Journal of the American Society of Questioned Document Examiners, Inc.
Volume 23, Number 1, June 2020

Cognitive Human Factors and Forensic Document Examiner Methods and Procedures

Mara L. Merlino, Ph.D.[a], Veronica B. Dahir, Ph.D.[b], Charles P. Edwards, M.A.[c], Derek L. Hammond[d], Tierra Freeman-Taylor, Ph.D.[e], Adrian Dyer, Ph.D.[f], and Bryan J. Found, Ph.D.[g]*

[a]Kentucky State University
[b]University of Nevada, Reno
[c]University of Nevada, Reno
[d]U.S. Army Defense Forensic Science Laboratory
[e]Kentucky State University
[f]Royal Melbourne Institute of Technology
[g]Victoria Police Forensic Services Department
* Deceased.

Here we report initial findings from an interdisciplinary study to empirically explore the reliability, measurement validity, and accuracy of established FDE methods and procedures, and to investigate the influence of possible sources of cognitive bias in the methods and procedures of forensic handwriting examination. This article reports findings of our analysis of the relationship between the position of the known signatures and the utilization of writing features in questioned/known signature comparison tasks. Forty-nine professional forensic document examiners from government labs and private participated in an eye-tracking experiment to investigate patterns in their visual behavior as they performed a series of handwriting comparisons. Eye-tracking results revealed a left-to-right and proximal-to-distal pattern of gaze behavior. Significant differences in mean gaze fixation counts were found according to whether the signatures were text-based or stylized; whether the signatures were high or low complexity; whether the signatures were genuine or non-genuine (ground truth); and whether FDE decisions were accurate or misleading. Binomial logistic regression analyses revealed that ground truth and utilization of the upper left known signature were significant predictors of FDE decision accuracy.

Introduction

The extensive scrutiny of the methods and findings of numerous areas of expert testimony has prompted acrimonious debate among academicians, forensic practitioners, and legal professionals concerning what has been referred to by the Forensic Science Committee of the National Academy of Sciences ("Committee") as "faulty forensic science analyses" [1, p. 4]. While acknowledging the importance and utility of the forensic disciplines, the Committee also addressed the perceived flaws in such evidence. For example, advances in technology in various forensic disciplines, especially in the field of DNA testing, show that erroneous or misleading forensic evidence has contributed to the wrongful conviction of innocent individuals [1]. The Committee called for improvements in forensic science practices, arguing that increased and demonstrated reliability and validity in forensics will help law enforcement investigations by improving the reliability of identifications. Additionally, homeland security efforts will also improve as advancements are made in the methods and procedures of the forensic disciplines [1].

The Committee specifically identified several important issues, including practitioner certification, accreditation, and the availability of skilled, well-trained personnel [1]. Many areas of forensic science lack of uniformity in training, accredita-

This research was supported by NIJ Award No. 2010-DN-BX-K271, 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 publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

tion, and practice standards. The report stated that operational principles and procedures for many disciplines are not standardized between or within jurisdictions, attempts at standardization are not viewed favorably in many instances, and that even protocols such as Scientific Working Group standards "often are vague and not enforced in any meaningful way…These shortcomings obviously pose a continuing and serious threat to the quality and credibility of forensic science practice" [1, p. 6].

The Committee also discussed the lack of demonstrated validity and reliability within the interpretation-based disciplines, stating "…no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source…The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity. This is a serious problem. Although research has been done in some disciplines, *there is a notable dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods"* [emphasis added] [1, p. 7-8].

The 2016 report of the President's Council of Advisors .on Science and Technology titled *Report to the President on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*[1] similarly identified what the Council concluded were two important research gaps: A need for clarity about the standards for reliability and validity in the forensic sciences, and the need for ongoing studies to establish the extent of the validity and reliability of the forensic sciences.

The challenges of implementing common standards for reliability and validity across forensic fields have been illustrated by the ongoing debates about which agencies or organizations should make these determinations. According to Dror, empirical observation is neither strictly objective nor subjective, but instead exists on a continuum along which most forensic inquiries fall [2]. Consequently, creating a dichotomy for fields of inquiry based on the extremes of pure objectivity and pure opinion is neither accurate nor fruitful. Dror stated that even subjective, experience-based expert opinion can be accurate and valuable, despite the possibility that such expert opinion may be at increased risk for error, bias, and contextual influences. He wrote "even with quantification and statistical tools, the human element still plays a critical role, and therefore cognitive issues may continue to play an important role even in the less subjective domains of forensic science" [2, p. 81].

Current Approaches to Understanding the Examination Process

Identifying and mitigating factors that may contribute to misleading conclusions has long been a matter of concern for many forensic fields. Forensic science is produced and consumed in the context of various systems developed by human actors. Recent approaches to understanding the impact of potentially biasing information on forensic conclusions have embraced a human factors perspective. The International Ergonomics Association defines human factors as "the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance" [3]. Cognitive ergonomics is a subfield of human factors in which cognitive processes such as memory, perception, reasoning, decision making, skilled performance, and human reliability are examined in the context of work and operational settings. The goal of cognitive ergonomics is to improve task performance by the systematic study of the interaction between human cognitive functioning and the systems or environments in which tasks are performed.

A human factors approach to understanding issues of reliability, validity, proficiency, expertise, and sources of bias involves an examination of sources of information that are both internal and external to the forensic examiner. It is therefore important to address the production of forensic science from multiple perspectives. For example, the report of the Expert Working Group on Human Factors in Latent Print Analysis presented a taxonomy of errors and human factors which identified multiple potential sources of error [4].

Recognizing that science is produced by humans who are acting within various systems

[1] https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science

of professional roles and standards, organizations such as the National Institute of Justice, the Department of Justice, the National Science Foundation, and the National Institute of Standards and Technology (NIST, see Organization for Scientific Area Committees, or OSAC)[2] have launched programs and initiatives intended to bring together stakeholders to address these issues using an interdisciplinary approach.

In February 2020, NIST published an extensive report prepared by the Expert Working Group for Human Factors in Handwriting Examination, titled *Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach* [5]. Among other recommendations, the report endorsed ongoing interdisciplinary research encompassing expertise from forensic practice, social and cognitive psychology, vision science, and other areas to establish the basis and extent of forensic document examiner (FDE) expertise, and to develop empirically validated and rigorous document examination protocols, measures, and education and training programs that consistently and comprehensively address the knowledge and skills required to establish expertise in forensic fields.

Dyer, Found, and Rogers utilized eye tracking to study visual attention given to signatures by FDEs and a control group, noting eye movement, response time, and opinions. They found that FDE opinions were significantly more accurate than the control group [6]. However, both FDEs and control participants viewed features of signatures similarly, leading the researchers to suggest that differences in accuracy may be due to different cognitive processes for approaching the task. Specifically, improved accuracy by FDEs involved a capacity to process elemental pieces of information into the context of the total or global information inspected in a signature; whilst well motivated control participants more frequently made decisions only based on elemental visual evidence as they terminated decisions when inspecting one potential error. Thus, FDEs showed initial evidence of processing specific evidence in context, which was also supported by control experiments where global information presented tachistoscopically for brief periods that prevented eye movements did permit correct classification of signatures, but accuracy significantly improved when detailed feature information was enabled by permitting the FDEs to move their eyes to inspect detail [6]. A subsequent forensic eye tracking study also showed a capacity of FDEs to reliably discriminate between forged and disguised or simulated signatures [7].

1. Using the eye-tracking method employed by Dyer and colleagues [6] in their seminal research on visual attention in document examination, our studies further explored how FDEs extract information from handwritten signatures. Our study specifically investigated the relationship between the arrangement of the questioned and known signature specimens, and the utilization of writing features in comparisons of questioned and known signatures.
2. The placement of the signatures during the examination created a visual context in which the FDE conducts the comparison, as demonstrated in Figure 1.

We were specifically interested in understanding how the position of the known signatures in relation to the questioned signatures (context effects) were related to the outcome of FDE decisions about whether the questioned signatures were genuine or non-genuine.

Methods

FDE Sampling, Participant Recruiting, and Data Collection Procedures

Forty-nine professional document examiners working in government laboratories or private practice participated in the eye-tracking study.[3] The target number of participants was determined by consideration of several factors, including how many participants would be needed to achieve sufficient power for statistical analyses.[4] We recruited participants by attending various professional meetings and presenting information about the research, and we also utilized snowball sampling whereby examiners who participated

[2] http://www.nist.gov/forensics/osac/index.cfm
[3] The data for one participant had to be excluded from the study due to a research protocol violation.
[4] Power analysis using techniques described by Cohen and Cohen (1983) were used to determine the number of observations needed per variable (120) to detect a medium effect size of 0.60 at a power of 0.80. The number of observations in each comparison reported here far exceeds the number required to reach this level of power.

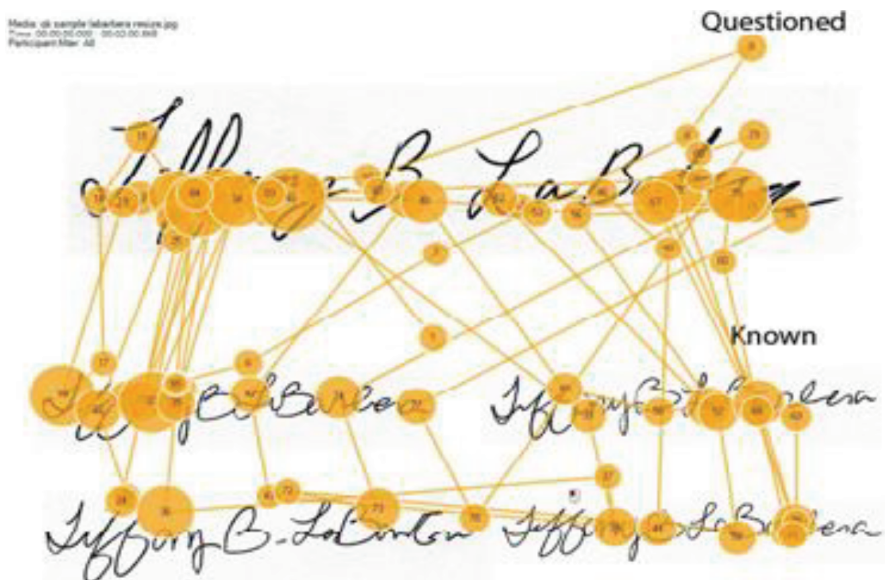

Figure 1. This "gaze plot" of a recorded comparison shows an example of a visual context effect. Note that the position of the two known signatures beneath the first name (top and bottom left [AGD1]) shows that this FDE fixated to a greater extent on those two known signatures, while the positioning of the other two known signatures under "LaBarbera" (top and bottom right) attracted a greater number of fixations on those known signatures.

recommended other members of the field whom they felt would also be willing to participate.

Materials and Equipment

Eye tracking devices track and record eye movements as people interact with the physical environment. Information about eye movements and gaze behavior allows researchers to answer questions about how people deploy their attention. For example, eye tracking can inform researchers about how people interact with web pages, with images, and with visual environments in a variety of situations [8, 9].

Eye trackers consist of three principal components. The *illuminators* (light source) direct near-infra red light to the pupil and cornea. The light creates highly visible reflections which are captured by the second component, the *camera*. The final component, the *image processing algorithms*, use the angle between the reflections of the cornea and pupil to calculate a vector. The vector and other features of the gaze reflections are then used to calculate the gaze direction. In this way the eye tracking unit can record what the participant is looking at, how long the participant's gaze rests in a particular area, the "scan path," or the sequence of positions where the participant's gaze rests ("gaze fixations"), and how long the fixation lasts. The raw fixation data is then filtered to display those areas where the eye position remains within a 50-pixel area for a time of greater than 100 MS. All gaze data were collected using Tobii® **model T-60** binocular eye tracking systems with 17″ thin film transistor (TFT) screens similar to the technology used in cellular phones, and 1280 x 1024 pixel displays (Tobii® Technology, Stockholm, Sweden), and Tobii Studio© software version 2.3.2.

Signature Stimuli

The signature stimuli were prepared to capture several different signature features that might be encountered as part of the FDE caseload. An objective classification system was utilized by the two FDE subject matter experts on our research team based upon the number of allographs (i.e., a letter of an alphabet in a specific shape, such as capital or lower case, italic, or various handwritten forms of the letter that fall within that letter template) that are present and legible within the signature [10, 11]. This classification scheme identified three types of signature: (1) text-based, in which each allograph of the name is clearly written; (2) mixed, in which two or more allographs can be read; and (3) stylized, in which one or fewer allographs are legible.

Our FDE subject matter experts determined signature complexity using a theory developed by Found and Rogers [12, 13]. Complexity theory states that as complexity increases, (as referenced by an increase in the number of features within the writing) the likelihood of someone being able to successfully simulate an image decreases [12, 13, 14]. Mohammed provides a detailed description of this method of determining the complexity of a signature, which is based on the evaluation of the number of turning points, line intersections, and retrace strokes in a signature [15].[5]

Signature Collection Procedure

Fifty signature writers participated in the signature collection process. All signature writers were unpaid volunteers who were over the age of 18 years. All volunteers were recruited by members of the research team from among friends, neighbors, colleagues, or students. None of the volunteers were police officers or investigators, and none of the volunteers had forensic training.

Writers were seated at a desk or table and provided with pen, paper, and backing sheets. Writers were then given four sheets labeled "Normal Signatures" and instructed to write their signature on the page using their normal signature style, as if they were signing a check or signing their name to routine office or schoolwork. Four signatures were collected per page, for a total of 16 genuine signatures. Writers were given one additional sheet of paper labeled "Model Signatures for Forgers," on which they used the same procedure to produce three model signatures to be used by other writers to create forged specimens.

Simulated signatures were then created by asking each writer except the writer who produced the original signature to simulate the genuine signature of a previous writer. Writers were not allowed to trace any of the known signatures, and that they were not allowed to practice simulating any of the signatures. The 49 simulators each produced three simulations using the genuine signatures provided as a model/guide.

The FDE subject matter experts then used an online random sequence generator[6] to selected 19 simulated and 11 genuine signatures. They classified the signatures as text-based, mixed, or stylized, and calculated the complexity of each writer's signature based on the classification model equations described above. Of these, they selected six signature specimens for each of the 11 selected signature writers, for a total of 66 questioned/known comparisons (22 genuine signatures, 9 disguised signatures, 22 simulated signatures, and 13 traced signatures). Sixty-two of the questioned signatures were randomly selected. The remaining four signatures were genuine writing specifically selected because of the presence of a feature or characteristic variation produced only in one genuine signature specimen produced by the writer (rare accidental or extraneous features). Figure 2 demonstrates an example of an "accidental" characteristic.

Procedure

The protocol consisted of 11 trials, each containing six questioned/known signature comparisons. The eye-tracking system was calibrated to each participant using a 9-point calibration grid. Each participant completed a practice trial for the protocol which included an additional calibration check. The material for each signature comparison was presented in the sequence demonstrated in Figure 3.

A fixation cross displayed for three seconds, then automatically changed to the "known signature" screen. The known signature screen displayed four examples of the signature writer's true, naturally written signature in the lower half of the screen. FDEs studied the four known signatures before conducting their comparison of the questioned and known signatures. No time limit was given for either their examination of the known signatures or the questioned/known comparisons so that accurate responses were enabled [6]. When the FDEs felt that they were ready to view the next stimulus they indicated this verbally to the researcher, who immediately exited to the next screen. Figure 4 illustrates the position of the FDE in relation to the eye-tracking equipment.

After each comparison we asked participants for a decision about whether the questioned signature was genuine, disguised, or simulated, then asked for their degree of certainty using the 9-point scale. If the participant responded that the signature was simulated, we asked them to indicate whether they thought that the simulation was a freehand copy or a tracing, and how

[5] Turning points are defined as the number of direction changes and the number of starting points and terminating points of any continuous line [16, 17]. Line intersections or retraced strokes are determined by counting the number of times a line either intersects or retraces a previously written stroke.

[6] www.random.org

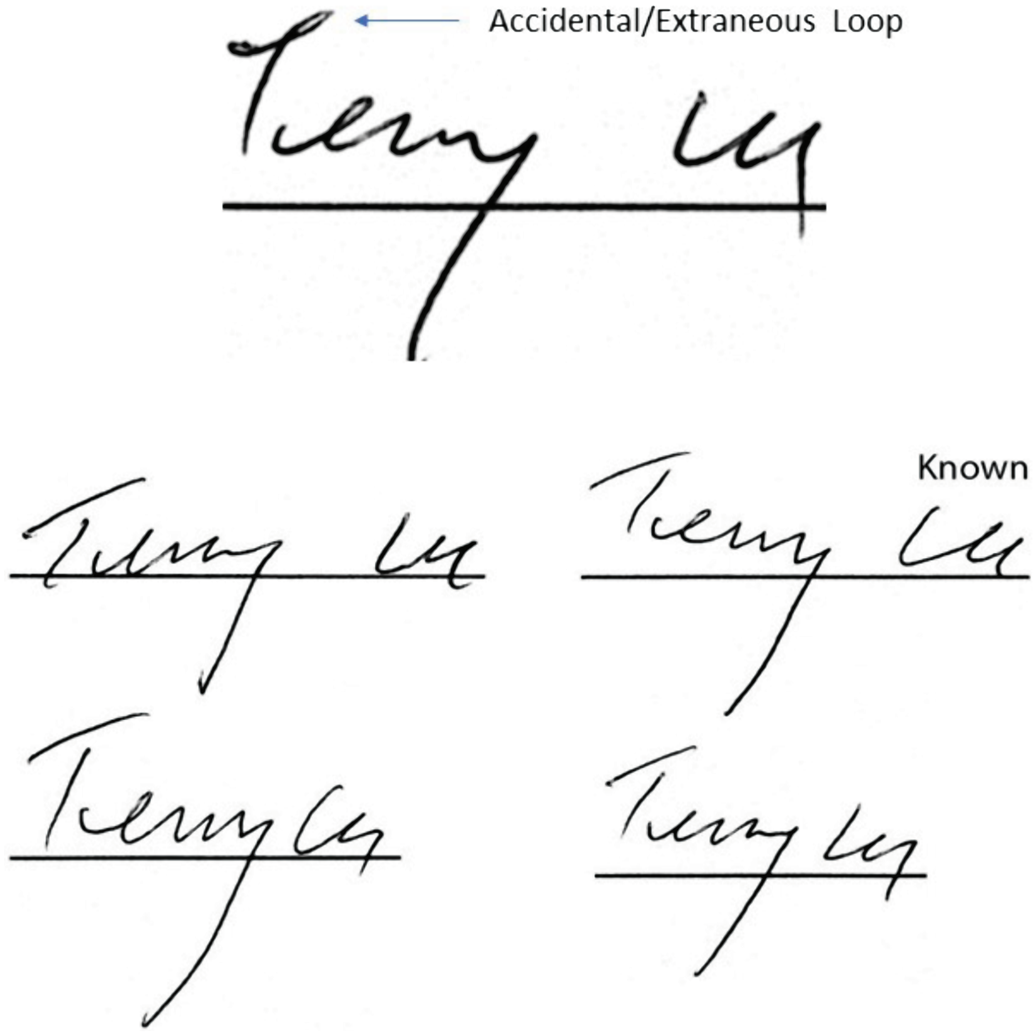

Figure 2. Genuine Terry Lu signature with "eyelet" in the T, which appeared only in this genuine questioned signature. The T in the specimen on the right is a two-stroke letter form in which the staff and the crossbar are not connected.

confident they were about this opinion on a 4-point Likert-type scale ranging from Not at all Confident to Extremely Confident. All responses were manually recorded by the researcher.

Results

These analyses are based on data from 36 signature comparisons of genuine and simulated signatures, conducted by 48 professional FDEs employed in private practice or government laboratories. Results represent 1,742 comparison decision observations. Of these, 723 (42.5%) observations are for genuine signatures and 1,019 (58.5%) are for simulated signatures. Of the 1,742 observations, 1,065 (61.1%) are for text-based signatures and 677 (38.9%) are for stylized signatures. A total of 729 (41.8%) observations are for high complexity signatures, and 1,013 (58.2%) are for low complexity signatures.

Although many metrics are available from our eye-tracking data, we focus here on the total fixation count (FC) metric for the genuine and simulated signature comparisons. A fixation is defined by the speed of the eye movement and the distance between adjacent data points. This metric measures the number of times the participant fixates in an "area of interest" (AOI).

AOIs for this study were created empirically using two visualizations of the cumulative gaze data for all participants. Figure 5 shows examples of a signature comparison stimulus, a gaze plot, a "heat map," and AOIs created from the information from the two data visualizations. The gaze

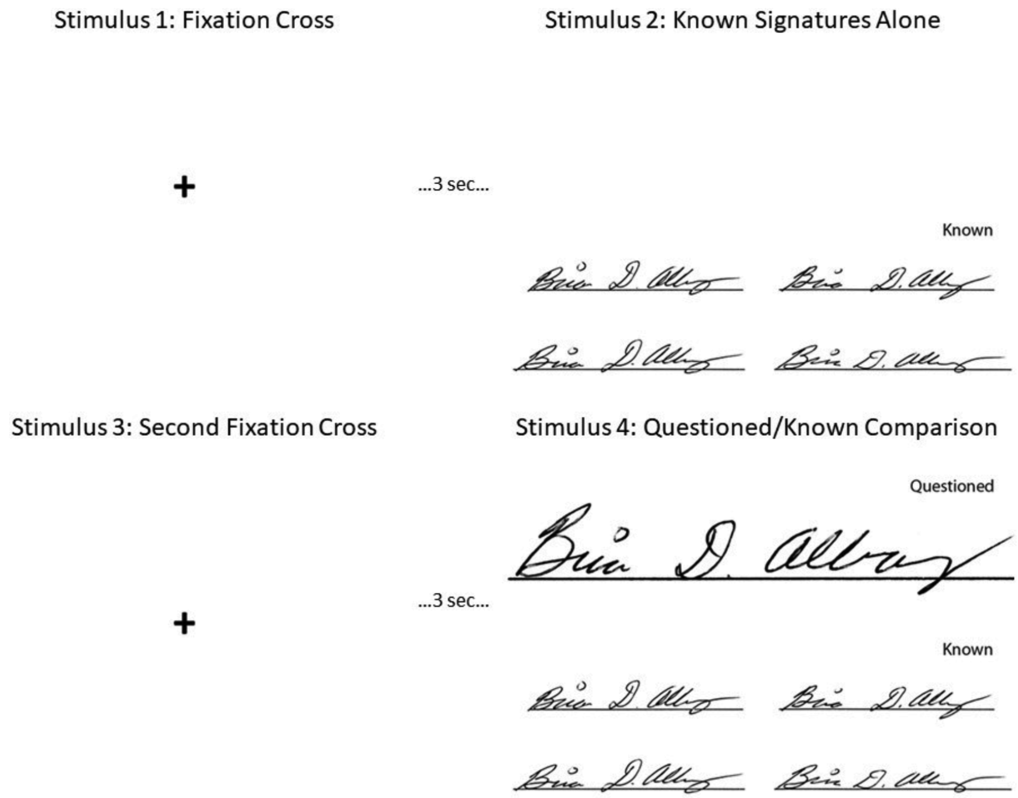

Figure 3. An example of the stimulus sequence displayed on the eye-tracking system. Each of the four images in the figure were displayed on separate screen. Stimulus 1 is a "fixation cross" where FDEs were instructed to center their gaze so that each comparison began in the same location. After being displayed for 3 seconds, the screen changed to Stimulus 2, the known signatures. FDEs viewed this screen for as long as they chose. When they were ready to complete the comparison, they indicated so verbally to the research assistant, who then changed the screen to Stimulus 3, a second fixation cross. After 3 seconds, the screen automatically changed to Stimulus 4, the questioned/known comparison. FDEs again viewed this screen for as long as they chose. When they completed each comparison they verbally indicated this to the research assistant, who exited the comparison screen. Fixations were recorded from the time the screen changed to the signatures until the research assistant exited that screen.

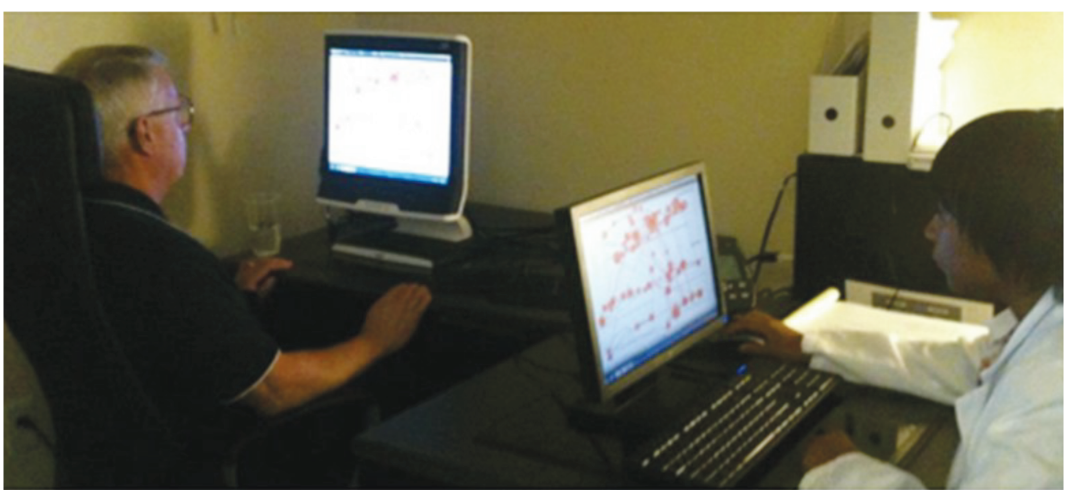

Figure 4. The participant is seated 57 cm away from the eye tracking unit. The operator, seated to the participant's right, controls the presentation of stimuli and records the participant's verbal responses. The eye-tracking protocols were conducted in low lighting for maximum pupil dilation, which improves the quality of the gaze data. In this image a gaze plot, or map of the fixation locations and the scan path, is displayed. The image displayed on the operator's screen is a gaze plot created from filtered fixation data.

plot shows the location of all fixations. The heat map shows a cumulative view of where the fixations are concentrated. Darker areas on the heat map show where the highest visual activity occurred.

Each AOI is identified by a different color. An AOI surrounds the questioned signature and separate AOIs surround the known signatures. Smaller AOIs were created within these five larger AOIs where the heat map and gaze plot indicate that fixation activity has been high. One large AOI encompasses all four known signatures. The largest AOI encompasses the entire stimulus. AOIs allowed us to compare the isolated areas separately from the other stimulus areas. The analyses we discuss here are based on the four AOIs isolating the known signatures.

Using these four AOIs, we created two new additional variables by combining the number of fixations into larger units. One variable isolated the fixations according to whether they occurred in the top two or the bottom two known signatures. The other variable isolated the fixations occurring in the two signatures on the left side or the right side of the stimulus. This results in the three AOI configurations demonstrated in Figure 6.

Utilization of Known Signatures

We performed a non-parametric Friedman test of differences among repeated measures[7] to investi-

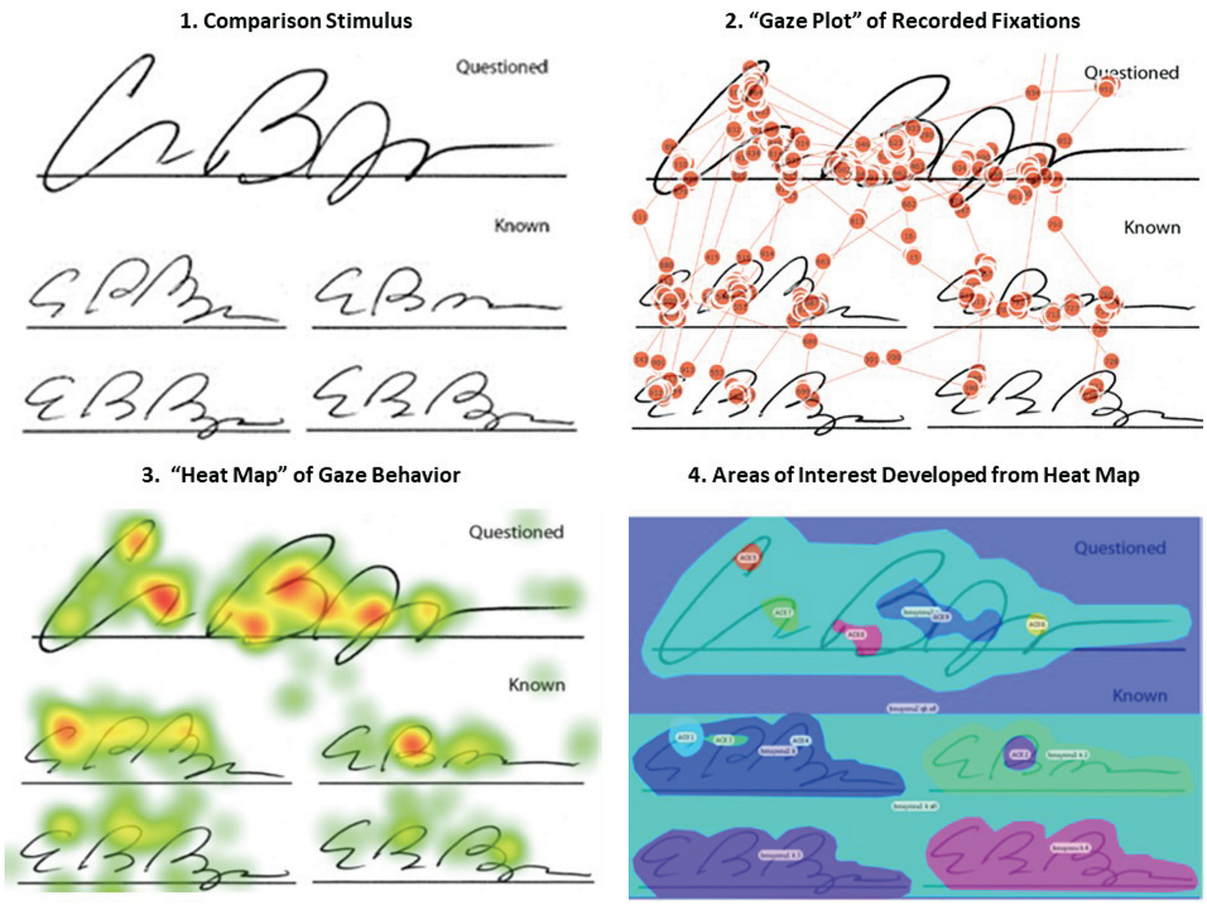

Figure 5. Image 1 (upper left) is an example of a signature comparison stimulus displayed on the eye-tracking screen. Image 2 (upper right) is a "gaze plot" visualization of the visual behavior of the FDE. Each dot represents a "fixation" recorded by the system when the FDE's eye movement drops below the threshold described above. The lines represent visual "saccades," or eye movement between fixations. The fixations are sequentially numbered by the eye-tracker to enable researchers to track the deployment of cognitive attention resources. Image 3 (lower left) is a "heat map" visualization of the areas that received to most attention from the FDE. Image 4 (lower right) demonstrates the areas of interest (AOIs) developed from the heat map.

[7] The Friedman test is a non-parametric equivalent of a one-sample related measures design. Test variables are ranked (in this case from 1 to 4). The chi-square test statistic is based on the ranks.

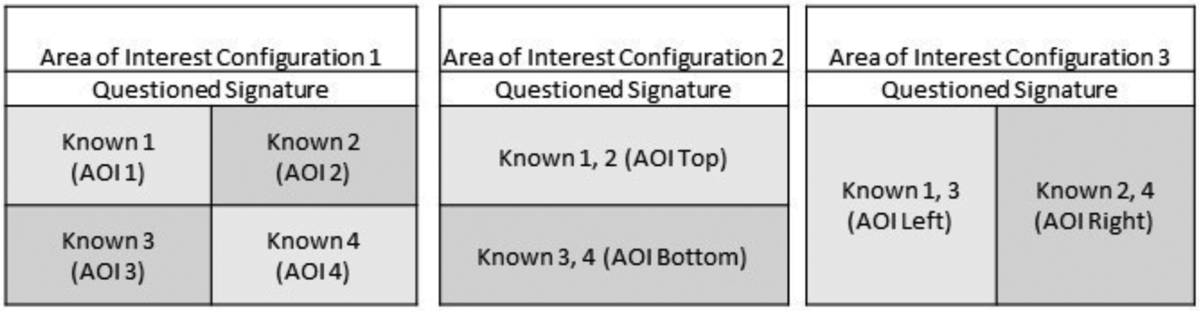

Figure 6. Configuration 1 includes all four known signatures. In configuration 2 the fixation counts and for known signatures 1 and 2 are combined in AOI Top, and 3 and 4 are combined in AOI Bottom. In configuration 3, fixation counts for signatures 1 and 3 are combined in AOI Left, 2 and 4 are combined in AOI Right.

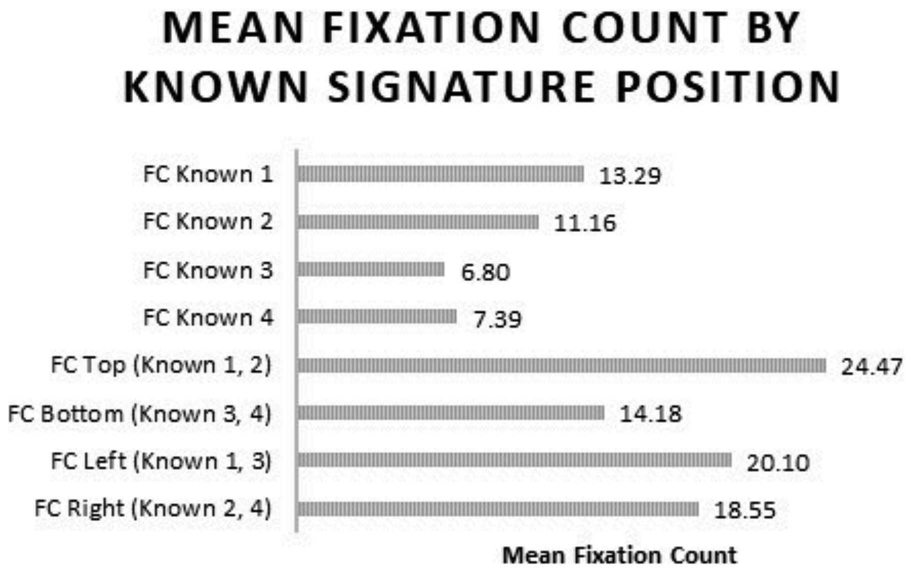

Figure 7. The statistically significant difference in the mean number of fixations demonstrated in this figure are consistent with a top-to-bottom and left-to-right viewing pattern. Fixations are highest in the upper left known signature (Known 1) then in the upper right signature (Known 2). Fixation count in Known 3 is significantly lower than Known 2, but significantly higher than in Known 4. Fixation count is significantly higher in the combined areas of Known 1 and 2 (Top), than in Known 3 and 4 (Bottom). Fixation count is also significantly higher in the combined areas of Known 1 and 3 (Left) than in Known 2 and 4 (Bottom). It is important to note that fixation counts are filtered data that identify where the FDE's gaze has slowed enough to meet the 50-pixel/100 ms threshold requirement. It should not be inferred from these data that FDEs failed to look at other areas where fixations were not indicated by the filter.

gate whether the differences in the mean number of fixations in the four known signatures were statistically significant. All analyses were conducted at an *alpha* level of 0.05. Mean FCs for each of the eight AOIs are presented in Figure 7.

We found that mean FCs varied significantly across Knowns 1-4, in a pattern suggesting a clockwise (left-to-right on top and right-to left on the bottom) deployment of attention due to the lower mean FC in Known 3 than in Known 4. A top-to-bottom pattern is clearly displayed by the higher number of fixations in the top Knowns compared to the bottom Knowns. A left-to-right pattern is also indicated in the higher number of fixations in the left Knowns than in the right. Significant results are highlighted in Table 1.[8]

[8] N = the number of observations on which the calculations are based. N for these observations is lower than N for the accuracy analyses because eye tracking data are not available for every examiner. We were unable to record gaze data for three FDEs whose corrective lenses interfered with the infrared illumination. These FDEs completed the comparisons and gave opinions, but no gaze data were collected. In several instances the eye tracking unit lost connection during a comparison, resulting in loss of an observation for that comparison. M = mean, and SD refers to the standard deviation of observations from the mean.

We compared the mean FCs for AOI configuration 2 (top Knowns versus bottom Knowns) and 3 (left Knowns versus right Knowns) by conducting paired-groups *t*-tests to investigate whether significant differences in mean FCs existed within the total observations for these AOIs.

Figure 8 highlights the AOIs where FCs were highest overall. Examination of the mean FCs from highest to lowest in each configuration reveals the clockwise, left-to-right, and proximal pattern of attention deployment.

Conclusion: Higher mean FCs in AOIs nearest to the known signature, and higher mean FCs in AOIs on the left side compared to the right side and means that decline in clockwise order indicate that use of the known signatures in this sample of observations reflects a left-to-right and proximal-to-distal pattern of attention.

Visual Behavior and Signature Characteristics

The differences observed above may also be influenced by characteristics of the signatures themselves. In this section we report analyses addressing the relationship between the type of signature (text-based or stylized), signature complexity (high or low complexity), and ground truth (whether the signature is genuine or non-genuine) and the gaze behavior of FDEs. We conducted a series of independent samples *t*-tests to investigate whether the mean FCs in each AOI differed significantly according the level of each signature characteristic (e.g., text-based vs. stylized, high vs. low complexity, genuine vs. non-genuine).

We also conducted a series of binomial logistic regression analyses to investigate which AOIs in combination might be significant predictors of the characteristics of the signatures (the outcome variables signature type, signature complexity, and ground truth).[9]

Table 1. Mean Fixation Count by Area of Interest

							95% Confidence Interval	
Fixation Count AOIs	**N**	***M***	**SD**	*t*	**df**	*p*	**Lower CI**	**Upper CI**
FC Knowns 1, 2 (top)	1650	24.47	22.95	26.92	1649	< .001	9.54	11.04
FC Knowns 3, 4 (bottom)	1650	14.18	19.07					
FC Knowns 1, 3 (left)	1650	20.1	21.5	4.4	1649	< .001	0.86	2.24
FC Knowns 2, 4 (right)	1650	18.55	20.25					
Fixation Count AOIs	N	*M*	SD	χ^2	df	*p*	Mean Rank	
FC Known 1	1652	13.29	13.05	1713.73	3	< .001	3.28	
FC Known 2	1650	11.16	11.23				2.95	
FC Known 3	1650	6.8	10.51				1.85	
FC Known 4	1650	7.39	11.33				1.92	

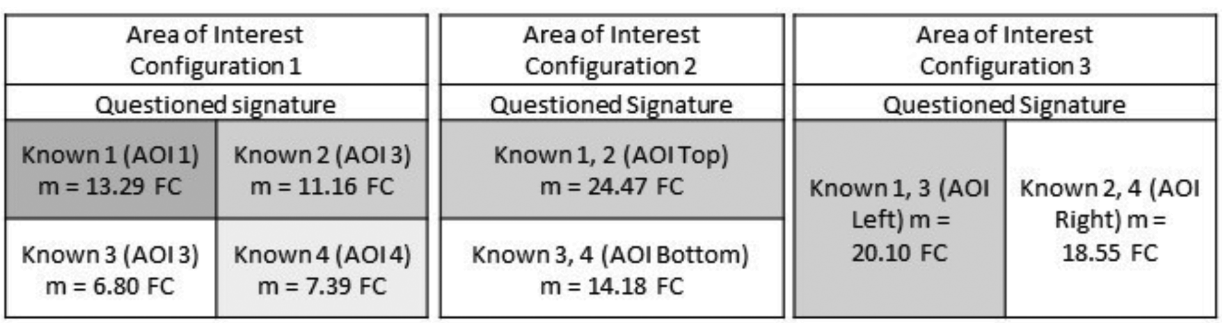

Figure 8. AOIs in which the highest mean fixation counts are highlighted. Areas of darker shading indicate AOIs with the higher fixation counts. All mean differences are statistically significant, as indicated in Table 1.

[9] While the t-tests measure whether the two group means are significantly different, a binomial logistic regression combines "predictor" variables (in this case, AOIs) in a model to predict which category of the two-category outcome variable (for example, high or low complexity) the observations fall under. The higher the number of correctly classified observations, the more reliable the model is considered. While the t-tests can tell us which AOIs are significantly different, they do not tell us anything about how the AOIs work in combination during an examination. Binomial logistic regression analyses reveal the changes in the odds of the observation falling into one category or the other, allowing us to identify the most influential predictor AOIs.

Signature Type Analyses

We performed a series of independent group t-tests to investigate whether the mean number of FCs differed according to whether the signature was text-based vs. stylized. The results are presented in Table 2. Statistically significant comparisons and the highest means are highlighted in gray.

These results reveal that in all comparisons but Known 2 and Known 3 the mean FC was higher in the text-based than in the stylized condition.

We performed a series of binary logistic regression analyses to investigate whether the observed differences in FC among the known signatures were related to whether the signatures were text-based or stylized. We analyzed each of the three AOI configurations identified above separately to avoid the influence of multicollinearity, which occurs when variables are highly correlated.[10]

Analysis 1: AOIs Top/Bottom. We combined the predictor variables FC Top (Knowns 1, 2), and FC Bottom (Knowns 3, 4) in a binomial logistic regression model to test how accurately the two factors together predicted category of signature type (the outcome variable).

The overall model was statistically significant (χ^2 (2) = 7.23, p = 0.027), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 0.59%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2198.29). The model correctly classified 61.1% of cases. *Wald* statistics indicated that when the variables were combined, neither predictor variable reached statistical significance.

Analysis 2 (Left/Right). We combined the predictor variables FC Left (Knowns 1, 3) and FC Right (Knowns 2, 4) in a binomial logistic regression model to test how accurately factors togeth-

Table 2. Mean Fixation Count by Signature Type

Signature Type								95% Confidence Interval	
	Sig Type	N	*M*	SD	*t*	df	*p*	Lower CI	Upper CI
FC Knowns 1, 2 (top)	Text	1008	25.64	24.33	2.701	1526	0.007	0.824	5.195
	Stylized	642	22.63	20.49					
FC Knowns 3, 4 (bottom)	Text	1008	15.03	20.98	2.429	1612	0.015	0.421	3.957
	Stylized	642	12.85	15.53					
FC Knowns 1, 3 (left)	Text	1008	21.08	22.88	2.416	1535	0.016	0.474	4.559
	Stylized	642	18.56	19.05					
FC Knowns 2, 4 (right)	Text	1008	19.60	22.11	2.788	1597	0.005	0.795	4.568
	Stylized	642	16.91	16.81					
FC Known 1	Text	1010	14.14	13.87	3.467	1538	0.001	0.948	3.419
	Stylized	642	11.95	11.51					
FC Known 2	Text	1008	11.48	11.84	1.454	1508	0.146	−0.278	1.874
	Stylized	642	10.68	10.19					
FC Known 3	Text	1008	6.91	11.36	0.574	1648	0.566	−0.737	1.346
	Stylized	642	6.61	9.02					
FC Known 4	Text	1008	8.12	12.99	3.654	1646	< .001	0.873	2.896
	Stylized	642	6.24	7.96					

Table 3. Regression Coefficients for Predictors of Signature Type

						95% Confidence Interval	
Analysis 1: Top/Bottom	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1, 2 (top)	−0.005	1.79	1	0.181	0.995	0.989	1.002
FC Knowns 3, 4 (bottom)	−0.002	0.275	1	0.6	0.998	0.99	1.006

[10] For example, AOI Top is a combination of AOI Knowns 1 and 3, so is highly correlated with Known 1 and Known 3.

er predicted category of signature type (the outcome variable). Results are presented in Table 4.

The overall model was statistically significant (χ^2 (2) = 7.36, p = 0.025), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 0.60%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2198.16). The model correctly classified 61.1% of cases. *Wald* statistics indicated that when the variables were combined, neither predictor variable reached statistical significance.

Analysis 3 (All Knowns). We combined the variables FC Known 1, FC Known 2, FC Known 3, and CF Known 4 in a binomial logistic regression model to test how accurately the four factors together were in predicting category of signature type (the outcome variable). Results are presented in Table 5

The overall model was statistically significant (χ^2 (4) = 29.47, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 2.4%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2176.05). The model correctly classified 61.2% of cases. *Wald* statistics indicated that when the variables were combined, all four variables were significant predictors of category of signature type. The odds ratios for all predictors were small, indicating low impact on the outcome.

Conclusion: The mean FC was higher in the text-based than in the stylized condition, although the differences in FCs in Known 2 and Known 3 did not reach statistical significance. Based on these analyses, we concluded that although the mean FCs for AOIs top/bottom and left/right were statistically significant, the known signatures were better predictors of signature type when their influence was measured separately, although analysis 3 explained only 2.4% of the variance. Odds ratios in all cases were very near to 1.0, indicating that change in the probability of the signature being text-based vs. stylized was very small.

Signature Complexity Analyses

We performed a series of independent group t-tests to investigate whether the mean number of FCs differed according to whether the signature was high complexity vs. low complexity. The results are presented in Table 6. Statistically significant comparisons and the highest means are highlighted in gray.

These results reveal that the only significant differences in mean FC were found in AOI Bottom (Knowns 3, 4) and AOI Known 4. In both instances the means were higher when the signatures were high complexity.

We performed a series of binary logistic regression analyses to investigate whether the observed differences in FC among the known signatures were related to whether the signatures were high or low complexity.

Analysis 1 (Top/Bottom). We combined the predictor variables FC Top (Knowns 1, 2), and FC Bottom (Knowns 3, 4) in a binomial logistic regression model to test how accurately the two factors together predicted whether signature complexity was high or low (the outcome variable). Results are presented in Table 7.

The overall model was statistically significant (χ^2 (2) = 23.85, $p < 0.001$), indicating that the variables improved the amount of variability ex-

Table 4. Regression Coefficients for Predictors of Signature Type

						95% Confidence Interval	
Analysis 2: Left/Right	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1, 3 (left)	−0.002	0.212	1	0.646	0.998	0.991	1.006
FC Knowns 2, 4 (right)	−0.005	1.749	1	0.186	0.995	0.986	1.003

Table 5. Regression Coefficients for Predictors of Signature Type

						95% Confidence Interval	
Analysis 3: All Knowns	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Known 1	-0.028	13.564	1	< .001	0.973	0.959	0.987
FC Known 2	0.021	6.774	1	0.009	1.021	1.005	1.038
FC Known 3	0.018	6.258	1	0.012	1.018	1.004	1.033
FC Known 4	-0.022	7.865	1	0.005	0.978	0.963	0.993

Table 6. Mean Fixation Count by Signature Complexity

	Comp	N	*M*	SD	*t*	df	*p*	95% Confidence Interval	
								Lower CI	Upper CI
FC Knowns 1, 2 (top)	High	690	24.22	22.69	−0.376	1648	0.707	−2.678	1.817
	Low	960	24.65	23.15					
FC Knowns 3, 4 (bottom)	High	690	15.79	22.67	2.750	1158	0.006	0.791	4.732
	Low	960	13.03	15.90					
FC Knowns 1, 3 (left)	High	690	20.33	22.75	0.362	1648	0.717	−1.716	2.494
	Low	960	19.94	20.56					
FC Knowns 2,4 (right)	High	690	19.68	22.36	1.866	1309	0.062	−0.100	3.984
	Low	960	17.74	18.55					
FC Known 1	High	690	13.35	13.16	0.168	1650	0.867	−1.168	1.386
	Low	962	13.24	12.97					
FC Known 2	High	690	10.87	10.95	−0.914	1648	0.361	−1.612	0.587
	Low	960	11.38	11.43					
FC Known 3	High	690	6.97	12.24	0.586	1648	0.558	−0.722	1.336
	Low	960	6.67	9.07					
FC Known 4	High	690	8.82	14.48	4.012	1007	< .001	1.254	3.655
	Low	960	6.36	8.23					

Table 7. Regression Coefficients for Predictors of Signature Complexity

						95% Confidence Interval	
Analysis 1: Top/Bottom	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1, 2 (top)	0.014	14.281	1	< .001	1.014	1.007	1.021
FC Knowns 3, 4 (bottom)	−0.021	20.298	1	< .001	0.979	0.971	0.988

plained (Nagelkerke $r^2 = 1.9\%$). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2194.79). The model correctly classified 60.7% of cases. *Wald* statistics indicated that when the variables were combined, both predictor variables reached statistical significance. The odds ratio for both variables were small, indicating low impact on the outcome.

Analysis 2 (Left/Right). We combined the predictor variables FC Left (Knowns 1, 3) and FC Right (Knowns 2, 4) in a binomial logistic regression model to test how accurately the factors together predicted category of complexity (the outcome variable). Results are presented in Table 8.

The overall model was statistically significant ($\chi^2(2) = 6.67$, $p = 0.036$), indicating that the variables improved the amount of variability explained (Nagelkerke $r^2 = 0.54\%$). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2236.33). The model correctly classified 59.0% of cases. *Wald* statistics indicated that when the variables were combined, FC Knowns 2, 4 (Right) reached statistical significance. The odds ratio for this variable was small, indicating low impact on the outcome.

Analysis 3 (All Knowns). We combined the variables FC Known 1, FC Known 2, FC Known 3, and CF Known 4 in a binomial logistic regression model to test how accurately the four factors together were in predicting category of complexity (the outcome variable). Results are presented in Table 9.

Table 8. Regression Coefficients for Predictors of Signature Complexity

						95% Confidence Interval	
Analysis 2: Left/Right	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1, 3 (left)	0.006	2.913	1	0.088	1.006	0.999	1.014
FC Knowns 2, 4 (right)	-0.01	6.325	1	0.012	0.99	0.983	0.998

Table 9. Regression Coefficients for Predictors of Signature Complexity

						95% Confidence Interval	
Analysis 3: All Knowns	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Known 1	< .001	< .001	1	0.990	1.000	0.986	1.014
FC Known 2	0.031	13.870	1	< .001	1.031	1.015	1.048
FC Known 3	0.003	0.189	1	0.664	1.003	0.990	1.016
FC Known 4	−0.046	31.831	1	< .001	0.955	0.939	0.970

The overall model was statistically significant (χ^2 (4) = 45.13, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 3.6%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2197.84). The model correctly classified 61.0% of cases. *Wald* statistics indicated that when the variables were combined, variables FC Known 2 and FC Known 4 were significant predictors of category of complexity. The odds ratios for all three predictors were small, indicating low impact on the outcome.

Conclusion: Only two means (AOI Bottom and AOI Known 4) reached statistical significance. We observed that when combined with other predictor variables, FC Top and Bottom were significant predictors of the category of signature complexity in model 1. FC Right was also a significant predictor in model 2, and Knows 2 and 4, were significant predictors in model 3.

Based on these analyses, we concluded that again, the known signatures were better predictors of signature type when their influence was measured separately, as demonstrated by the amount of variability explained in Analysis 3. Odds ratios in all cases were still very near to 1.0, indicating that changes in the probability of the signature being high vs. low complexity were very small.

Ground Truth Analyses

We performed a series of independent group t-tests to investigate whether the mean number of FCs differed according to whether the signature was genuine vs. non-genuine. The results are presented in Table 10. Statistically significant comparisons and the highest means are highlighted in gray.

These results reveal that all differences in mean FC were statistically significantly higher for genuine signatures than for non-genuine signatures.

We performed a series of binary logistic regression analyses to investigate whether the observed differences in FCs among the known signatures were related to whether the signatures were genuine or non-genuine (ground truth).

Table 10. Mean Fixation Count by Ground Truth

Ground Truth								95% Confidence Interval	
	Truth	**N**	*M*	**SD**	*t*	**df**	*p*	**Lower CI**	**Upper CI**
FC Knowns 1, 2 (top)	Genuine	686	28.64	25.18	6.100	1291	< .001	4.844	9.437
	Non	964	21.50	20.73					
FC Knowns 3, 4 (bottom)	Genuine	686	17.67	20.65	6.172	1316	< .001	4.077	7.876
	Non	964	11.70	17.45					
FC Knowns 1, 3 (left)	Genuine	686	24.32	24.49	6.504	1214	< .001	5.038	9.391
	Non	964	17.10	18.53					
FC Knowns 2,4 (right)	Genuine	686	22.00	21.04	5.811	1396	< .001	3.910	7.896
	Non	964	16.10	19.31					
FC Known 1	Genuine	688	15.17	14.11	4.852	1331	< .001	1.918	4.521
	Non	964	11.95	12.06					
FC Known 2	Genuine	686	13.43	12.55	6.740	1250	< .001	2.749	5.006
	Non	964	9.55	9.89					
FC Known 3	Genuine	686	9.10	12.90	7.103	1056	< .001	2.860	5.042
	Non	964	5.15	8.01					
FC Known 4	Genuine	686	8.57	9.96	3.592	1648	< .001	0.920	3.132
	Non	964	6.55	12.15					

Analysis 1 (Top/Bottom). We combined the predictor variables FC Top (Knowns 1, 3) FC Bottom (Knowns 3, 4) in a binomial logistic regression model to test how accurately the three factors together predicted category of ground truth (the outcome variable). Results are presented in Table 11.

The overall model was statistically significant (χ^2 (2) = 45.54, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 3.7%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2118.03). The model correctly classified 60.2% of cases. *Wald* statistics indicated that when the variables were combined, FC Top and FC Bottom were significant predictors of category of ground truth. The odds ratios for these variables were small, indicating low impact on the outcome.

Analysis 2 (Left/Right). We combined the predictor variables FC Left (Knowns 2, 4, 6) and FC Right Knowns 1, 3, 5) in a binomial logistic regression model to test how accurately the three factors together predicted ground truth (the outcome variable). Results are presented in Table 12.

The overall model was statistically significant (χ^2 (2) = 41.17, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 3.8%). Results indicated that the overall model fit was poor (-2 Log Likelihood = 2193.15). The model correctly classified 60.4% of cases. *Wald* statistics indicated that when the variables were combined, only FC 1, 3 (left) was a significant predictor of category of ground truth. The odds ratios for this variable was small, indicating low impact on the outcome.

Analysis 3 (All Knowns). We combined the variables FC Known 1, FC Known 2, FC Known 3, FC Known 4, FC Known 5, and FC Known 6 in a binomial logistic regression model to test how accurately the six factors together were in predicting category of ground truth (the outcome variable). Results are presented in Table 13.

The overall model was statistically significant (χ^2 (4) = 84.56, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 6.7%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 2155.77). The model correctly classified 62.1% of cases. *Wald* statistics indicated that when the variables were combined, FC Known 1, FC Known 2, and FC Known 3 were significant predictors of category of ground truth. The odds ratios for these variables were small, indicating low impact on the outcome.

Conclusion: All FC means comparisons reached statistical significance, and in all comparisons were greater for genuine than for non-genuine signatures. We observed that when combined with other predictor variables, FC Top and

Table 11. Regression Coefficients for Predictors of Ground Truth

						95% Confidence Interval	
Analysis 1: Top/Bottom	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1, 2 (top)	−0.007	4.432	1	0.035	0.993	0.986	0.999
FC Knowns 3, 4 (bottom)	−0.011	5.998	1	0.014	0.989	0.981	0.998

Table 12. Regression Coefficients for Predictors of Ground Truth

						95% Confidence Interval	
Analysis 2: Left/Right	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1, 3 (left)	−0.014	12.289	1	< .001	0.986	0.978	0.994
FC Knowns 2, 4 (right)	−0.004	0.824	1	0.364	0.996	0.989	1.004

Table 13. Regression Coefficients for Predictors of Ground Truth

						95% Confidence Interval	
Analysis 3: All Knowns	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Known 1 (top left)	0.021	8.341	1	0.004	1.022	1.007	1.036
FC Known 2 (top right)	−0.033	16.553	1	< .001	0.967	0.952	0.983
FC Known 3 (bottom left)	−0.050	30.359	1	< .001	0.951	0.934	0.968
FC Known 4 (bottom right)	0.012	3.550	1	0.060	1.012	1.000	1.024

Bottom were significant predictors of the category of signature complexity in model 1. FC Left was also a significant predictor in model 2, and Knowns 1, 2, and 4 were significant predictors of the category of ground truth in model 3.

Based on these analyses, we again concluded that the known signatures were better predictors of signature type when their influence was measured separately, as demonstrated by the amount of variability explained in Analysis 3 (6.7%). Odds ratios in all cases were still very near to 1.0, indicating that changes in the probability of the signature being genuine vs. non-genuine were very small.

FDE Decision Accuracy Analyses

We performed a series of independent group t-tests to investigate whether the mean number of FCs differed according to whether the FDE decisions were accurate vs. misleading. The results are presented in Table 14. Statistically significant comparisons and the highest means are highlighted in gray.

These results reveal that all differences in mean FC were statistically significantly higher for misleading than for accurate FDE decision.

We performed a series of binary logistic regression analyses to investigate whether the observed differences in FCs among the known signatures were related to whether the FDE decisions were accurate or misleading.

Analysis 1 (Top/Bottom)

We combined the predictor variables FC Top (Knowns 1, 2), and FC Bottom (Knowns 3, 4) in a binomial logistic regression model to test how accurately the two factors together predicted category of decision accuracy (the outcome variable). Results are presented in Table 15.

The overall model was statistically significant (χ^2 (2) = 38.27, p < 0.001), indicating that the variables improved the amount of variability

Table 14. Mean Fixation Count by FDE Decision Accuracy

	FDE Decision Accuracy								
								95% Confidence Interval	
	Accuracy	N	*M*	SD	*t*	df	*p*	Lower CI	Upper CI
FC Knowns 1, 2 (top)	Accurate	1413	22.90	22.26	−6.380	303	< .001	−14.315	−7.566
	Mislead	237	33.84	24.78					
FC Knowns 3, 4 (bottom)	Accurate	1413	13.21	19.06	−5.079	1648	< .001	−9.354	−4.142
	Mislead	237	19.96	18.15					
FC Knowns 1, 3 (left)	Accurate	1413	18.71	21.07	−6.234	311	< .001	−12.699	−6.606
	Mislead	237	28.37	22.22					
FC Knowns 2,4 (right)	Accurate	1413	17.40	19.96	−5.573	315	< .001	−10.874	−5.199
	Mislead	237	25.43	20.64					
FC Known 1	Accurate	1414	12.44	12.72	−6.135	308	< .001	−7.773	−3.998
	Mislead	238	18.33	13.85					
FC Known 2	Accurate	1413	10.45	10.87	−5.844	300	< .001	−6.665	−3.307
	Mislead	237	15.43	12.36					
FC Known 3	Accurate	1413	6.26	10.44	−5.057	321	< .001	−5.137	−2.259
	Mislead	237	9.96	10.41					
FC Known 4	Accurate	1413	6.95	11.50	−4.284	352	< .001	−4.451	−1.650
	Mislead	237	10.00	9.90					

Table 15. Regression Coefficients for Predictors of Decision Accuracy

						95% Confidence Interval	
Analysis 1: Top/Bottom	*B*	Wald	df	*p*	Odds	Lower CI	Upper CI
FC Knowns 1 and 2 (top)	0.017	18.015	1	< .001	1.017	1.009	1.024
FC Knowns 3 and 4 (bottom)	< .001	0.002	1	0.961	1.000	0.991	1.009

explained (Nagelkerke $r^2 = 4.1\%$). Results indicated that the overall model fit was poor (−2 Log Likelihood = 1319.16). The model correctly classified 85.5% of cases. *Wald* statistics indicated that when the variables were combined, FC Top was a significant predictor of category of decision accuracy. The odds ratios for FC Top was small, indicating low impact on the outcome.

Analysis 2 (Left/Right)

We combined the predictor variables FC Left (Knowns 1, 3) and FC Right (Knowns 2, 4) in a binomial logistic regression model to test how accurately the factors together predicted category of decision accuracy (the outcome variable). Results are presented in Table 16.

The overall model was statistically significant (χ^2 (2) = 35.18, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke $r^2 = 3.8\%$). Results indicated that the overall model fit was poor (−2 Log Likelihood = 1322.80). The model correctly classified 85.3% of cases. *Wald* statistics indicated that when the variables were combined, the only predictor variable to reach significance was FC Left. The odds ratio for FC Left was small, indicating low impact on the outcome.

Analysis 3 (All Knowns)

We combined the variables FC Known 1, FC Known 2, FC Known 3, and FC Known 4 in a binomial logistic regression model to test how accurately the four factors together were in predicting category of decision accuracy (the outcome variable). Results are presented in Table 17.

The overall model was statistically significant (χ^2 (4) = 39.39, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke $r^2 = 4.2\%$). Results indicated that the overall model fit was poor (−2 Log Likelihood = 1318.60). The model correctly classified 85.3% of cases. *Wald* statistics indicated that when the variables were combined, none of the predictor variable reached significance.

Conclusion: Although all mean comparisons were statistically significantly higher for misleading calls, none of the individual known signatures were significant predictors of the category of decision accuracy. Analysis 1 revealed that FC Top significantly predicted decision accuracy, and Analysis 2 revealed the FC Left was also a significant predictor. All three models explained approximately the same amount of variability. Taken together, these analyses suggest that Known 1, which is common to both AOI Top (Knowns 1, 2) and AOI Left (Knowns 1, 3), accounts for these statistically significant differences.

Combined Signature Characteristic and Fixation Count Analyses

We conducted *chi-square* (χ^2) analyses to determine whether significant differences in FDE decision accuracy occurred for text-based versus stylized signatures, high complexity versus low complexity signatures, and genuine versus non-genuine signatures. The difference in decision accuracy for text-based (86.4%) versus stylized signatures (85.4%) was not statistically significant, (χ^2 (1) = .349, $p = 0.554$). Decision accuracy was significantly greater for high complexity (89.8%) than for low complexity signatures (83.2%), (χ^2(1) = 262.92, $p < 0.001$). Decision accuracy was sig-

Table 16. Regression Coefficients for Predictors of Decision Accuracy

						95% Confidence Interval	
Analysis 2: Left/Right	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Knowns 1 and 3 (left)	0.013	8.101	1	0.004	1.013	1.004	1.022
FC Knowns 2 and 4 (right)	0.005	1.352	1	0.245	1.005	0.996	1.015

Table 17. Regression Coefficients for Predictors of Decision Accuracy

						95% Confidence Interval	
Analysis 3: All Knowns	*B*	**Wald**	**df**	*p*	**Odds**	**Lower CI**	**Upper CI**
FC Known 1	0.016	3.776	1	0.052	1.016	1.000	1.033
FC Known 2	0.017	3.063	1	0.080	1.017	0.998	1.036
FC Known 3	0.005	0.374	1	0.541	1.005	0.99	1.019
FC Known 4	−0.004	0.272	1	0.602	0.996	0.981	1.011

nificantly greater for non-genuine (97.4%) than for genuine signatures (69.9%). This means that in this sample of signatures, both complexity and ground truth are significant predictors of FDE decision accuracy.

As a final step, we combined all the predictor variables (signature type, signature complexity, ground truth, and all FC variables) in a binomial logistic regression model to investigate which factors together predicted the category of decision accuracy (the outcome variable). Because of high multicollinearity (correlations among the FC variables), we again included only the individual signature variables (FC Known 1, FC Known 2, FC Known 3, and FC Known 4) in the final decision accuracy analysis. Results are presented in Table 18.

The overall model was statistically significant (χ^2 (7) = 292.17, $p < 0.001$), indicating that the variables improved the amount of variability explained (Nagelkerke r^2 = 28.9%). Results indicated that the overall model fit was poor (−2 Log Likelihood = 1065.81). The model correctly classified 85.5% of cases. *Wald* statistics indicated that when the variables were combined, the only significant predictors of category of decision accuracy were ground truth and FC Known 1. Although this model explained considerably more of the variability (28.9%) than any of the previous models, the odds ratios for the significant predictor variables were still small, indicating low impact on the likelihood of predicting the decision accuracy outcome.

Summary and Conclusions

Use of Known Signatures. We investigated FDE use of known signatures by comparing the mean FCs for three different AOI configurations. FCs were higher for known signatures nearest to the questioned signature. Signatures on the left attracted more fixations than did the signatures on the right. Higher mean FCs in AOIs nearest to the known signature, higher mean FCs in AOIs on the left side compared to the right side, and means that decline in clockwise order indicate that use of the known signatures in this sample of observations reflects a left-to-right and proximal-to-distal pattern of attention during the comparisons.

Use of Known Signatures by Signature Type. We investigated the relationship between several signature characteristics and FDE use of the known signatures by comparing mean FCs and conducting binomial logistic regression analyses for each of the three different AOI configurations. These analyses revealed that the mean FC was higher in the text-based than in the stylized condition for all comparisons but FCs in Known 2 and Known 3. Although the mean FCs for AOIs top/bottom and left/right were statistically significant, the known signatures were better predictors of signature type when their influence was measured separately.

Use of Known Signatures by Signature Complexity. We investigated FDE use of known signatures when the signatures were high complexity versus low complexity. We again concluded that the known signatures were better predictors of signature type when their influence was measured separately, as demonstrated by the amount of variability

Use of Known Signatures by Ground Truth. We performed an analysis of means and a series of binary logistic regression analyses to investigate whether the observed differences in FCs among the known signatures were related to whether the signatures were genuine or non-genuine (ground truth). All FC means comparisons reached statistical significance, and in all comparisons the means were greater for genuine than

Table 18. Regression Coefficients for Decision Accuracy Predictors

						95% Confidence Interval	
Predictor Variables	*B*	**Wald**	**df**	*p*	**Odds Ratio**	**Lower CI**	**Upper CI**
Signature Type	−0.090	1.293	1	0.256	0.914	0.782	1.068
Complexity	0.229	1.878	1	0.171	1.258	0.906	1.746
Ground Truth	−2.699	153.035	1	< .001	0.067	0.044	0.103
FC Known 1	0.027	8.125	1	0.004	1.027	1.008	1.047
FC Known 2	< .001	< .001	1	0.989	1.000	0.980	1.021
FC Known 3	−0.010	1.281	1	0.258	0.990	0.973	1.007
FC Known 4	0.003	0.082	1	0.775	1.003	0.983	1.023

for non-genuine signatures. We again concluded that the known signatures were better predictors of signature type when their influence was measured separately, as demonstrated by the amount of variability explained in the regression model including only the individual known signatures (6.7%).

Use of Known Signatures and FDE Opinion Accuracy. We performed a series of binary logistic regression analyses to investigate whether the observed differences in FCs among the known signatures were related to whether the FDE decisions were accurate or misleading (decision accuracy). Although all mean comparisons were statistically significantly higher for misleading calls, none of the individual known signatures were significant predictors of the category of decision accuracy. All three regression models explained approximately the same amount of variability (4.1%). Taken together, these analyses suggest that Known 1, which is common to both AOI Top (Knowns 1, 2) and AOI Left (Knowns 1, 3), accounts for these statistically significant differences.

Signature Characteristics and FDE Decision Accuracy. We conducted *chi-square* (χ^2) analyses to determine whether significant differences in FDE decision accuracy occurred for text-based versus stylized signatures, high complexity versus low complexity signatures, and genuine versus non-genuine signatures. Decision accuracy was significantly related to signature complexity and ground truth, but not significantly related to signature type.

Decision Accuracy, Signature Characteristics, and Known Signature AOI Fixation Count. After combining all signature characteristic variables and the AOI variables FC Known 1, FC Known 2, FC Known 3, and FC Known 4 in a final binary logistic regression model to investigate which of these variables were significant predictors of the category of FDE decision accuracy. Taken in combination, the only significant predictors of FDE decision accuracy were ground truth and FC Known 1 (upper left).

Implications

The movement of expert testimony from the status of "proffer" to that of "admissible evidence" is a social process in which experts, attorneys, and judges all participate. It is a negotiated movement from "science," which is itself a social construction [18], to "legal science" [19], which is mediated by the rhetoric and discourse of attorneys, judges, and academicians. Transparency of methods is an important component of the admissibility of FDE testimony. Eye tracking methodology, physiological data, the diagnostic value of the evidential features of handwriting, and descriptions of the decision-making process will help increase the transparency of the examination process, improving the quality of performance of attorneys, judges, and experts.

Practitioners from many forensic fields have taken seriously the need for standardized training and proficiency testing, and through organizations such as the NIST OSAC are working nationally and internationally to define and establish valid and reliable measures of certainty, proficiency, and error. Forensic experts around the world are striving to ensure that their methods are transparent to the courts, and that judges are given the information they need to make their decisions. Efforts to organize and present information effectively, which are important goals of OSAC, have been an important consequence of the *Daubert* trilogy[11] and the 2009 National Academies of Science report [1]. Forensic scientists are also seeking opportunities to collaborate with judges, attorneys, and scientists from other fields on research and education projects.

These findings have implications not only for Questioned Document Examination, but also for other areas of Pattern and Physics Evidence identified by NIST OSAC and other organizations. Our research methods can be adapted to other disciplines, which will increase the understanding of cognitive human factors in those fields and provide information about possible sources of cognitive bias, such as the visual or cognitive context of the examination, top-down/bottom-up processing of information, order of presentation

[11] Daubert v. Merrell-Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). The four pronged "test" for the evidence/expert includes the following guidelines: (1) general acceptance, (2) falsifiability, (3) error rate, and (4) peer review and publication. Daubert also gave birth to two other controversial and historical cases that with Daubert, became known as the "Daubert Trilogy": Gen. Elec. Co. v. Joiner, 522 U.S. 136 (1997) and Kumho Tire Co. v. Patrick Carmichael, 526 U.S. 137 (1999).

effects, word-superiority effects, and other relevant cognitive phenomena.

Limitations and Future Directions

Although these findings are based on 1,742 FDE decisions that afford good statistical power, these experimental findings do not approximate the working document examination laboratory conditions under which signature comparisons are conducted. The participants were constrained by the experimental environment and the amount of control necessary to ensure that possible confounding circumstances were held as constant as possible. FDEs were limited to only four known signatures for their comparisons, and they did not have access to the tools or resources that are typically available to them during an examination. Future research could address not only the underlying cognitive and physiological aspects of the examination process, but also the broader laboratory environment of the examiner, where the examinations really occur.

Conclusion

Despite these constraints, experimental data provide valuable insights about the mechanisms underlying the decision-making process. Eye movement studies about human gaze behavior have revealed that visual behavior changes depending on the visual task, whether we are casually interacting with our natural environment, whether our attention is captured by salient features, or whether we are engaging in goal-directed behavior in which our attention is strategically focused on what is relevant to accomplishing the tasks we perform [20, 21, 22].

The results of this study are consistent with the findings of these and other studies, as this study demonstrated that context effects do impact the visual behavior of FDEs. Mitigating any misleading effects due to the relative positions of the questioned and known signatures may be as simple as shifting the position of the questioned and known materials during the examination to ensure that all the available data are utilized. Better understanding the physiological and cognitive contexts of document examination offers important opportunities to find potential sources of bias and error, and to develop evidence-based protocols and procedures to help mitigate their influence. Ongoing research in both experimental academic laboratories and forensic laboratories where examinations occur is needed to gain a full understanding of the relationship between human factors and forensic science.

Acknowledgements

NIJ Award No. 2010-DN-BX-K271: Kentucky State University: Undergraduate assistants Sandra Keene, Mickail Buster-Jones, Inna Malyuk, Cierra Alexander, Kara Francis, Nick Williams, Laurice Jackson, La'Quida Smith, Ivan Duvall, Savada Smothers, and Melissa Pickett; University of Nevada, Reno: Christopher Sanchez, Survey Manager; Martha Rodriguez, Assistant Survey Manager; Christopher Swinger, Supervisor; and the survey interview staff of the Center for Research Design and Analysis; Denise Schaar-Buis, Research Faculty Assistant, Grant Sawyer Center for Justice Studies; and graduate assistants Mauricio Alvarez, Lindsay M. Perez, and Vicky Springer of the Grant Sawyer Center for Justice Studies, University of Nevada, Reno.

References

1. National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: The National Academies Press.

2. Dror, I. E. (2013). The ambition to be scientific: Human expert performance and objectivity. *Science & Justice*, Vol. 53 (2), pp. 81–82.

3. Human Factors and Ergonomics Society. (2020). *The Definitions of HF/E*. Retrieved from https://www.hfes.org/resources/educational-and-professional-resources/new-item

4. Expert Working Group on Human Factors in Latent Print Analysis. (2012). *Latent print examination and human factors: Improving the practice through a systems approach*. U.S. Department of Commerce, National Institute of Standards and Technology.

5. Expert Working Group for Human Factors in Handwriting Examination. (2020). *Forensic handwriting examination and human factors: Improving the practice through a systems approach.* U.S. Department of Commerce, National Institute of Standards and Technology. doi:10.6028/NIST.IR.8282

6. Dyer, A., Found, B., & Rogers, D. (2006). Visual attention and expertise for forensic signature analysis,

Journal of Forensic Sciences, Vol. 51, pp. 1367–1404.

7. Dyer, A., Found, B., and Rogers, D, (2008). An insight into forensic document examiner expertise for discriminating between forged and disguised signatures. *Journal of Forensic Science*, Vol. 53, pp. 1154–1159.

8. Dyer, A., Found, B., Merlino, M. L., Pepe, A. L., Rogers, D. K., & Sita, J. (2014). Eye movement evaluation of signature forgeries: Insights to forensic expert evidence In M. E. Horsley, M. Eliot, B. Knight, & R. Reilly (Eds.), *Current Trends in Eye Tracking Research* (pp. 211–223).

9. Tatler B. (2014). Eye Movements from Laboratory to Life. In M. E. Horsley, M. Eliot, B. Knight, & R. Reilly (Eds.), *Current Trends in Eye Tracking Research* (pp. 17–35).

10. Mohammed, L., Found, B., & Rogers, D. (2008). Frequency of signatures styles in San Diego County, *Journal of the American Society of Questioned Document Examiners*, Vol. 1, pp. 9–13.

11. Nguyen, D. M. T., Hammond, D. L., & Salyards, M. J. (2011). Signature frequency and classification in the military. In American Academy of Forensic Sciences, *Annual Meeting 2011, Chicago, IL.*

12. Found, B., & Rogers, D. (1996). The forensic examination of signature complexity. In M. L. Simmer, C. G. Leedham, & A. J. W. M. Thomassen (Eds.), *Handwriting and Drawing Research: Basic Issues* (pp.483–492). Amsterdam: IOs Press.

13. Found, B., & Rogers, D. (1998). A consideration of the theoretical basis of forensic handwriting examination. *International Journal of Forensic Document Examiners*, Vol. 4, pp. 109–118.

14. Dewhurst, T., Found, B., & Rogers, D. (2008). Are expert penmen better than lay people at producing simulations of a model signature? *Forensic Science International*, Vol. 180, pp. 50–53.

15. Mohammed, L. A. (2019). Assessment of signature complexity. In L. Mohammed (Ed.), *Forensic Examination of Signatures* (pp. 31–42).

16. Found, B., Rogers, D., Rowe, V., & Dick, D. (1998). Statistical modelling of experts' perceptions of the ease of signature simulation. *Journal of Forensic Document Examination*, Vol. 11, pp. 75–102.

17. Alewijnse, L. C., van den Heuvel, E. C., & Stoel, R. D. (2011). Analysis of signature complexity. *Journal of Forensic Document Examination*, Vol. 21, pp. 37–49.

18. Jasanoff, S. (1993). What judges should know about the sociology of science. *Judicature,* Vol. 7, pp. 77–82.

19. Caudill, D. S. (2001). Law and science: An essay on links and socio-natural hybrids. *Syracuse Law Review,* Vol. 51, pp. 841–862.

20. Foulsham T. (2015). Eye movements and their functions in everyday tasks. *Eye (London, England)*, Vol. 29, pp. 196–199.

21. Dicks, M., Button, C., & Davids, K. (2010). Examination of gaze behaviors under in situ and video simulation task constraints reveals differences in information pickup for perception and action. *Attention, Perception, & Psychophysics*, Vol. 72, pp. 706–720.

22. Hayhoe, M. M., McKinney, T., Chajka, K., & Pelz, J. B. (2011). Predictive eye movements in natural vision. *Experimental Brain Research*, Vol. 217, pp. 125–136.