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

Document Title:	Kinematic Validation of FDE Determinations About Writership in Questioned Handprinting and Handwriting
Author(s):	Michael P. Caligiuri, Ph.D.
Document Number:	304651
Date Received:	April 2022
Award Number:	2017-DN-BX-0148

Agency Name: NATIONAL INSTITUTE OF JUSTICE
Agency Award Number: 2017-DN-BX-0148

Project title: Kinematic validation of FDE determinations about writership in questioned handprinting and handwriting

PI: Michael P. Caligiuri, Ph.D.; Professor, UCSD; mcaligiuri@ucsd.edu; 9500 Gilman Drive, San Diego, CA 92093; 760-492-2291

Submitting Official: Sennite Meche, Contract & Grant Officer, smeche@ucsd.edu (858) 822-6554

Submission Date: 9/30/2019

DUNS and EIN numbers:

Recipient Organization: Regents of the University of California

Recipient Identifying Number: 956006144

Project Period: January 1, 2018 – December 31, 2019

Reporting Period: January 1, 2018 – December 31, 2019

Report Term: 24-month

Final Report: Yes

Signature of Submitting Official:

Sennite Meche
Contract & Grant Officer
University of California, San Diego

1. Purpose

Since *Daubert*, several challenges to the reliability, validity, and subsequent admissibility of handwriting evidence have been raised in Federal court. Of particular relevance to the problem addressed by this research are the conclusions reached by Judge Rakoff in *Almeciga v Centers for Investigative Reporting*. In his 2015 ruling, Judge Rakoff excluded expert testimony on handprinting following a Daubert hearing concluding that without a refined methodology, forensic document examination "is virtually untestable, rendering it an unscientific endeavor". Rakoff noted further that the peer reviewed literature "has not been sufficiently robust or objective to lend credence to the proposition that handwriting comparison is a scientific discipline".

The aim of this research is to examine the foundational validity of FDE writership determinations by examining the relationships between FDE strength of support concerning writership propositions and the differences in the kinematic feature dynamic distributions of the two writing samples. We employed univariate and multivariate regression procedures to estimate the degree to which kinematic feature dissimilarities account for variability in FDE strength of support for alternate propositions. A finding that FDE writership opinions are statistically associated with these kinematic feature dissimilarity scores would support the scientific basis for handwriting examination and strengthen the validity of the FDE decision-making process.

2. Project Subjects

This project involved two groups of subjects. Thirty-five writers recruited from the San Diego Sheriff's Crime Laboratory were instructed to write six phrases from the London Letter and to repeat each phrase five times using both print and cursive writing styles. The second group of subjects consisted of 41 FDEs recruited from North America, Europe, and Australia/New Zealand to participate in a survey of these handwriting samples to obtain strength of support for alternate propositions regarding writership.

3. Methods and Procedures

3.1. Handwriting Sample Collection

Table 1 shows the handwriting samples used for this study. Subjects wrote each of the six phrases five times with an inking pen on lined paper placed on a Wacom digitizing tablet. The stimulus phrase was shown on the top of each page and repetitions were written vertically, five per page. Seven subjects returned to the lab two weeks later and repeated the writing experiment.

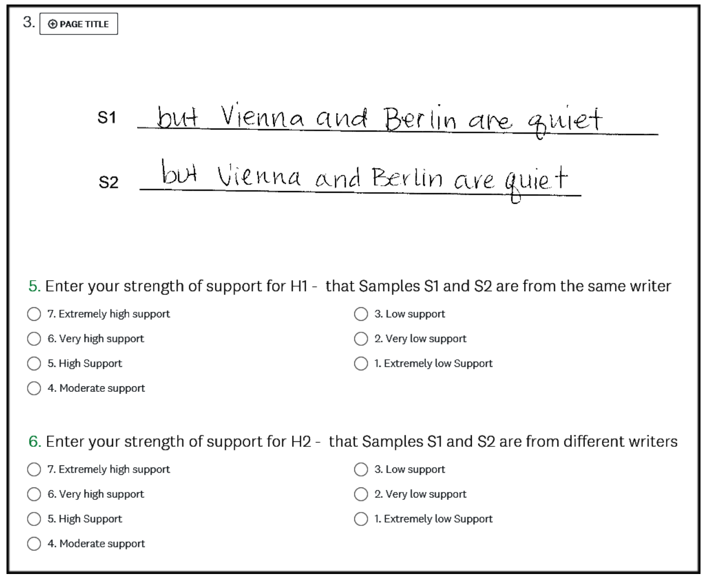
Table 1. The six phrases from the London Letter used in the study
Our London business is good
but Vienna and Berlin are quiet
Mr. Lloyd has gone to Switzerland
and I hope for good news
He will be there for a week
Turin and Rome and will join

Paper and ink copies along with the digitized samples were available for the research. Digitized samples were recorded using Movalyzer (Neuroscript, LLC, Tempe AZ) software. The digitized samples were processed to extract multiple dynamic features from vertical and horizontal strokes. Each page of the hard copy ink samples was scanned at 600 dpi, segmented into individual phrases and saved as separate 16-bit TIFF files. We subjected all possible pairing (for the same phrase and same style) to an automated feature recognition program (FlashID) to rank order the between-writer pairs having lower distance scores and the within-writer pairs having larger distance scores among the thousands of possible pairs. Thirty between-writer pairs and ten within-writer pairs were selected from this larger pool for inclusion in the survey. All remaining pairs were used to calculate the distribution of kinematic feature differences at a population level.

3.2. FDE Opinion Survey

The writership survey was sent to 60 FDEs from North America, Europe, and Australia or New Zealand. The survey consisted of different 40 pairs: 20 print samples (15 between-writer and 5 within-writer) and 20 cursive samples (15 between-writer and 5 within-writer). FDEs were asked to score their strength of support for each of two propositions (for a total of 90 responses). Proposition 1 (H1) pertained to the examiner's strength of support for the proposition that the two samples were written by the same writer (i.e. prosecution hypothesis). Proposition 2 (H2) pertained to the examiner's strength of support for the proposition that the two samples were written by different writers (i.e. defense hypothesis). Examiners indicated their strength of support using a 7-point scale rating from extremely strong support (7) to extremely low support (1). An example of a survey pair with scoresheet is shown in Figure above. For each respondent to the survey, there were 90 scores available for analysis (two from each of 40 sample pairs + 5 repeats).

Figure. A sample scoresheet from the survey.


3. PAGE TITLE

S1 but Vienna and Berlin are quiet

S2 but Vienna and Berlin are quiet

5. Enter your strength of support for H1 - that Samples S1 and S2 are from the same writer

- 7. Extremely high support
- 6. Very high support
- 5. High Support
- 4. Moderate support
- 3. Low support
- 2. Very low support
- 1. Extremely low Support

6. Enter your strength of support for H2 - that Samples S1 and S2 are from different writers

- 7. Extremely high support
- 6. Very high support
- 5. High Support
- 4. Moderate support
- 3. Low support
- 2. Very low support
- 1. Extremely low Support

Forty-one FDEs submitted responses to the survey (68.3% response rate). Six were from North America, 9 from Australia/New Zealand, and 26 from European countries. The average time to complete the 90-item survey was 66 minutes. Five survey items were repeated to examine intra-examiner reliability. Absolute difference scores between the repeated pairs of items were calculated for each of the 41 FDEs. Five FDEs had absolute difference scores greater than 1.5 and were removed from further analyses.

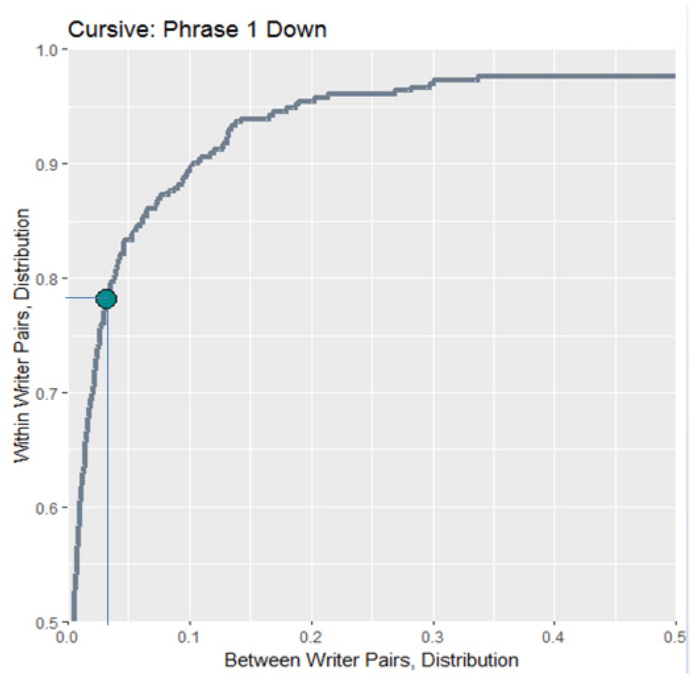
4. Data Analysis

4.1 Calculation of Kinematic Dissimilarity Scores

The dynamic handwriting samples were processed to extract kinematic features from each pen stroke for each sample written by each writer. For a given writing sample we extracted several temporal (stroke duration, velocity, acceleration; pen contact duration, road length), spatial-geometric (stroke amplitude, straightness error, slant), fluency (normalized jerk, number of acceleration inversions) and pen pressure variables for both upstrokes and downstrokes using Movalyzer software.

In order to control the multidimensionality problem necessary to reduce the kinematics to feature sets, we first calculated a kinematic dissimilarity score from the difference between the two members of a sample pair for each kinematic feature. We developed a new dissimilarity score to measure the difference between two documents by writing style (print or cursive). This new dissimilarity score is constructed by first dividing the writing sample into up-strokes and then projecting the feature set into a new dimension that maximizes the separation between the two sets of points; one set for each of the two writing samples being compared. Once we have projected these two sets into the univariate space, we then measure the similarity between the two distributions of points by looking at the integrated squared error difference of the corresponding quantile functions. This procedure resulted in a measure of the dissimilarity between two quantile functions or the Wasserstein distance score (WDS; Del Barrio et al., 1999). An analogous set of steps are then repeated for the down strokes.

Two additional covariates were calculated from the cumulative distribution of the WDS relative to the population distribution (all other between or within-writer WDS not used in the survey). These covariates reflect the probability (ranging from 0-1) that a given WDS calculated from the population of available sample between-writer and within-writer pairs would be less than or equal to the WDS of the survey sample pair and is referred to as the empirical cumulative distribution function or ECDFb and ECDFw, respectively. An example of the ECDF scores for within and between writer distributions for a WDS score of 0.25 is shown in the figure to the right. Higher ECDF values indicate that a large number of WDS in the population of scores are less than or equal to the WDS for the given sample pair. Conversely, an ECDF value near zero indicates that a very small number of WDS in the population of scores are less than or equal to the WDS for the given sample pair. An ECDF near 1.0 indicates that the WDS for the



given pair is in the upper tail of the population distribution, whereas ECDF values near 0 indicates that the WDS for the given pair is in the lower tail of the population distribution. In this example, the ECDFs for within and between population distributions was 0.78 and 0.03, respectively. For this project, two different population distributions (ECDF) served as covariates for modelling FDE response: one for known within-writer scores and another for known between-writer scores (i.e. the alternate population variable).

4.2 Statistical Analyses and Modeling of FDE Response

FDE responses were examined in pair-wise fashion by comparing mean strength score for Hp vs Hd. This was a necessary first step to confirm that FDEs understood the instructions of the survey and provided reasonable responses to alternate propositions. These tests were performed on all 40 pairs from the survey.

Two procedures were conducted to understand the relationship between FDE responses and handwriting kinematic dissimilarity scores. First, univariate correlational analyses were performed to examine the individual contribution of the three kinematic covariates across style and stroke direction. Significant univariate associations would justify the more complex modeling involving multivariate procedures. Multiple linear regression models were used to predict FDE scores reflecting their strengths of support for either the prosecution hypothesis (Hp) that the two samples come from a common source or the defense hypothesis (Hd) that the two samples come from different sources. The WDS or ECDF scores for the temporal and spatial-geometric feature set and style (print or cursive) served as independent predictor variables. For these analyses, we focused on the 10 within-writer pairs, where ground truth supported the prosecution hypothesis that two sets of handwriting have a common source. All statistical analyses were conducted using R software.

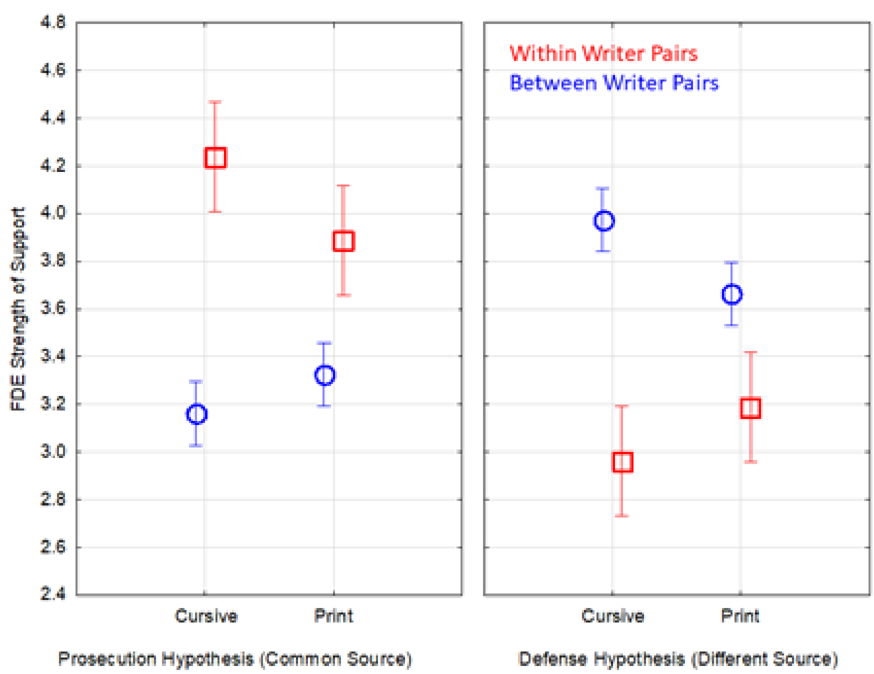
5. Findings

5.1 FDE Support for Alternate Propositions

FDE scores indicating strength of support for two alternate propositions were examined for all survey pairs. Ten were from a common source and 30 from different sources. Post-hoc results from a three-way ANOVA are shown in the Table and Figure below. All pair-wise differences between Hp and Hd reached statistical significance @ $p<0.05$ and are in the expected direction.

	Common Source (within writer)			Different Source (between writer)		
	Cursive	Print	Combined	Cursive	Print	Combined
Hp	4.23 (1.10)	3.89 (0.79)	4.06 (0.87)	3.15 (0.37)	3.32 (0.44)	3.24 (0.41)
Hd	2.96 (0.80)	3.19 (0.56)	3.07 (0.66)	3.97 (0.29)	3.66 (0.45)	3.82 (0.41)

Mean (with SEM bars) FDE strength score in support of the prosecution (single writer) and defense (different writers) hypothesis for 10 within- and 30 between-writer pairs.



For sample pairs from the same writer (common source problem), FDEs offered stronger support for the proposition that the samples were from a common source than when written by different writers and weaker support for the proposition that the samples were from different sources. The opposite was observed for sample pairs from different writers.

5.2 Relationship between FDE response and Kinematic Dissimilarity Scores and Distributions

Univariate correlation analyses were conducted to examine the statistical associations between FDE strength of support for alternate propositions and kinematic dissimilarity scores for the 10 within-writer pairs (5 each for print and cursive). Pearson correlation coefficients are shown in the table below. Coefficients with $p<0.05$ are italicized in red.

	Upstrokes				Downstrokes			
	Hp		Hd		Hp		Hd	
Variable	Print	Cursive	Print	Cursive	Print	Cursive	Print	Cursive
WDS	*-0.91*	-0.15	*0.87*	-0.21	-0.69	-0.29	0.79	-0.06
ECDFb	*-0.96*	-0.33	*0.93*	-0.02	-0.67	-0.63	0.78	0.36
ECDFw	*-0.99*	-0.88	*0.97*	0.65	-0.66	-0.79	0.70	0.65

The univariate results support a strong association between FDE strength of support and kinematic upstroke feature set dissimilarity scores for printed handwriting samples. As support for a common source proposition (Hp) increased, dissimilarity in spatial-geometric kinematic feature set decreased. Coefficients for upstroke cursive samples show a trend toward significant ($p<0.10$) for ECDFw and Hp only, while coefficients for downstroke kinematic features and associations for Hd were statistically non-significant.

5.3 Multivariate Kinematic Models of FDE response

5.3.1 Prosecution Hypothesis

Results from a multiple regression analysis with kinematic feature difference scores and their distribution functions revealed two factors derived from upstroke kinematics accounted for 95% of the variability in FDE support score for the prosecution hypothesis ($R^2 = 0.95$; $F_{2,7} = 67.78$; $p = 0.0001$). Table 3 shows the results of this model. As strength of support that the sample pair came from a common source decreases, the between-writer ECDF scores increases. That is, stronger support for Hp was associated with kinematic dissimilarity scores that are less common among all between-writer comparisons.

Factors	Estimate	S.E.	t-value	p-value
Intercept	9.64	0.50	19.37	<0.0001
$ECDF_w$ for Upstr	-7.07	0.65	-10.81	<0.0001
$ECDF_b$ for Upstr	1.31	0.34	4.01	0.005

5.3.2 Defense Hypothesis

Results from a multiple regression analysis with kinematic feature difference scores and their distribution functions revealed three factors derived from upstroke kinematics accounted for 94% of the variability in FDE support score for the prosecution hypothesis ($R^2 = 0.94$; $F_{3,6} = 29.75$; $p < 0.001$). Table 4 shows the results of this model. As strength of support that the sample pair was written by different writers increases, the between-writer ECDF scores also increase. That is, stronger support for Hd was associated with kinematic dissimilarity scores that are more common among all between-writer comparisons.

Factors	Estimate	S.E.	t-value	p-value
Intercept	-1.45	0.48	-3.03	0.02
Style	-0.38	0.15	-2.53	0.04
$ECDF_w$ for Upstr	6.21	0.69	9.05	0.0001
$ECDF_b$ for Upstr	-2.04	0.36	-5.71	0.001

5.4. Multivariate Modeling of FDE Responses Using Automated Feature Recognition (FlashID)

For this analysis, the feature dissimilarity scores and their ECDFs from an automated feature recognition program were used to model FDE response variability. Unlike the results from the kinematic predictors, the multiple regression models with FlashID scores as independent covariates were not statistically significant. It is likely that the ranking procedures used to select the 40 sample pairings for the FDE survey biased these results. Sample pairs were selected based on FlashID feature dissimilarities in order to challenge the examiners. Specifically, for the within-writer pairs we purposefully selected samples with larger FlashID dissimilarity scores and conversely for the between-writer pairs we selected samples with smaller FlashID dissimilarity scores. Short of selecting sample pairs randomly and conducting a second FDE survey, we decided that the better approach at this stage is to perform an analysis of the relationships between FlashID dissimilarity scores and kinematic WDS for all possible within and between writer pairings. These analyses are currently underway.

5.5 Summary of Findings

For handwriting samples written by a single writer, we found that FDEs offered significantly stronger support for a common source proposition than the alternate proposition that the samples came from different writers. Conversely, for handwriting samples written by different writers, we found that FDEs offered significantly stronger support for a different source proposition than the alternate proposition that the samples came from the same writer. These findings support the validity of the survey instrument as a tool for obtaining writership opinions under two alternate propositions.

FDE strength of support scores were significantly associated with kinematic feature difference scores from handprinted samples under the common source proposition. Lower dissimilarity scores for spatial-geometric kinematic features were associated with increased strength of support for a single writer proposition. These patterns were predominantly observed for upstroke kinematics. The small number of within-writer hand printed samples (n=5) likely contributed to the lack of statistical significance despite the large coefficients for cursive samples. To remedy this limitation, we subjected these data to multivariate analyses.

Results from a multiple regression analysis revealed that the distribution functions derived from population data on kinematic dissimilarity scores accounted for between 94%-95% of the variability in FDE support score for the prosecution and defense hypotheses. The rarity of a given handwriting feature difference score relative to the population of writers (i.e. the distribution function) was a strong predictor of FDE writership opinion within our 2-proposition framework. Style of handwriting was not a significant factor when modeling FDE responses to the prosecution hypothesis; whereas the predictive model for FDE responses to different source proposition included style as a significant factor suggesting that the model held for print and not cursive handwriting.

Spatial-geometric features, particularly for pen upstrokes were found to be highly significant in our predictive models; whereas temporal features, such as stroke duration or velocity were not significant factors in any of the predictive models. This was not unexpected considering that temporal information such as stroke duration cannot be directly measured from static traces and movement velocity can only be inferred.

6. Implications for criminal justice policy and practice in the United States.

This study provides compelling support for the foundational validity grounded in principles of motor control of FDE writership decisions. When experts follow accepted best practices, they compare specific features between questioned and known samples to reach a likelihood or probability opinion about writership. While such comparative methods are largely subjective and depend on training and experience, research supporting the validity of this approach against objective quantitative standards have been lacking. The results of the present study are the first to directly address this scientific limitation. We found that examiner opinions of support for or against a common source proposition can be estimated to a large degree by sets of independent handwriting features derived from dynamic analyses of pen movements. As such, the results of this study provide empirical evidence to support the validity of expert opinions for admissibility in courts of law under *Daubert.*

7. References

Almeciga v Centers for Investigative Reporting. 121 F. Supp. 3d 379 (S.D. N.Y.) 2015

Del Barrio E, Gine,E, Matran C. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. The Annals of Probability 27 (2) (1999) pp. 1009-1071.