Received: 30 June 2021 | Revised: 1 September 2021 | Accepted: 27 September 2021

DOI: 10.1111/1556-4029.14914

JOURNAL OF FORENSIC SCIENCES AAFS

PAPER

Questioned Documents

Elucidating the relationships between two automated handwriting feature quantification systems for multiple pairwise comparisons

Cami Fuglsby MS[1] | Christopher Saunders PhD[1] | Danica M. Ommen PhD[2] | JoAnn Buscaglia PhD[3] | Michael P. Caligiuri PhD[4]

[1]Department of Mathematics and Statistics, South Dakota State University, Brookings, South Dakota, USA

[2]Department of Statistics, Iowa State University, Ames, Iowa, USA

[3]FBI Laboratory, Research and Support Unit, Quantico, Virginia, USA

[4]Department of Psychiatry, University of California at San Diego, La Jolla, California, USA

Correspondence
Michael P. Caligiuri, Department of Psychiatry, University of California at San Diego, La Jolla, CA, USA.
Email: mcaligiuri@health.ucsd.edu

Funding information
This research was supported by the National Institute of Justice, 2017-DN-BX-0148. Cami Fuglsby received additional support from the National Science Foundation, DGE-1828492.

Abstract

Recent advances in complex automated handwriting identification systems have led to a lack of understandability of these systems' computational processes and features by the forensic handwriting examiners that they are designed to support. To mitigate this issue, this research studied the relationship between two systems: FLASH ID®, an automated handwriting/black box system that uses measurements extracted from a static image of handwriting, and MovAlyzeR®, a system that captures kinematic features from pen strokes. For this study, 33 writers each wrote 60 phrases from the London Letter using cursive writing and handprinting, which led to thousands of sample pairs for analysis. The dissimilarities between pairs of samples were calculated using two score functions (one for each system). The observed results indicate that dissimilarity scores based on kinematic spatial-geometric pen stroke features (e.g., amplitude and slant) have a statistically significant relationship with dissimilarity scores obtained using static, graph-based features used by the FLASH ID® system. Similar relationships were observed for temporal features (e.g., duration and velocity) but not pen pressure, and for both handprinting and cursive samples. These results strongly imply that both the current implementation of FLASH ID® and MovAlyzeR® rely on similar features sets when measuring differences in pairs of handwritten samples. These results suggest that studies of biometric discrimination using MovAlyzeR®, specifically those based on the spatial-geometric feature set, support the validity of biometric matching algorithms based on FLASH ID® output.

KEYWORDS
automated handwriting system, black box system, handwriting, questioned documents, statistical modeling, validity, white box system

Highlights

- For cursive and print writing, over 81,000 pairwise scores were calculated for analysis.
- Relationships of feature dissimilarity scores of two automated handwriting systems were assessed.

The preliminary findings were presented at the 71st Annual Scientific Meeting of the American Society of Questioned Document Examiners, August 4–8, 2019, in Cary, NC, and at the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences, February 17–22, 2020, in Anaheim, CA. Statistical aspects of this research were presented at the 2021 Joint Statistical Meetings, August 8–12, 2021, held virtually.

- Relationships of scores based on spatial-geometric and graphical features were significant.
- Statistically significant relationships were observed for print and cursive handwriting samples.
- Construct and convergent validity of the studied handwriting feature systems is supported.

1 | INTRODUCTION

In forensic science, examiner-based black box studies "evaluat[e] the examiners' accuracy and consensus in making decisions, rather than attempting to determine or dictate how those decisions are made." [1] More broadly, an examiner-based black box study is "an empirical study that assesses a subjective method by having examiners analyze samples and render opinions about the origin or similarity of samples" ([2]; p. 48). Typically, the examiner is viewed as a black box, and the aim of the research is to measure the degree to which the output or response from the black box examiner conforms with ground truth. Conversely, white box studies "are detailed assessments of the bases of examiners' decisions, focused not just on the end decisions but the features and attributes used by the examiners in rendering conclusions" [3]. Although the concepts of black box and white box methods of examiner testing in forensic science have become well-known in recent years, black box and white box methods have their roots in computer systems testing. With advances in automated feature recognition systems for forensic science applications, the forensic focus on black box methods should include both machine-based decision systems and human examiners, with increasing emphasis on interpretable artificial intelligence.

Approaches to automated handwriting identification and verification have been developed since the mid-1980s [4]. Several systems have emerged over the years including CEDAR-FOX [5], Forensic Information System for Handwriting (FISH), WANDA [6], and FLASH ID® (Sciometrics, LLC). FLASH ID® is an automated handwriting feature extraction program designed for closed-set identification of writers [7]. FLASH ID® relies on complex algorithms using graph theory to skeletonize and segment handwriting from a scanned document into graphemes (or subgraphs) having nodes and edges. Each grapheme is assigned an "isomorphism class" based on the connectivity structure and a "shape class" based on a set of rules centered on each grapheme's geometry. Each grapheme also has a feature vector of physical measurements within the geometric-spatial domain. Similar to FLASH ID®, other automated systems segment handwriting into smaller pieces in order to extract meaningful measurements from a larger handwriting sample. The responses produced by FLASH ID® involve multiple decisions for segmenting and classifying features based on graphemes, but the precise methods of doing so are not disclosed to the system's users. In this sense, FLASH ID® may be considered a black box evaluative system because the transfer function between input and output response is not transparent.

In contrast, MovAlyzeR® (Neuroscript, LLC) is a program that records and analyzes dynamic pen movements. MovAlyzeR® captures the digitized writing sample and then segments the writing sample into individual strokes based on change in stroke direction; it encodes the on-line pen strokes to generate spatial-geometric and temporal metrics (i.e., kinematics) and pen pressure to characterize the handwritten features. The on-line decoding of pen strokes and reduction of feature metrics by the MovAlyzeR® system is fully transparent to the user and, as such, we considered it to be a white-box evaluative system.

The process of disentangling the inner workings of an automated black box system may not be trivial and, in some cases, the user may only have access to the input objects and their outputs but not complete access to the black box system. Using the inputs, a white box system can deconstruct each object and gain a broader/deeper understanding of the closed black box system. These details may be used to model the black box system and determine if the features measured are significant in predicting the outputs of the black box system. The black box and white box systems chosen for modeling are FLASH ID® and MovAlyzeR®, respectively.

The first goal of this study is to use MovAlyzeR® to elucidate the informative characteristics of a black box automated handwriting feature recognition system used in forensic handwriting comparisons (i.e., FLASH ID®). The second goal is to determine the strength of associations (if any) of feature differences between the two systems for handprinting and cursive styles of handwriting across different features. Finally, the third goal is to provide empirical support for the validity of the two automated handwriting feature analysis systems used.

To accomplish the first study goal, both systems are deployed on the same handwriting sample pairs, and feature dissimilarity scores are calculated and used to evaluate the relationship between these two systems. Specifically, we are interested in determining whether feature differences between two samples of handwriting obtained from a black box automated system are associated with feature differences obtained from a white box automated system.

The second goal is to determine the strength of these associations (if any) for handprinting and cursive styles of handwriting across multiple feature sets. Based on preliminary power studies (see the Appendix) and some knowledge about each system's capabilities, we formed four expectations. First, knowing that FLASH ID® uses a static image, we expect to observe a relationship between FLASH ID® dissimilarity scores and the scores for static spatial-geometric MovAlyzeR® features. Second, as FLASH ID® does not accept dynamic pen features as input, we did not expect to observe a relationship between FLASH ID® dissimilarity scores and scores for the dynamic temporal MovAlyzeR® features. Third, because FLASH ID® uses static images, we did not expect to observe a relationship between FLASH ID® dissimilarity scores and scores for the dynamic pen pressure features from MovAlyzeR®. Fourth, we expected these relationships to hold for both writing styles (i.e., cursive writing vs. handprinting).

There is evidence that both MovAlyzeR® and FLASH ID® are considered valid instruments when applied to their designed purpose. Regarding MovAlyzeR®, support comes from controlled validation studies designed to assess the accuracy of spatial-geometric and temporal kinematic features and pen pressure in distinguishing genuine from simulated signatures [8,9], measuring signature complexity [10] and for distinguishing handwriting samples from two unknown writers [11]. Several studies summarized in Miller et al. [7] support the validity of several versions of FLASH ID®. Walch and colleagues [12] reported performance rates from two experiments of FLASH ID® deployed in a pairwise comparison of topological and geometric classes extracted from handwritten samples. They found 100% correct classification from 194 test documents (100 writers) in the first experiment and 100% correct classification from 590 test documents (300 writers) in the second. Another study by Walch et al. [13] used grapheme-based shape codes processed from 200 test documents to test the performance of FLASH ID®. They reported 99.5% accuracy in correctly identifying same-source documents. These studies motivated the third goal of this study, namely, to provide further empirical support for the validity of MovAlyzeR® and FLASH ID® as measures of handwriting feature and pattern analysis systems. A fundamental principle in scientific measurement validation is that one of the instruments under study exhibits performance characteristics that are consistent with the expected response pattern of the behavior being measured [14]. The third aim extends this principle to forensic measurement validation, as recommended in the PCAST Report ([2]; p. 14) as applied to handwriting feature and pattern analysis systems.

2 | METHODS

2.1 | Study participants and handwriting sample collection

The study recruited 33 volunteer writers from the San Diego Sheriff's Crime Laboratory; each subject was asked to write six phrases from the London Letter [15] and to repeat each phrase five times using both handprinting and cursive writing styles (for a total of 60 writing samples per subject). Handwriting data from these subjects were used in two prior studies aimed at further understanding the decision-making process of forensic document examiners [16,17]. Subjects were asked to write the handwriting sample phrases with an inking pen on lined papers placed on top of a Wacom (Intuos Pro, model PTH-660) digitizing tablet. The stimulus phrase was shown on the top of each page, and repetitions were written vertically, five per page. A total of 1980 separate handwriting samples were collected on both paper (for processing in FLASH ID®) and digital forms (for processing in MovAlyzeR®) from 33 writers. The 60 handwritten samples from each subject collected on paper were scanned to digital format and underwent feature extraction via FLASH ID®, whereas the 60 digital samples collected on the Wacom tablet underwent direct feature extraction via MovAlyzeR®. Then, for any given stimulus phrase and style of writing, the comparison of the features between all pairs of samples resulted in a large set of dissimilarity scores, as described later.

2.2 | FLASH ID® feature dissimilarity scores

For this study, we modified the scoring output (but not the feature extraction) of FLASH ID®, as previously described in Fuglsby et al. [16]. The output of FLASH ID® encodes all the graphemes in a document relative to a reference set of writers (in this case, 50 writers from the "FBI100" data set, described in Saunders et al. [18]; the reference set is a term used in FLASH ID® to denote a list of possible writers of interest for the original recommendation system). The graphemes used for this encoding were derived from a base set of 50 different writers (in this case, the remaining 50 writers from the "FBI100" data set). The FLASH ID® system uses the idea of reward functions to construct an omnibus score for the corresponding recommendation system. We use the idea of a reward function to construct our Vector of Scores (VOS); that is, each grapheme receives a set of rewards based on the recommender algorithm built by the reference set documents (one reward per grapheme for each reference set writer). Although the specific mechanism for assigning rewards is not revealed to the user, it is known that a larger reward indicates a greater similarity of that grapheme to the reference writer's samples (M. Walch, D. Gantz, J. Miller, J. Buscaglia, personal communication, September 8–11, 2009). For each reference writer, these rewards are then summed over all the graphemes in a document, resulting in an omnibus VOS (comparable with the vector of counts method in Gantz et al. [19], for which the rewards are split among a reference set of writers) for each document. Calculating the Euclidean distance between the two VOSs (one per writing sample in a pair) yielded the dissimilarity score between the pair of writing samples. Larger Euclidean distance scores between two VOSs reflect larger feature dissimilarities. This was repeated for all possible sample pairs within a given phrase (from the London letter) and writing style. With 33 writers and five repeats for each of six phrases, this procedure yielded dissimilarity scores for 81,180 possible pairs for each writing style. The structure of this class of score functions leaves much to be desired in terms of how to interpret and explain the resulting dissimilarity.

To the best of our knowledge, the 33 writers who participated in this study are not part of the "FBI100" data set, given that they were collected approximately 15 years apart in different collections. However, as part of our ethical obligation to protect the privacy of study subjects, we could not cross-compare identity between the two groups.

2.3 | MovAlyzeR® kinematic feature dissimilarity scores

Handwriting samples were automatically segmented into upstrokes and downstrokes using MovAlyzeR®. Pen stroke segmentation

points were determined based on the zero-axis crossing of the vertical velocity curve over time. The zero velocity points along the curve reflect a momentary absence of vertical pen movement just prior to a direction change. The segmentation criterion is a user-defined property that was applied to all samples consistently. Several spatial-geometric, temporal, and pressure features were then automatically extracted from each upward and downward pen stroke. The set of spatial-geometric features included vertical and horizontal stroke amplitude, slant, loop surface, and trace length. The set of temporal features included stroke duration, peak velocity, and average velocity. Pen pressure was treated as a third feature set with only a single feature: the average pen pressure during the stroke.

These features characterize handwriting movement in multiple dimensions. The multidimensional kinematic features were transformed into a single score representing the dissimilarity between two handwriting samples, as in Ommen et al. [17]. First, using the kinematic features for all upstrokes in a pair of handwriting samples, a dissimilarity score is constructed by determining the direction of maximum separation by applying linear discriminant analysis (LDA). The LDA method uses this direction to classify each upstroke to either the first or second sample in the pair by providing an estimated posterior probability of belonging to the first sample. For handwriting pairs produced by two different writers, every upstroke from the first sample should have posterior probabilities near one, and all upstrokes from the second sample should have posterior probabilities near zero. For sample pairs produced by the same writer, both samples should have posterior probabilities anywhere between zero and one (depending on the range of natural within-writer variation). Then, the integrated squared error difference of the two quantile functions for estimated posterior probabilities of upstrokes between the pair of handwriting samples is computed. This calculation is a measure of the dissimilarity between two quantile functions and is known as the Wasserstein distance score (WDS) [20,21]. The WDS values range from zero to one, where values near zero indicate more overlap in the posterior probabilities for the two samples, and values near one indicate less overlap. The level of dissimilarity between the measured features of each pairwise comparison is therefore determined by the corresponding WDS value. An analogous set of steps are repeated to obtain kinematic dissimilarity scores for the downstrokes.

2.4 | Regression models of pairwise comparisons

A total of 1980 separate handwriting samples were collected on both paper and in digital forms from 33 writers. Hard copy samples were digitally scanned at 600 pixels per inch (ppi). The MovAlyzeR® feature dissimilarity scores for each pair were used to model the FLASH ID® feature dissimilarity score as follows. Separate simple linear regression models were run for each kinematic feature set (n = 3; spatial-geometric, temporal, and pressure), for upstrokes and downstrokes (n = 2), for each writing style (n = 2; handprinting and cursive), and for each phrase (n = 6) for a total of 72 regression models ($3 \times 2 \times 2 \times 6 = 72$). We established that the large number of potential co-dependences across multi-writer input samples can inflate the Type I error (see Appendix). To minimize the threat stemming from multiple comparisons involving the same writer, we developed a robust statistical approach that takes the comparison/dependence structure into account.

We assume that the collections of writing samples (with one collection per writer) are independent and identically distributed random elements; in effect, we have a simple random sample of writers, and from each writer, we have observed one collection of writing samples. For each of the writing samples, we have measured two sets of features: one corresponding to the FLASH ID® VOS dissimilarity score and a second set of features extracted from the MovAlyzeR® system. We further reduced the features from the MovAlyzeR® system into six sets of subfeatures: spatial-geometric, temporal, and pen pressure feature sets for both upstrokes and downstrokes.

For each of these seven sets of measurements (one FLASH ID® score and six kinematic feature scores), we developed a pairwise dissimilarity score to represent a document-level comparison. Following Ommen et al. [17], the pairwise dissimilarity is computed using a modification of the WDS (see the Appendix for further details). The goal was to create six different regression models to assess the marginal relationship between the MovAlyzeR® features and the FLASH ID® features, where the WDS for one of the six kinematic feature sets is used as the explanatory variable and the FLASH ID® dissimilarity score is used as the response variable. However, this became difficult because the observations (i.e., document-level dissimilarity scores) are not independent, although the assumption of independence is required to perform regression.

When the original set of samples are assumed to be a simple random sample (as in this case), the act of performing pairwise comparisons to produce a score introduces a dependence structure that must be accounted for before any statistical tests can be performed at the desired nominal level. If the full dependency structure (i.e., covariance matrix) is known up to a constant, then the generalized least-squares (GLS) approach can be used. Unfortunately, in this setting, there are three distinct terms that are needed before we can perform a GLS-based analysis. We do have the advantage of being able to solve out for the eigenvectors, but not the eigenvalues, of the pairwise dissimilarity scores covariance matrix. These issues are explored in greater detail in Appendix.

To address the issue of independence, the regression approach was modified. A summary measure was obtained for each pair of writers by averaging their 25 between-writer document-level dissimilarity scores. This resulted in a reduction of the 13,530 document-level dissimilarity scores for each phrase and style of writing to 528 writer-level dissimilarity scores. (See the Appendix for further details.) We performed the modified regression analyses for each of the six phrases, handprinting and cursive separately, and only considered one of the kinematic feature sets at a time. This resulted in a total of 72 tests and corresponding p-values.

3 | RESULTS

Scatterplots with regression lines-of-best fit are shown in Figure 1 for the phrase "Our London business is good" for the set of upstrokes for cursive (top row) and handprinting (bottom row) styles, respectively. The points on the scatterplots represent the average dissimilarity scores across all pairwise comparisons between a pair of writers. Each plot contains 528 averaged dissimilarity scores; for a detailed description of these averaged dissimilarity scores, see Appendix. The red regression lines are fit using the averaged pairwise scores, and the black line is the average of the red lines in each plot. Each plot shows the relationships between individual FLASH ID® VOS dissimilarity scores (*y*-axis) and MovAlyzeR® spatial-geometric, temporal, and pressure feature dissimilarity

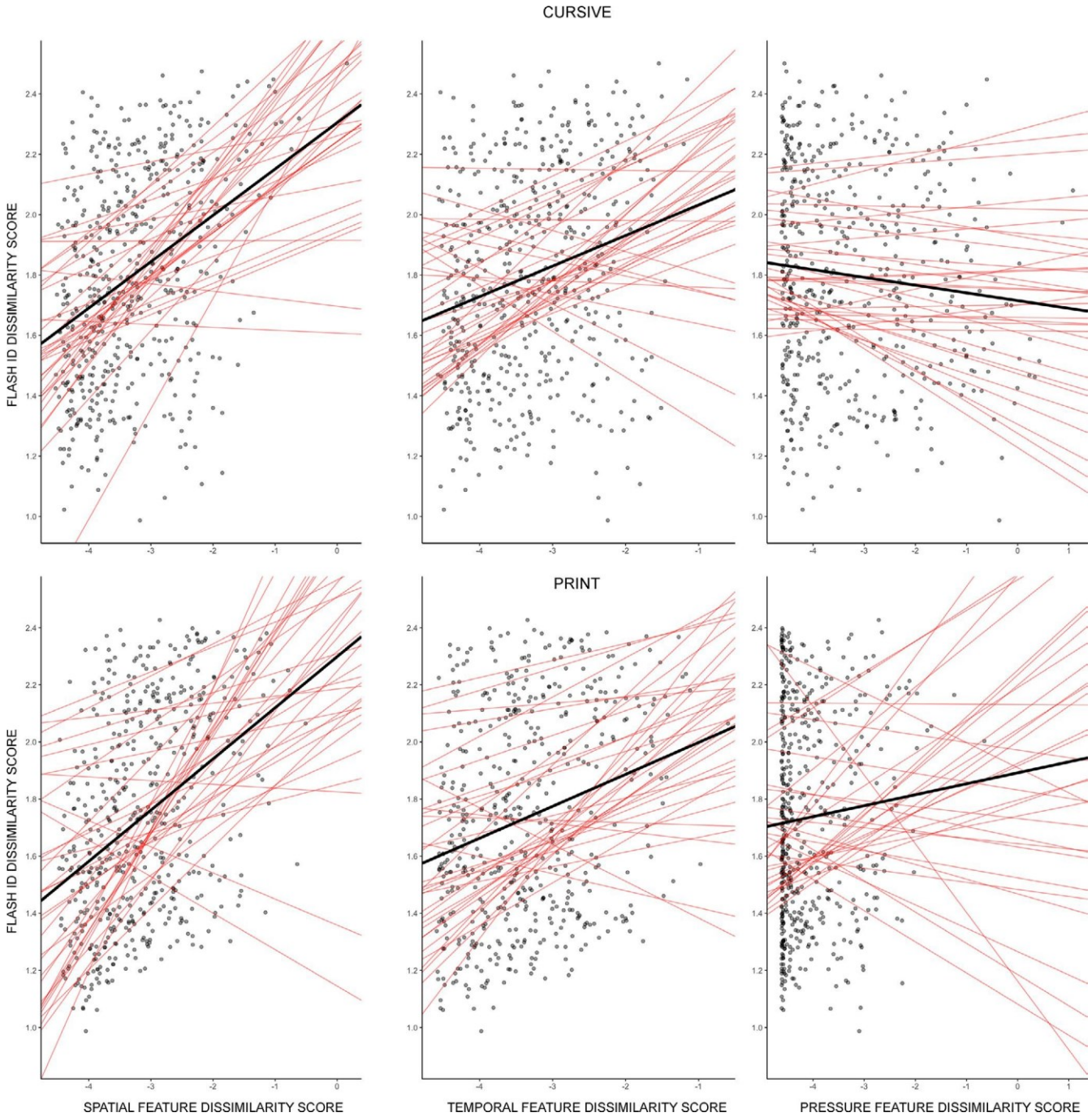

FIGURE 1 Scatterplots with individual (red) and average (black) lines of best fit for cursive (top row) and handprinting (bottom row) handwriting showing the relationship between FLASH ID® dissimilarity score (y-axis) and the dissimilarity scores for spatial-geometric (left), temporal (center), and pressure (right) features for upstrokes for the phrase "Our London business is good." The red regression lines are fit using the averaged pairwise scores—one score per pair of writers, each line representing the 33 scores with one fixed writer for a total of 33 red lines. The thick black line is the average of the red lines in each plot [Color figure can be viewed at wileyonlinelibrary.com]

scores (*x*-axis) for the set of upstrokes for cursive (top row) and handprinting (bottom row) for all possible pairs for this phrase. The more negative the kinematic feature dissimilarity score (along the *x*-axis) is, the less dissimilarity there is in that feature between a given pair of writers.

Inspection of the scatterplots reveals a strong positive relationship for spatial-geometric feature dissimilarities between the two systems. Surprisingly, a modest positive relationship between the FLASH ID® VOS dissimilarity score and temporal feature dissimilarity score from MovAlyzeR® was observed. Lower FLASH ID® VOS dissimilarity scores were associated with lower kinematic spatial-geometric and temporal feature-based dissimilarity scores for cursive samples, whereas only spatial-geometric feature-based dissimilarity scores were significantly associated with FLASH ID® VOS dissimilarity scores for handprinted samples. Similar plots were obtained for downstrokes and for all phrases other than phrase 3.

Results from the regression models for average dissimilarity scores for the relationships between FLASH ID® VOS dissimilarity scores and MovAlyzeR® spatial-geometric, temporal, and pen pressure feature dissimilarities across all pairs of writers for cursive writing and handprinting are shown in Tables 1–3, respectively. Results show that spatial-geometric dissimilarity scores were significant ($p < 0.05$) in predicting FLASH ID® VOS dissimilarity scores for both handprinting and cursive sample pairs as well as upstrokes and downstrokes. The relationships between temporal feature dissimilarity scores and FLASH ID® VOS dissimilarity scores were significant ($p < 0.05$) for cursive sample pairs only, whereas the average pen pressure dissimilarity scores across samples between two writers was not a significant factor ($p > 0.05$) in predicting FLASH ID® VOS dissimilarity scores. With the exception of phrase 3, these patterns were consistent across stroke direction and across the different phrases from the London Letter. Phrase 3 differed from the other 5 phrases as it contains unfamiliar words such as "Mr. Lloyd" and "Switzerland," which may have contributed to greater dysfluencies and subsequently more variability in feature sets across writers as writers self-checked spelling and punctuation of this phrase.

4 | DISCUSSION

In the present study, we expected to observe three patterns. First, we expected that we would observe a relationship between these instruments for spatial-geometric features. We found that dissimilarity scores calculated from spatial-geometric stroke kinematics were significantly associated with dissimilarity scores calculated from an independent, automated feature recognition system in support of our hypothesis. As expected, the relationships between FLASH ID® VOS and MovAlyzeR® dissimilarity scores for spatial-geometric features were generally consistent, regardless of handwriting style. This finding implies that the spatial-geometric features detected and used by the FLASH ID® algorithm in its feature quantification may be robust to writing style.

For our second expectation, we did not expect to observe a relationship between dissimilarity scores produced by FLASH ID® VOS and those produced by kinematic analyses of temporal features. For handprinting, we did not find statistically significant relationships. However, contrary to this, we found significant relationships in the temporal domain for *cursive* handwriting. This is likely due to the well-established relationship between stroke velocity and stroke amplitude for limb movement in general [22] and handwriting specifically [14]. FLASH ID® relies upon complex algorithms to skeletonize and segment writing into graphemes, classify these graphemes using the resulting nodes and edges, and calculate the physical measurements exclusively within the spatial-geometric domain. Although it is a black box system, the static input (i.e., digital scan of a document) contains no temporal components for the algorithms to utilize. Because two of the three parameters that make up the temporal feature set are velocity measures, it is possible that the temporal features were correlated with at least one of the spatial-geometric features driving the FLASH ID®–kinematic relationship. Thus, at least for cursive handwriting, velocity and amplitude are probably not independent features.

For the third expectation, we did not expect to observe a relationship between dissimilarity scores produced by FLASH ID® VOS and those associated with pen pressure. This expectation holds as we did not find any statistically significant relationships. As a static

TABLE 1 Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® spatial-geometric dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes

	Downstrokes				Upstrokes			
	Print		Cursive		Print		Cursive	
Phrase	Slope Coefficient	*p*-value	Slope coefficient	*p*-value	Slope coefficient	*p*-value	Slope coefficient	*p*-value
1	0.259	0.001	0.204	0.003	0.197	0.043	0.155	0.016
2	0.315	<0.001	0.309	<0.001	0.315	0.001	0.283	<0.001
3	0.306	<0.001	0.076	0.174	0.243	0.012	0.062	0.167
4	0.233	<0.001	0.212	<0.001	0.215	0.017	0.172	0.001
5	0.185	0.005	0.250	<0.001	0.030	0.740	0.220	<0.001
6	0.263	<0.001	0.187	0.001	0.250	0.001	0.180	<0.001

TABLE 2 Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® temporal dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes

	Downstrokes				Upstrokes			
	Print		Cursive		Print		Cursive	
Phrase	Slope coefficient	*p*-value	Slope coefficient	*p*-value	Slope coefficient	*p*-value	Slope coefficient	*p*-value
1	0.051	0.528	0.106	0.232	0.113	0.218	0.167	0.057
2	0.094	0.373	0.197	0.043	0.173	0.083	0.332	<0.001
3	0.020	0.857	−0.006	0.910	0.084	0.482	0.039	0.449
4	0.005	0.948	0.166	0.016	0.032	0.762	0.179	0.004
5	−0.063	0.464	0.277	0.002	−0.074	0.381	0.183	0.004
6	0.138	0.174	0.176	0.023	0.134	0.127	0.177	0.004

TABLE 3 Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® pen pressure dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes

	Downstrokes				Upstrokes			
	Print		Cursive		Print		Cursive	
Phrase	Slope coefficient	*p*-value	Slope coefficient	*p*-value	Slope coefficient	*p*-value	Slope coefficient	*p*-value
1	−0.032	0.666	−0.064	0.193	0.087	0.501	−0.062	0.27
2	−0.034	0.645	−0.078	0.138	−0.008	0.952	−0.016	0.813
3	−0.053	0.508	−0.022	0.636	0.092	0.568	0.032	0.530
4	−0.053	0.498	−0.015	0.738	−0.066	0.610	0.012	0.823
5	−0.041	0.631	−0.079	0.119	0.071	0.635	−0.011	0.835
6	0.003	0.972	−0.078	0.133	0.017	0.897	−0.024	0.637

feature encoding system, FLASH ID® was not designed to encode pressure features in handwriting. However, considering that pen pressure often affects line thickness in the static handwriting sample, it is possible that pressure variation could affect the skeletonization and attribution of some grapheme structures in FLASH ID® (e.g., lower case "e" and "i"). Although line thickness can also be impacted by writing instrument (e.g., ballpoint pen vs marker), in the present study, all writers used the same writing instrument.

The kinematic feature dissimilarity scores for upstrokes behaved similarly to downstrokes with regard to their correlations with FLASH ID® VOS dissimilarity scores. This observation is not surprising, given that some of the graphemes used in the FLASH ID® system will contain both upstrokes and downstrokes. Further research may disentangle a stroke–direction effect that this study did not capture. There are strong correlations between the upstroke and downstroke dissimilarity scores (for both spatial-geometric and temporal); thus, seeing the significant p-values of these models with respect to the FLASH ID® VOS dissimilarity scores is not surprising.

The third goal of the present study was to provide empirical support for the validity of two automated handwriting feature analysis systems, MovAlyzeR® and FLASH ID®. Our results support both the construct and convergent validity of MovAlyzeR® and FLASH ID® as instruments capable of detecting differences in handwriting features between two samples written by different writers. The construct itself is a "process or characteristic believed to account for individual or group differences in behavior" ([23]; p. 1) where construct validity refers to how well an instrument measures that behavior or characteristic [24,25]. Handwriting consists of a series of individual pen movements or strokes, each characterized by multiple features in the spatial-geometric, temporal, and pressure domains. These characteristics form the construct used by examiners to understand variability within and across writers. Based on the robust statistical relationships between dissimilarity scores measured by our two instruments, especially in the spatial-geometric domain, we may conclude that both instruments are valid as measures of the construct that handwriting is a series of spatial-geometric parameters or patterns.

Convergent validity reflects the relationship among different measures of the same construct [23]. The present study demonstrated empirically that different measures of the same construct were statistically related. Dissimilarity scores derived from two different approaches to measuring handwriting converged along with some (but not all) features. Specifically, we observed convergence for spatial-geometric features such as vertical and horizontal stroke

amplitude, slant, and trace length; however, such convergence was not observed for pen pressure. Where present, convergent validity held for both handprinting and cursive writing styles.

Within a statistical framework, validity can be defined as the absence of both random and systematic measurement error [14]. Although it is unreasonable to expect the complete absence of random or unexplained error between two independent measurement systems, minimizing systematic error is an attainable goal. Results from the present study demonstrate that there is at least a linear relationship between the FLASH ID® VOS dissimilarity scores and the previously noted subsets of the kinematic dissimilarity scores. In the present study, individual regression models for each of the kinematic feature scores were used, which ignores any possible interactions between the kinematic features. In the future, a single model that incorporates all the kinematic features could be developed using more sophisticated statistical tools. However, before these methods can be applied, they must be fully developed for pairwise comparison data [26].

Last, the guidance document published by the Presidents' Council of Advisors on Science and Technology [2] on ensuring scientific validity of forensic feature comparison methods recognizes a valid scientific instrument as one that "has shown, based on empirical studies, to be reliable with levels of repeatability, reproducibility, and accuracy that are appropriate to the intended application." (p. 48). The PCAST position on scientific validity is that if a measurement of a feature (or in this case, feature-based dissimilarity scores) produced accurate results (based on some accepted standard) and these results can be reproduced, then one can claim that the measurement system is valid within the context of legal discourse. Results from the present study demonstrate the scientific validity that is accepted in legal discourse for our intended application of both MovAlyzeR® and FLASH ID® as biometric verification systems.

Computational algorithms used in proprietary automated forensic biometric identification systems are considered black box systems and, therefore, pose a challenge for proper discovery in the U.S. judicial system. To increase their transparency and interpretability, many have called for the release of algorithm source code, potentially infringing the intellectual property of the algorithm developers. Our approach offers an alternative to the access to intellectual property while addressing the need for transparency and interpretability of such algorithms by developing techniques to characterize the performance of a black box algorithm in terms of a transparent system.

The present research focused on two systems, and any extension of the results of this research to other systems is not warranted at this time. Further research is needed to test whether the correlations observed in the present study between a black box system designed for writer verification and an open handwriting kinematic feature analysis system generalize to other automated systems such as CEDAR-FOX [5] or WANDA [6]. Such studies would strengthen the construct and convergent validity of these and other automated handwriting feature recognition systems.

In conclusion, the present study demonstrated that a white box system has the potential to inform the user of, and to validate, a black box system. Using handwriting data, the results of the testing showed a significant relationship between the FLASH ID® system and the spatial-geometric kinematic features measured by MovAlyzeR®, robust to writing content and writing styles.

ACKNOWLEDGEMENTS

The authors acknowledge Sierra Lutz for her contribution to the handwriting sample formatting, and Brenda Lanners and Gina Hunger for their assistance in collecting the handwriting samples. We wish to thank the volunteers who participated in providing the handwriting sample. The authors would like to acknowledge Sciometrics for providing access to the FLASH ID® system to use for this research.

DISCLAIMER

This is publication number 21–21 of the Laboratory Division of the Federal Bureau of Investigation (FBI). The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. government. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI.

ORCID

Danica M. Ommen https://orcid.org/0000-0001-9955-3817
JoAnn Buscaglia https://orcid.org/0000-0002-8148-2018

REFERENCES

1. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci USA. 2011;108(19):7733–8. https://doi.org/10.1073/pnas.1018707108
2. President's Council of Advisors on Science and Technology (PCAST). Report to the President. Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Washington, DC: PCAST; 2016. Accessed 27 Sept 2021 https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf
3. Hicklin RA, Ulery B, Roberts MA, Buscaglia J. Black box and white box forensic examiner evaluations: Understanding the details. In: Proceedings of the 69th Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2017. p. 480.
4. Plamondon R, Lorette G. Automatic signature verification and writer identification – The state of the art. Pattern Recognit. 1989;22(2):107–31. https://doi.org/10.1016/0031-3203(89)90059-9
5. Srihari SN, Srinivasan H, Desai K. Questioned document examination using CEDAR-FOX. J Forensic Doc Exam. 2007;18:1–20. https://doi.org/10.31974/jfde28-15-26
6. Franke K, Schomaker L, Vuurpijl L, Giesler S. FISH-New: a common ground for computer-based forensic writer identification. In: Proceedings of the Third European Academy of Forensic Science Triennial Meeting; 2003 Sept 22-27; Istanbul, Turkey. Rome, Italy: Eur Acad Forensic Sci. 2003;136(S1-S432):84.

7. Miller JJ, Patterson RB, Gantz DT, Saunders CP, Walch MA, Buscaglia J. A set of handwriting features for use in automated writer identification. J Forensic Sci. 2017;62(3):722–34. https://doi.org/10.1111/1556-4029.13345
8. Caligiuri MP, Mohammed LA, Found B, Rogers D. Nonadherence to the Isochrony Principle in forged signatures. Forensic Sci Int. 2012;223:228–32. https://doi.org/10.1016/j.forsciint.2012.09.008
9. Mohammed L, Found B, Caligiuri M, Rogers D. The dynamic character of disguise behavior for text-based, mixed, and stylized signatures. J Forensic Sci. 2011;56(Suppl 1):S136–41. https://doi.org/10.1111/j.1556-4029.2010.01584.x
10. Angel M, Cavanaugh M, Caligiuri MP. Kinematic models of subjective complexity in handwritten signatures. J Am Soc Quest Doc Exam. 2017;20(2):3–10.
11. Caligiuri M, Mohammed L, Lanners B, Hunter G. Kinematic validation of FDE determinations about writership in handwriting examination: a preliminary study. J Am Soc Quest Doc Exam. 2018;21(1):3–12.
12. Walch M, Gantz D, Miller J, Saunders C, Lancaster M, Buscaglia J. Evaluation of the individuality of handwriting using FLASH ID – A totally automated language-independent system for handwriting identification. In: Proceedings of the 60th Annual Scientific Meeting of the American Academy of Forensic Sciences; 2008 Feb 18-23; Washington, DC. Colorado Springs, CO: American Academy of Forensic Sciences. 2008. p. 388.
13. Walch M, Gantz D, Miller J, Buscaglia J. Evaluation of the language-independent process in the FLASH ID system for handwriting identification. In: Proceedings of the 61st Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2009. p. 381–2.
14. Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. Psychol Rev. 2004;111(4):1061–71. https://doi.org/10.1037/0033-295X.111.4.1061
15. Osborn AS. Questioned documents. 2nd edn. New York, NY: Boyd Printing Co.; 1929. p. 34.
16. Fuglsby C, Saunders C, Ommen DM, Caligiuri MP. Use of an automated system to evaluate feature dissimilarities in handwriting under a two-stage evaluative process. J Forensic Sci. 2020;65(6):2080–6. https://doi.org/10.1111/1556-4029.14547
17. Ommen D, Fuglsby C, Caligiuri MP. Advances toward validating examiner writership opinion based on handwriting kinematics. Forensic Sci Int. 2021;318:110644. https://doi.org/10.1016/j.forsciint.2020.110644
18. Saunders C, Davis L, Lamas A, Miller J, Gantz D. Construction and evaluation of classifiers for forensic document analysis. Ann Appl Stat. 2011;5(1):381–99. https://doi.org/10.1214/10-AOAS379
19. Gantz DT, Miller JJ, Saunders CP, Walch MA, Buscaglia J. New results for addressing the open set problem in automated handwriting identification. In: Proceedings of the 62nd Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2010. p. 431–2.
20. del Barrio E, Cuesta-Albertos JA, Matrán C, Rodriguez-Rodriguez JM. Tests of goodness of fit based on the L2-Wasserstein distance. Ann Stat. 1999;27(4):1230–9.
21. del Barrio E, Cuesta-Albertos JA, Matrán C, Csörgö S, Cuadras CM, de Wet T, et al. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. Test. 2000;9:1–96. https://doi.org/10.1007/BF02595852
22. Viviani P, Terzoulo C. Space–time invariance in learned motor patterns. In: Stelmach GA, Requin J, editors. Tutorials in motor behavior. Amsterdam, Netherlands: North-Holland Publishing Company; 1980. p. 525–33.
23. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. Ann Rev Clin Psychol. 2009;27:1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639
24. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281–302. https://doi.org/10.1037/h0040957
25. Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas: Issues Pract. 1995;14(4):5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881
26. Rosen SL, Saunders CP, Guharay SK. A structured approach for rapidly mapping multilevel system measures via simulation metamodeling. Syst Engin. 2015;18:87–101. https://doi.org/10.1002/sys.21290

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Fuglsby C, Saunders C, Ommen DM, Buscaglia J, Caligiuri MP. Elucidating the relationships between two automated handwriting feature quantification systems for multiple pairwise comparisons. J Forensic Sci. 2022;67:642–650. https://doi.org/10.1111/1556-4029.14914