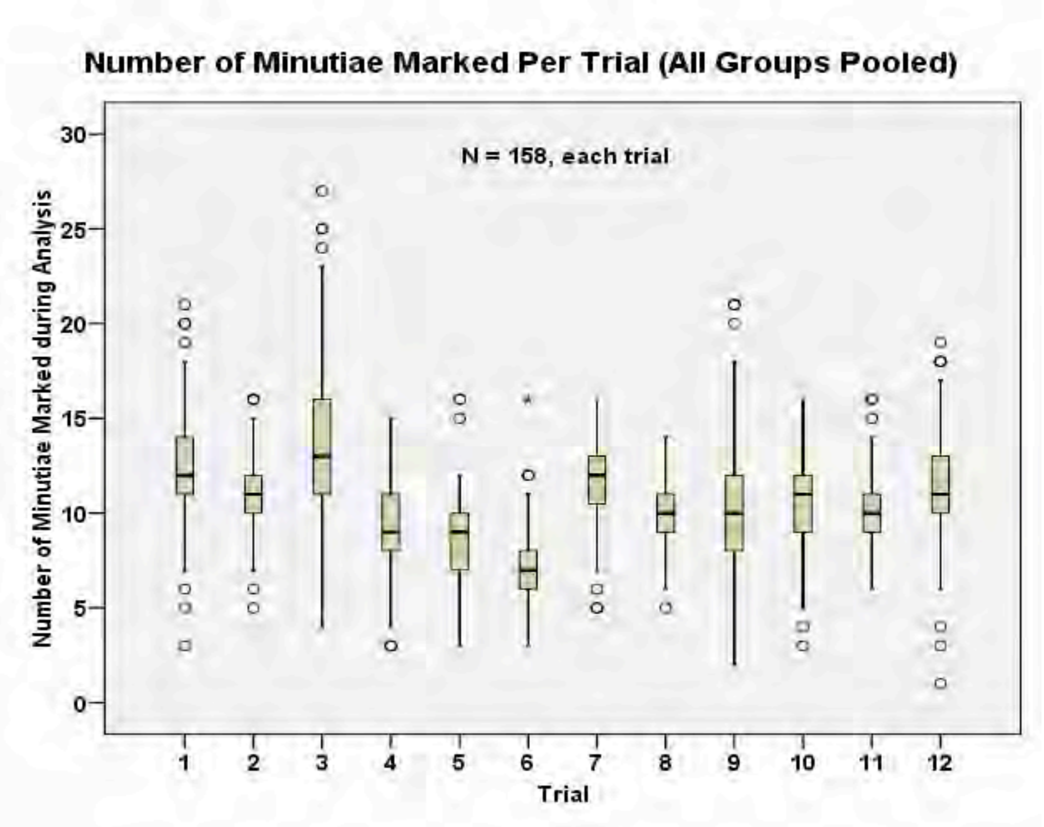


Figure 9: Boxplots showing the number of minutiae marked by experts (all groups combined) for the twelve trials. It is clear that some of the latent print images produced a larger variance of annotated minutiae than other images.

Number of Minutiae	Mean	Std. Dev.	N
A_{min} | "ID"	10.9	2.3	727
A_{min} | "Inc"	9.2	3.6	287
C_{min} | "ID"	11.6	4.2	657
C_{min} | "Inc"	8.9	4.8	233

Table 5: The mean and standard deviation for the number of minutiae annotated during the Analysis phase (without an exemplar present) and during the Comparison phase (annotated in the exemplar print and presumably corresponding to minutiae selected in the latent print). The data have been stratified according to the decision reported by the expert. The data above only represent the seven same source trials.

The distributions for these conditions (Pr[number of minutiae annotated | decision reported, when the images *are coming from the same source*) were plotted in Figure 10. From Figure 10, it can be observed that the likelihood ratio of the analyst reporting an "Identification" decision versus an "Inconclusive"[14] decision changes from less than 1 to greater than 1 approximately between 8 and 9 minutiae. Thus we can infer that, at least under the conditions of this study, experts had an *operational decision threshold* around 8 or 9 minutiae. While experts are trained to not solely base their decision on minutiae alone, we can still predict that above 8 or 9 minutiae in correspondence, experts will be more likely to report a positive attribution.

We did not examine the correlation of decision and number of minutiae reported in cases where the images came from different sources. The main reason for not doing so was that participants normally did not completely annotate or marked few minutiae in comparison after reaching an "Exclusion" decision. However, we did examine the 23 false positive cases and overlaid the results on the Figure 10 distributions. Remember, in these 23 cases, the participants believed the images were coming from the same source and therefore annotated what they believed to be valid correspondences[15]. Because of the relatively few data points (compared to the correct "Identification" decisions and "Inconclusive" decisions), the 6 trainee false positive errors were included in these data. The distribution of minutiae annotated in the Comparison phase for these erroneous decisions are shown in Figure 11. It is interesting to observe that the mean number of Comparison phase minutiae (mean = 9.0, SD = 3.5)[16] for the erroneous "Identification" decisions is found at the previously discussed *operational decision threshold*.

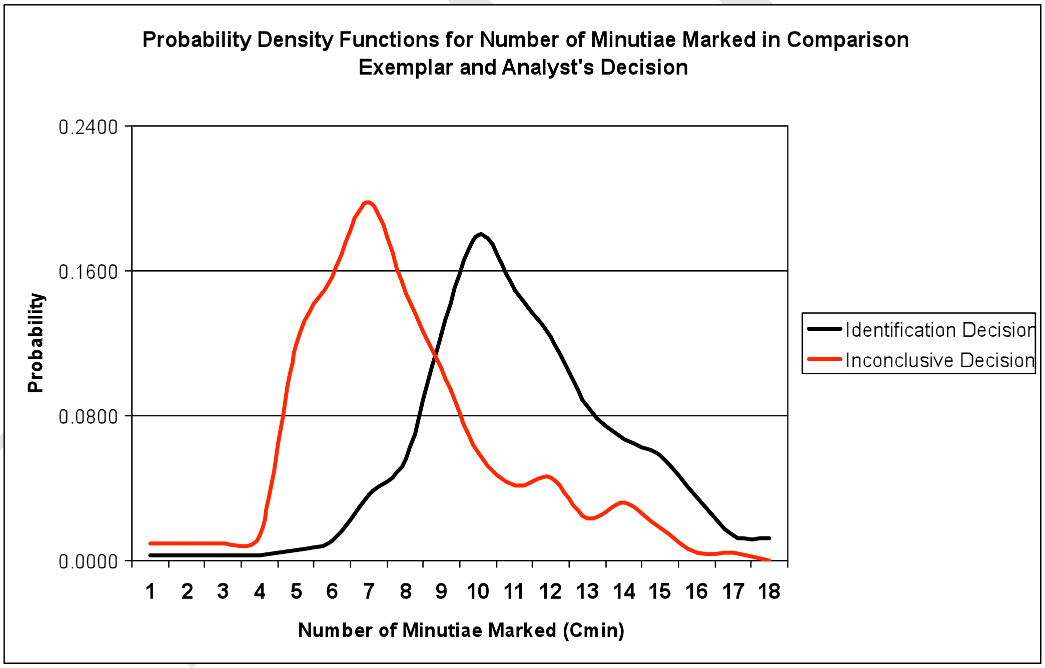


Figure 10: Probability distributions for the number of minutiae annotated in the exemplar during the Comparison phase in the seven same source trials. The distributions are separated according to the decision reported by the expert ("Inconclusive" versus "Identification").

[14] In some agencies, analysts would report "No value" or "not of value for identification purposes" instead.

[15] However, one of the 23 false positives was not annotated at all.

[16] Removing the 6 trainees, who generally had higher numbers of minutiae annotated, the mean and standard deviation change to 8.4 and 3.6, respectively.

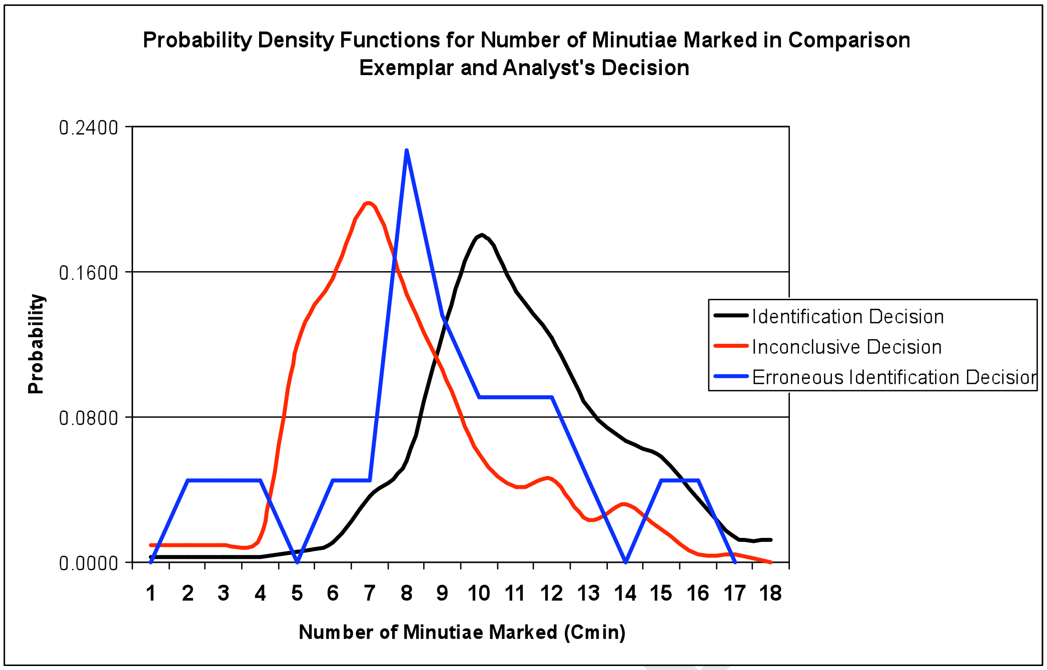


Figure 11: Distribution of minutiae marked in Comparison phase (C_{min}) when an erroneous "Identification" decision was made. Note that these data represent only 22 data points (including 6 data points from trainees) compared to hundreds of data points for the distributions of C_{min} for "Identification" and "Inconclusive" decisions.

Inter-Observer Variation

If two analysts each annotate a total of 11 minutiae in the same trial, it may be tempting to think these two analysts are in agreement. However, Figure 12 illustrates just how incorrect this notion is. The images in Figure 12 depict the Analysis phase annotations of two analysts in the same group (both were classified as experts). While both have 11 minutiae annotated, they only share 6 minutiae between them. The participant on the left has 5 different minutiae than the participant on the right and the participant on the right has 5 different minutiae than the participant on the left. Therefore, it can be said there is a difference of 10 minutiae between these two participants. This type of measurement is called a ***Euclidean Squared Distance*** (hereafter **ESD**). ESDs offer deeper insight into the inter-observer differences between analysts.

Pairwise comparisons can be performed for all pairs of participants in each experimental group and across all groups. Figure 13 is a histogram of the ESDs across all groups for all 158 expert participants (although not all 158 experts annotated minutiae in each trial). This generated ½ N x (N – 1) pairwise comparisons. For each group, this resulted in approximately 300-400 expert pairwise comparisons for each trial. Across all groups, for each trial, this resulted in approximately 11,000-12,000 pairwise comparisons.

Recall from Figure 9 that the mean number of A_{min} annotated by experts in Trial 1 was 12.2. If we divide the mean ESD (5.24) by the mean A_{min} (12.2), we see that the relative ratio of ESD difference to the average total number of minutiae marked is 0.43. In other words, on average, 43% of the experts' annotations differed between randomly selected pairs of experts. A summary of this ESD ratio for each trial, and other relevant statistics can be found in Table 6.

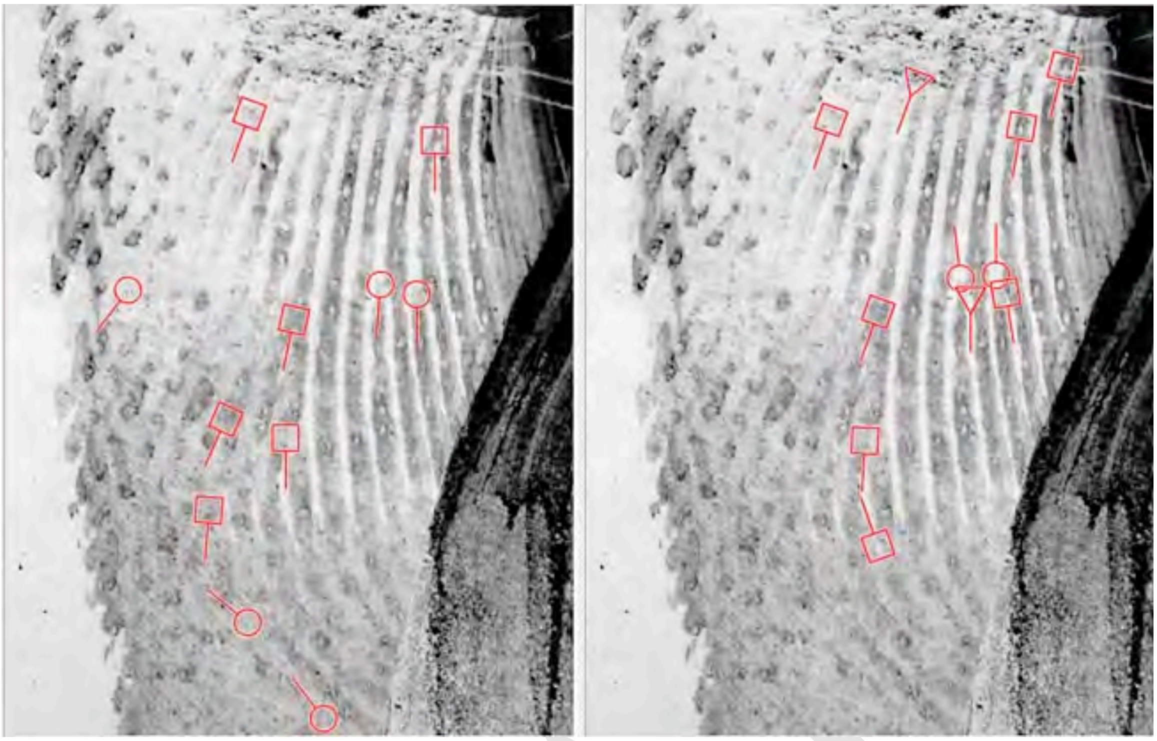


Figure 12: A screenshot of the Analysis minutiae annotations between two experts from the same group. While both participants have 11 minutiae, they share 6 minutiae in common and have 10 different minutiae between them.

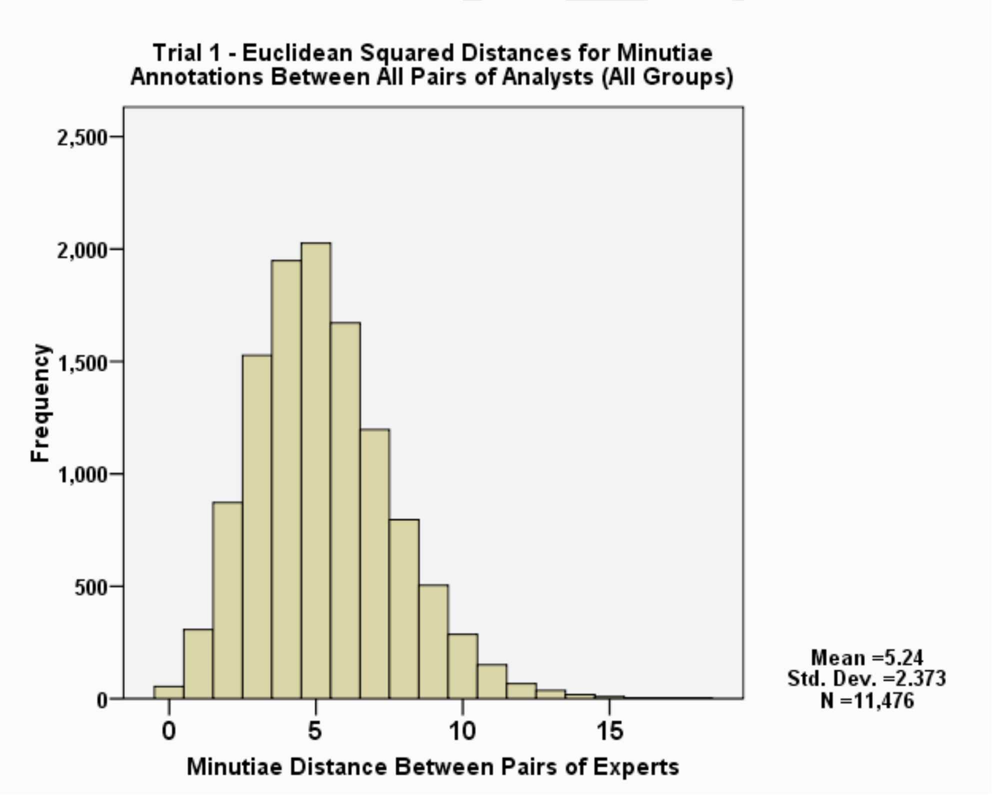


Figure 13: Euclidean Squared Distances (ESDs) for Analysis phase minutiae for all pairwise comparisons across all experimental groups for Trial 1.

It can be seen from Table 6, that for this study, the average difference between analysts, across trials was approximately 44% (SD = 9%). Therefore, on average, selecting any 2 expert analysts, one would expect approximately between 30%-60% of the minutiae annotated in the Analysis phase to differ between the analysts for the fingerprint comparisons in this experiment.

	Amin			ESD Statistics						
	mean	SD	% False Marked	mean	SD	N	min	med	max	ESD ratio
Trial 01	12.2	2.7	3%	5.2	2.4	11,476	0	5	18	0.43
Trial 02	10.9	2.0	3%	3.6	1.8	11,628	0	3	11	0.33
Trial 03	14.1	3.8	3%	7.9	3.6	11,781	0	7	27	0.56
Trial 04	9.2	2.2	4%	3.9	2.1	11,476	0	4	12	0.43
Trial 05	8.6	2.1	2%	4.5	2.2	11,781	0	4	18	0.52
Trial 06	7.1	2.0	12%	3.8	2.1	11,935	0	4	15	0.53
Trial 07	11.6	2.1	5%	4.0	2.0	11,781	0	4	12	0.34
Trial 08	10.0	1.6	3%	3.3	2.0	11,781	0	3	17	0.33
Trial 09	10.3	3.3	7%	5.7	2.7	11,781	0	5	19	0.55
Trial 10	10.6	2.5	11%	5.3	2.3	11,935	0	5	17	0.50
Trial 11	10.4	1.8	4%	3.3	2.1	11,781	0	3	18	0.32
Trial 12	11.1	2.8	2%	4.6	2.7	11,935	0	4	18	0.42
									Mean	0.44
									SD	0.09

Table 6: The left half of the table shows statistics for the number of minutiae marked during Analysis phase (A_{min}) for each trial (including the percentage of incorrectly marked minutiae according to the ground truth) by 158 experts. The right half of the table shows statistics for the Euclidean squared distances measuring the differences in A_{min} annotations between all expert pairwise comparisons.

Minutiae Map Effects

Participants in Groups 2, 4, and 6 were exposed to "minutiae suggestion" maps during the Analysis phase. Groups 4 and 6 had the quality map and Group 2 had the expert consensus map—although all three groups were suggested to look at the exact same minutiae. These groups received two different colored minutiae prompts: one color represented minutiae marked by 75% or more of the pre-tested experts and the other color represented minutiae marked by 50% or more of the pre-tested experts[17]. We investigated whether these groups had an increase in the number of analysts marking the minutiae that corresponding to the "suggested minutiae". No statistically significant increase was noted in any group for the 75+% minutiae. Of course, this is somewhat understandable since these were minutiae that most experts observed naturally, without prompting, during the pre-testing. However, for the 50+% minutiae, a statistically significant increase was observed in all three of the "suggested minutiae" map groups (Chi-square test for all 12 trials compared against control group, 11 d.f., Chi-square stat = 98, 43, 54, p values = all <0.001). Group 3 did not show a significant increase in the number of 50+% minutiae. Group 5, unexplainably, did.[18] It is unknown why this group would have had such a significant increase in the number of 50+% minutiae for one of the twelve trials. In summary, the minutiae maps appeared to help

[17] In Group 2, these colors corresponded to green (75+%) and yellow (50+%). In Groups 4 and 6, these colors corresponded to black (75+%) and white (50+%).

[18] A possible explanation comes from an outlier value. One of the Group 5 trials had one outlier value that made the p-value significant. (Chi-square stat = 27, p = 0.004). Over half of this statistic came from the contribution of Trial 1 where inexplicably a large number of 50+% minutiae were marked. Similar values weren't seen in the other trials for this group.

direct the attention of the analyst to minutiae that they might not have normally noticed or annotated. ***This resulted in greater consistency and conformity in their annotations in those groups.***

Furthermore this effect can be observed directly in the ESD ratio values when the data in Table 6 are stratified according to experimental group. Table 7 shows the ESD ratios for each group per trial. The ESD differences are significantly lower[19] for all three groups which had a minutiae suggestion map (Groups 2, 4, and 6). The ESD differences are slightly lower for Groups 3 and 5. It is unknown why these groups would have been slightly lower, since they received the latent print images exactly as Group 1 (control) during the Analysis phase. Perhaps because they were told they would be viewing a new tool in the Comparison phase, this created some "conservative minutiae selection" experimental bias.

From Table 7 we can conclude that the Expert Minutiae Consensus map in Group 2 had the strongest effect of reducing analyst variation in minutiae selection/annotation. Groups 4 and 6 had comparable reductions as well. Figure 14 provides a snapshot of the effect. The decrease in ESD ratio is shown in Figure 14, by setting Groups 2 through 6 relative to Group 1. The groups have been reordered by increasing effect.

	Group 1 Ratio	Group 2 Ratio	Group 3 Ratio	Group 4 Ratio	Group 5 Ratio	Group 6 Ratio
Trial 01	0.51	0.36	0.47	0.40	0.43	0.46
Trial 02	0.37	0.30	0.32	0.31	0.34	0.29
Trial 03	0.65	0.50	0.53	0.49	0.65	0.53
Trial 04	0.42	0.36	0.44	0.40	0.47	0.49
Trial 05	0.58	0.41	0.54	0.49	0.56	0.50
Trial 06	0.60	0.46	0.54	0.44	0.66	0.47
Trial 07	0.39	0.30	0.37	0.31	0.35	0.31
Trial 08	0.41	0.30	0.32	0.29	0.34	0.26
Trial 09	0.63	0.56	0.55	0.50	0.57	0.46
Trial 10	0.56	0.44	0.53	0.44	0.52	0.45
Trial 11	0.40	0.32	0.30	0.28	0.31	0.29
Trial 12	0.52	0.36	0.42	0.40	0.41	0.36
Mean	0.50	0.39	0.44	0.40	0.47	0.41
SD	0.10	0.09	0.10	0.08	0.12	0.10

Table 7: Euclidean squared distances (ESD) ratios (ESD/Mean A_{min}*) for each group per trial.*

When the false minutiae[20] annotation error rates were stratified by group, a similar trend as that in Figure 14 was observed. Groups 6, 4, and 2 all showed a reduction in the error rate for false minutiae annotations (see Figure 15). The groups have been ordered from highest error rate to lowest. The order is identical to the order of the ESD ratios.

[19] A Mann-Whitney test found the distributions of ESD ratios for Groups 1, 3, and 5 to be statistically significantly different from Groups 2, 4, and 6 (p = 0.003).

[20] False minutiae annotations are defined in this study as the presence of an annotated feature where none existed according to the ground truth exemplar. These were counted manually and we used the following guideline for determining a false minutiae: we gave a "one ridge leeway" in their markings (if they within a ridge of a true feature we counted it as accurate) and incipient ridges and dots were not scored as false minutiae). Accuracy of minutiae type was not considered in these determinations.

Finally, when we attempted to tie accuracy and consistency of A_{min} to decision error rates, we did not see this trend. In Figure 16, the groups were ordered from highest to lowest false positive discovery rates (for reference and completeness, false negative discovery rates were superimposed on the chart). Compared to Figures 14 and 15, the order of the groups has changed, although Group 2 still has the lowest error rate.

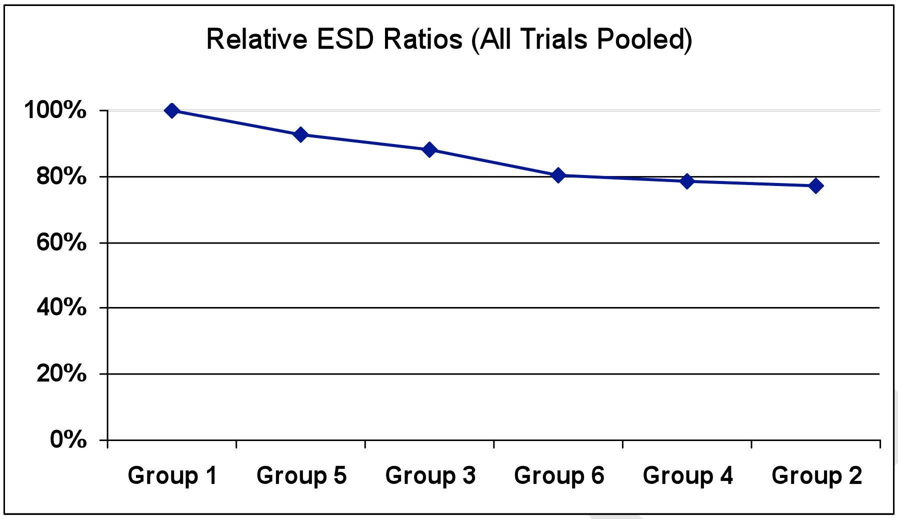


Figure 14: Relative (to Group 1, control) ESD ratio, showing a decreasing effect for groups with a minutiae suggestion map. Here, a decreasing effect represents less differences in annotations between paired experts during the Analysis phase.

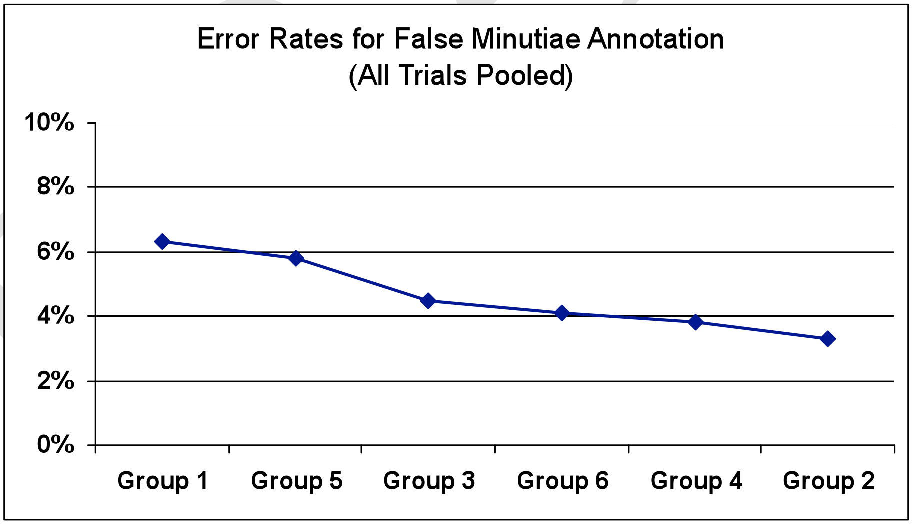


Figure 15: The percentage of Analysis phase minutiae annotations (for all trials pooled) that reflected false minutiae (i.e. they did not exist in the ground truth exemplar).

If the results of Group 2 were a direct result from using the Expert Consensus Minutiae map, then this suggests that observing the markings of other analysts was very useful information. This group exhibited the most consistency in minutiae selection, the lowest false minutiae error rates, and the lowest false discovery rates. Group 5, the Expert Consensus Decision table, was comparable in false discovery rates and also resulted in fewer inconclusive decisions and more definitive decisions (without a corresponding increase in error rates). However Group 5 did not show the consistency and accuracy of minutiae selection during the Analysis phase. Ideally combining the consistency and accuracy of minutiae selection in Group 2 and the utility and accuracy of Group 5 decisions would appear to be an improvement to expert performance. Additional testing could be done here with a group that receives the Expert Consensus Minutiae map during the Analysis phase, followed by the Expert Consensus Decision table during the Comparison phase. Those results could be compared against a control group that receives the same images without that information.

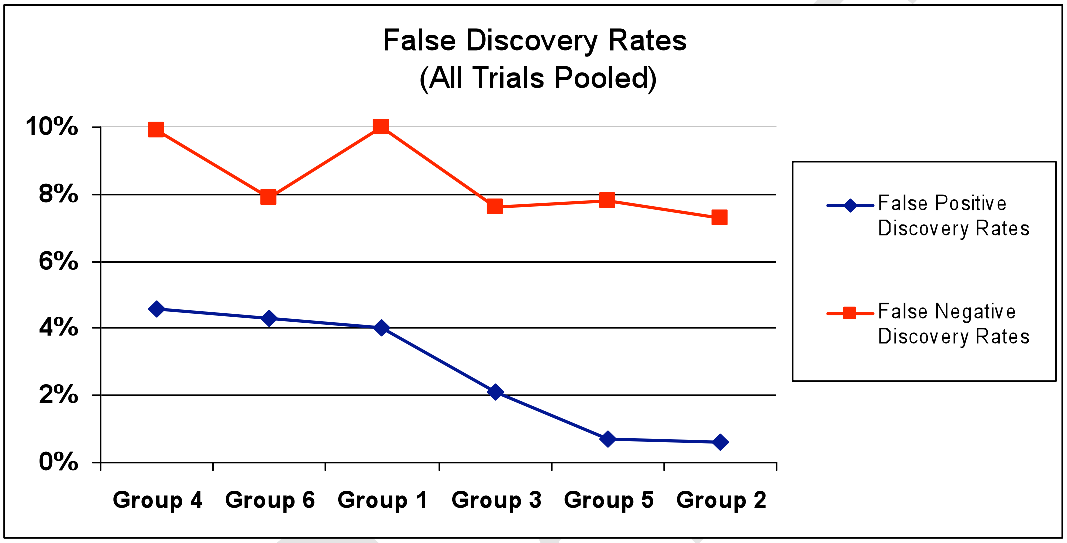


Figure 16: False discovery error rates (i.e. Pr [Not Source] | "Identification" decision by analyst] ordered according to descending rates and groups.

SUMMARY OF MAJOR CONCLUSIONS

Further research is warranted in this area. It appears from the results of this study, experts can improve performance, as measured by the consistency and accuracy of minutiae selection and the reproducibility and accuracy of decisions. The tools that showed the greatest effect came from providing the results of other experts that had previously examined the fingerprint trials using PiAnoS software. If experts could be trained to be more consistent in their feature selection, this may result in more reproducible decisions. Furthermore, based on data here, it would appear that operational guidelines for minutiae threshold decisions could be provided, but that these guidelines would need to account for specificity and clarity of minutiae, in addition to inclusion of non-minutiae, discriminatory friction ridge features (e.g. scars, creases).

The following bullet points are the major conclusions that can be derived from the results of this experiment.

- Error rates, false discovery rates, and performance measurements (i.e. sensitivity, selectivity, rate of detection, etc.) were calculated for all participants (N =176). Overall, given the difficulty of the cases, the false positive and false negative error rates were quite low (2.6% and 5.7%, respectively).
- Error rates were reduced in Group 2 (received the consensus minutiae map) and Group 5 (received the consensus decision tables). These tools were generated from other analysts previously viewing these cases. This evidence suggests that the natural collaboration and exchange of information between fingerprint experts is beneficial, if properly balanced against potential dangers of bias.
- False positive error rates for trainees were significantly higher than the rates for experts.
- The highest false positive error rates (15.0%) were observed in a small subset of experts (N = 4) that did not annotate their ridge features or document the ACE process.
- The latent print image (i.e. the available quantity and quality of ridge detail) itself was clearly impacting the mean number and variance of minutiae selected.
- Experts that analyzed and compared a higher number of minutiae (approximately above 8 or 9 minutiae) were more likely to report a positive source attribution (when the images were coming from the same source).
- Experts in the groups that were presented with minutiae suggestion maps (Expert Consensus Minutiae map in Group 2 and Quality Map in Groups 4 and 6) showed less variation than experts in the other groups with respect *to which minutiae were annotated* during the Analysis phase.
- These tools did not appear to significantly impact the mean or variance of the number of minutiae annotated, but the Group 2, 4, and 6 tools did appear to impact the percentage of false minutiae marked (i.e. the accuracy of the selected minutiae).
- Group 2 (Expert Consensus Minutiae map) reported the most consensus and accuracy of minutiae annotations and the lowest false discovery rates. We cannot say if the more accurate and consistent minutiae selection during Analysis phase led directly to lower false discovery rates, but it is a possibility. Group 5 (Expert Consensus Decision table) exhibited low false discovery rates too, but did not have a tool during the Analysis phase.

For additional information, please contact:
Glenn Langenburg
Minnesota Bureau of Criminal Apprehension
1430 Maryland Avenue East
Saint Paul, MN 55106
(651) 793-2967
glenn.langenburg@state.mn.us

We do intend to submit these findings for publication. Please use caution when citing and distributing, since these results have not yet been formally peer reviewed and published. This study was funded by a grant from the Midwest Forensic Resource Center (MFRC), Iowa State, Ames, Iowa (SC-10-339).