Moshe Kam,[1] *Ph.D.; Joseph Wetstein,*[1] *B.S.E.E.; and Robert Conn,*[1] *M.S.E.E.*

Proficiency of Professional Document Examiners in Writer Identification

REFERENCE: Kam, M., Wetstein, J., and Conn, R., "**Proficiency of Professional Document Examiners in Writer Identification,**" *Journal of Forensic Sciences,* JFSCA, Vol. 39, No. 1, January 1994, pp. 5–14.

ABSTRACT: A comprehensive writer identification test was designed and administered to a group of professional document examiners and to a control group of nonprofessionals. The professional group consisted of seven document examiners from the Federal Bureau of Investigation. The control group consisted of ten graduate students in the areas of engineering and business. The main finding is that the professional document examiners performed significantly better than the members of the control group. The hypothesis that professionals and nonprofessionals are equally proficient in performing writer identification was found via the Kruskal-Wallis test to have probability of less than 0.001. These findings give indication that handwriting identification expertise indeed exists, and calls into question the conclusions of previous studies in this area.

KEYWORDS: questioned documents, document examiners

Writer screening and writer identification are problems of great interest in law enforcement and paleography [*1–3*]. Several criminal justice systems use extensive analysis of handwritten samples for association of questioned documents with the writers of those documents [*4,5*]. Surprisingly, there are only a few studies that examine the reliability of writer screening by document examiners, a fact noted by Risinger et al. [*6*]. In a 1989 article, Risinger et al. have studied past reports that pertain to document examiner proficiency. On the basis of this study, and without conducting any independent tests, they have questioned the legitimacy of handwriting identification as a law-enforcement discipline. In addition to collecting statistical data from past tests, the main tool that Risinger and his co-workers have used in analyzing past proficiency reports was the *Kruskal-Wallis one-way analysis of variance by ranks.* The Kruskal-Wallis analysis is a test to determine if k samples originated from different populations with respect to averages [*7*]. This test can be used to check whether there are genuine differences between writer identifications that are performed by professional document examiners and those performed by nonprofessionals.

In this paper we revisit the issue of document-examiner proficiency. Unlike Risinger

Received for publication 1 March 1993; revised manuscript received 17 June 1993; accepted for publication 29 June 1993.

This research was funded, in part, by the Federal Bureau of Investigation through contract #J-FBI-91-352.

[1]Associate Professor and Research Assistants, respectively, ECE Department, Drexel University, Philadelphia, PA.

et al., we do not limit our study to the re-examination of past reports. Rather, we have designed and administered a comprehensive writer identification test to professional Federal Bureau of Investigation Questioned-Document Examiners, and to college-educated nonprofessionals (mostly graduate students in engineering and business). In our tests, the professional document examiners performed significantly better than members of the control group. In fact, the hypothesis that professionals and nonprofessionals are equally proficient in performing writer identification was found via the Kruskal-Wallis test to have probability of less than 0.001. For practical reasons, we were restricted to seven professionals in the test group and to ten nonprofessionals in the control group. Although these modest sample sizes may limit the significance of the computed probabilities, the differences in performance between the two groups are striking. These differences indicate that handwriting identification expertise exists, and that the generally negative conclusions of Risinger and his co-workers (cited widely by others in the legal community) may have been premature. Our conjecture is that the negative impressions obtained by Risinger et al. stem from the fact that the tests that they have examined were not based on well designed, consistent, and controlled experiments.

A Summary of Findings by Risinger, Denbeaux, and Saks

In 1989, Risinger, Denbeaux, and Saks [*6*] conducted a comprehensive literature search for empirical evaluations of handwriting identification. They found three published reports [*8–10*], and four studies of the Forensic Science Foundation (FSF) [*11–14*]. The FSF documents describe several different tests administered through the mail from 1984 to 1987 to document-examination laboratories who volunteered to participate.[2] Risinger and his co-workers have analyzed these studies with the following major conclusions (Risinger, pp. 741–749:)

- An early study (Inbau, 1939 [*8*]) had methodological defects that "prevent it from being used as a basis to draw virtually any conclusion."
- A 1973 discussion of document-expert testimony (Todd, 1973 [*9*]) "presents only uncontrolled impressionistic and anecdotal information not qualifying as data in any rigorous sense."
- A Department of Justice test (Peterson et al., 1978 [*10*]) and the FSF studies [*11–14*] "were not presented to control groups of non-experts to determine if the problems presented were too easy." This flaw is considered a "major omission" by Risinger.
- The Peterson et al. study [*10*] and the four FSF studies [*11–14*] reveal that "examiners who returned reports on the analysis disagreed among themselves a good deal of the time suggesting limited reliability, and many of the opinions offered were incorrect suggesting limiting validity." Risinger et al. proceed to attempt different aggregations of the data, arriving at the general conclusion that the laboratories participating in the tests show low proficiency in writer identification.

Our interpretation of the data that Risinger et al. describe and analyze is that the tests that they have reviewed are not very relevant to the question of whether handwriting identification expertise exists. The 1939 test indeed was too flawed to be considered at all. The FSF tests were marred by several meaningful difficulties:

- they changed from year to year in methodology and substance;
- they were conducted in uncontrolled and inconsistent environments;
- they were based on voluntary cooperation;

[2] Since then the FSF has conducted and analyzed another test [*15*], and has conducted, but not yet analyzed, an additional one [*16*].

- they did not use control groups; and
- they often used test problems and samples that professional document examiners considered inadequate for making meaningful decisions.

It is well documented (([*12*], p. 11) and ([*15*], p. 19)) that laboratories that participated in the FSF tests often criticized these tests as invalid. We do not know whether dissatisfaction with the test methodology was the reason why some laboratories that were contacted by the FSF did not return the tests.

If there is a conclusion that can be drawn from the comprehensive literature search performed by Risinger et al. (and from the more recent FSF study [*15*]), it is that good tests for determining the existence or nonexistence of handwriting expertise need to be devised and that there is a lamentable lack of empirical evidence about the subject in the forensic literature. This study is a modest step in addressing this deficiency.

Description of the Writer Identification Test

Database

A database of 86 documents created by 20 writers was used for the test. The database was selected at random from a larger database of 238 documents, created at Drexel University by 45 individuals. These individuals, who were not informed of the purpose of the database, transcribed five samples from a large screen. The texts were selected randomly from books and magazines. A large number of writing utensils were used during the transcription sessions, and writing utensils were often swapped among transcribers. There was no intended deception on the part of the writers (no writer was attempting to forge the writings of another or to disguise his or her own writing during transcription).

Table 1 shows how many samples of each one of the texts were included in the

TABLE 1—*The database.*

	Sample #1	Sample #2	Sample #3	Sample #4	Sample #5	Total	Pen Used[a]
Writer #1	1					1	U
#2	1	1	1	1	1	5	U
#3		1		2		3	D
#4	1	1	1	1	1	5	U
#5	1	1	1	1	1	5	D
#6	1	1	1	1	1	5	U
#7		1		1	1	3	D
#8	1	1	1	1	1	5	D
#9	1	1	1	2	1	6	D
#10	1	1	1	1	1	5	U
#11	1		1	1	1	4	D
#12	1	1	1	1	1	5	U
#13			1	2	1	4	D
#14	1	1	1	1	1	5	U
#15	1	1	1	2	1	6	D
#16	1	1		1	1	4	D
#17	1			1		2	U
#18	1	1		2	1	5	D
#19	1	1	1	2	1	6	D
#20				1	1	2	D
Total	15	15	13	25	17	86	

[a] D = different ball-point and fountain pens were used by the participants during transcription. U = the same blue medium point pen was used by the participants throughout the transcription.

database for each one of the 20 writers. It also gives information about the writing utensils used.

The Test

Each participant was given the same 86 documents and was instructed to sort them into separate piles. Each pile should have included documents created by the same writer. No more than one pile was to be created for each writer.

The participants were not given any information about the structure of the database (such as, number of writers, number of documents per writer, etc.).

The tests were anonymous; no rewards or penalties were involved, and no time limits were imposed. However, administration of the tests through supervisors of both test groups was used to impress on the participants the importance that their respective institutions attach to these experiments. Our impression was that participants in the tests have performed at the peak of their abilities.

The Test Group

The test group consisted of seven professional document examiners, trained and employed by the FBI laboratory in Washington, D.C.

The Control Group

The control group consisted of ten graduate students from the College of Engineering and the College of Business at Drexel University. No member of the control group took part in the preparation of the database.

Evaluation Methods

There are several different ways to assess and score errors in classification of documents. We used two methods to evaluate the participants: the number of *refinement errors*; and the *confusion index* (which is based on the trace of the *confusion matrix*).

Refinement Errors

Two types of refinement errors were defined: *under-refinement* errors and *over-refinement* errors.

Suppose that there were W writers in the data base, and that a document examiner has divided them into P piles. Let w_i ($i = 1, 2, \ldots, P$) be the number of writers whose samples appear in pile i and p_j ($j = 1, \ldots, W$) be the number of piles which contain a sample generated by writer j.

An *under-refinement error* occurs when two documents generated by two different writers are assigned to the same pile.
The number of under-refinement errors is therefore

$$\epsilon_{ur} = \sum_{i=1}^{P} (w_i - 1). \tag{1}$$

An *over-refinement error* occurs when two documents generated by the same writer are assigned to two different piles.

The number of over-refinement errors is therefore

$$\epsilon_{or} = \sum_{i=1}^{W} (p_i - 1). \tag{2}$$

Both ϵ_{or} and ϵ_{ur} are ideally zero, and have an (unachievable) upper bound of $N(M - 1)$. Over-refinement errors indicate that an extra pile has been created for a writer. Under-refinement errors are perhaps more significant, because they indicate that two different writers have been confused. We have used the sum of both types of refinement errors, $\epsilon_{or} + \epsilon_{ur}$, as one measure for writer-screening skills.

Confusion Index

Let D be the total number of documents in the database. In order to construct $C = \{c_{ij}\}$, the confusion matrix for participant k, we define first

$$\hat{c}_{ij} = \frac{\text{number of documents created by writer } i \text{ and assigned by participant } k \text{ to pile } j}{D} \tag{3}$$

$\hat{C} = \{\hat{c}_{ij}\}$ has W rows. If it has less than W columns, we augment it to a $W \times W$ matrix by adding columns of zeros.

We then permute the columns of $\hat{C}$ such that the trace of C, the $W \times W$ square matrix formed by the permuted $\hat{C}$'s first W columns, is maximized. Ideally, for C we should get

$$c_{ii} = \frac{\text{number of documents from writer } i \text{ in the database}}{D}, \tag{4}$$

and

$$\text{trace } (C) = \sum_{i=1}^{W} c_{ii} = 1. \tag{5}$$

We use

$$\rho = 1 - \text{trace } (C) = 1 - \sum_{i=1}^{W} c_{ii} \tag{6}$$

as the *confusion index* of participant k. This confusion index satisfies $0 \leq \rho \leq 1 - \frac{W}{D}$, and the lower the index, the better the classification.

TABLE 2—*Test-group refinement-error scores.*

Participant number	ϵ_{ur}	ϵ_{or}	$\epsilon_{ur} + \epsilon_{o}$
1	1	1	2
2	0	2	2
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
Total number of errors:	1	3	4
Mean number of errors:	0.14285	0.42857	0.5714

Test Results

Tables 2 and 3 show the refinement-error scores for the test group and for the control group, respectively. Tables 4 and 5 show the confusion index scores for the participants in the two groups.

Comparison of the Control Group to the Test Group

To obtain a more direct comparison of our results with those of Risinger et al., we have used the Kruskal-Wallis one-way analysis of variance by ranks to determine whether there exists a statistically significant difference between the test group and the control group ([7], pp. 189–193). Let the null hypothesis be H_0 = {there is no difference between the average error level of professionals and nonprofessionals in writer identification}, and let the alternative hypothesis be H_1 = {there exists a significant difference between professionals and nonprofessionals in writer identification with respect to average error levels}. Each score in the table is replaced by its associated rank for the purposes of this test: the smallest score is replaced by rank 1, the next smallest by rank 2, and the largest by rank H (H = total number of independent observations in the two

TABLE 3—*Control-group refinement-error scores.*

Participant number	ϵ_{ur}	ϵ_{or}	$\epsilon_{ur} + \epsilon_{o}$
8	2	12	14
9	13	18	31
10	2	7	9
11	24	21	45
12	8	19	27
13	2	9	11
14	3	19	22
15	6	38	44
16	3	7	10
17	17	17	34
Total number of errors:	80	167	247
Mean number of errors:	8	16.7	24.7

TABLE 4—*Test-group confusion index.*

Participant number	Confusion index
1	0.034883
2	0.058139
3	0
4	0
5	0
6	0
7	0
Mean confusion index	0.01328

groups). The statistic used in the test is:

$$H = \frac{\dfrac{12}{N(N+1)}\displaystyle\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(N+1)}{1 - \dfrac{\sum T}{N^3 - N}} \qquad (6)$$

Where,

k = number of the groups being compared (in our case, $k = 2$)
n_j = number of members in the jth group (in our case, $n_1 = 7$, $n_2 = 10$)
N = the number of members in all groups combined ($n_1 + n_2 = 17$)
R_j = sum of ranks in jth group
$T = t^3 - t$, where t is the number of tied observations in a tied group of scores
ΣT = the sum of T over all groups of ties

H is distributed as Chi Square with the number of degrees of freedom $df = k - 1$, provided that the sizes of the various groups are not too small (see [7] for additional details).

Using refinement errors as the basis for comparison, $H = 11.974$ ($df = 1$). Using a table of critical values of the Chi Square distribution (for example, [7], p. 249), we find that p, the probability of obtaining this H assuming the null hypothesis, is less than 0.001. With this result, it is necessary to dispose of the null hypothesis and conclude

TABLE 5—*Control-group confusion index.*

Participant number	Confusion index
8	0.186047
9	0.26745
10	0.12791
11	0.313954
12	0.2442
13	0.162791
14	0.27907
15	0.5466
16	0.12791
17	0.2094
Mean confusion index	0.2465

that there is statistical significance to the hypothesis that professional document examiners are better capable of performing writer identification than nonprofessionals.

We repeat the analysis for the *confusion index* scoring. Replacing the confusion index with the associated ranks and applying the test, we discover that $H = 11.9597$ ($df = 1$), and again $p < 0.001$ of obtaining this H under H_0. The same conclusion is reached, we dismiss H_0, and accept H_1. There is a significant performance difference between the test group and the control group.

Some Insights into the Methods Used by Professional Document Examiners

Following our tests, we have interviewed several document examiners at length, as well as nonexperts, to obtain insights into the methodology that experts use in document screening, and in order to use these insights later in developing automatic classification tools. It is very likely that many examiner decisions and associations are difficult to verbalize, and that some verbal explanations are post factum re-creations of the reasoning process. It is nevertheless of interest to discover the framework of document examination through the eyes of its practitioners, and to identify the reasons that nonexperts err where experts perform the correct classification.

When first encountering a document of an unknown writer, (the *questioned document* (*QD*)), the expert often examines it first from a global perspective, then collects "document features," and eventually focuses on specific letters and letter combinations to collect specific "letter features."

When extracting global features, the examiner often determines the skill level of the QD's writer. The skill level is a subjective measure of the artistic sense of the writer, the level of variation in the writing, and the amount of pen control exerted. Experts often show an appreciation for artistic writing, while determining the writer's specific style, which is a counterpart to the skill. The style of a writer is manifested in the use of capital letters, printing, cursive writing, a mixture of printed and cursive letters, types of the connecting strokes, beginning and ending strokes, and the manner in which individual letters are formed. The style and skill are determined by a global search through the document, but not necessarily by specific letter analysis at this stage. The flow of writing along the page, or rhythm, is also an important feature, as are the size and spacing of the words on the page, and the uniformity of the handwriting. These characteristics are used to allow the expert to develop a sense of the style and skill level of the writer.

Examiners are usually looking for unique features. Noticeable are breaks within words, size variations within words, changes from printing to cursive writing, and so forth. Experts emphasize that the ways in which the same letters and letter combinations are formed typically change within a document. The existence of these changes (whether slight or major) contributes to defining a writer. *Variation* is the term used by examiners to describe these changes. Variations are attributes of a writer that experts can identify and nonexperts often gloss over. We found that when pointed out by an expert to a nonexpert, departures from a "norm" become more obvious, but that nonexperts are much less proficient in discovering them.

The examiner usually continues with a closer examination of the document, looking for these variations. Variations will indicate spontaneous writing, and will define the range of expected writing characteristics from the writer.

Next, examiners often pay attention to the spacing, size, slant, shading, speed, shape, and slope of specific letters ("the seven 'S's"). A critical analysis is performed to determine the pen pressure applied, as this often reveals whether the document was freely and naturally prepared (a document that is not freely and naturally prepared may be a forgery or disguised writing). It is at this stage that magnification and extra lighting may be most useful.

During a detailed analysis of the text, there is a deliberate attempt to note and record all existing features. Letter dotting and crossing are noticed, every letter is analyzed, and variation is determined by comparing letters and letter combinations to identical letters and combinations in the same locations within words. Comparisons are made of letters and letter combinations which occur at the same place within a word (either the beginning, the middle, or the end). Additionally, expert document examiners have the ability to find correlations between nonadjacent portions of text, sometimes correlations between letters that are several lines apart. Indeed, some professional examiners are capable of discovering high-order correlations between letters and word-fragments in different areas of the handwritten note. If this capability was automated, it would require an inordinate computation time due to combinatorial explosion of the correlation calculations. The ability of trained examiners to perform such correlations and to use them to make decisions is tied to another ability that examiners possess; namely, the ability to select (in about 1 minute per a 100-word document) "meaningful" patterns over "meaningless" patterns to correlate and analyze.

Finally, it is interesting to note that, when comparing different documents for association, expert document examiners are often looking first for evidence that the two examined documents emanated from two *different* writers, while nonprofessionals often concentrate on *similar* characteristics first. Attention to variations and ability to perform high-order correlations were almost completely absent in the nonprofessional group.

Conclusion

Using a proficiency document examination test, we provide indication that professional document examiners from the Federal Bureau of Investigation are significantly better in performing writer identification than college-educated nonexperts. Using standard hypothesis-testing statistical tools, the hypothesis that professionals and nonprofessionals are equally proficient in performing writer identification was found in our test to have probability of less than 0.001. These findings give indication that handwriting identification expertise indeed exists, and that the generally negative conclusions about this issue by Risinger et al. [6] may have been premature.

Acknowledgments

We would like to thank the reviewers for many insightful and useful comments. These have significantly improved our manuscript.

We are grateful to the Document Examiners at the Federal Bureau of Investigation for the many hours of instruction and interviews that they have provided us.

References

[1] Krantz, K. A., "Handwriting Exemplars," *Naval Law Review*, 1985, pp. 185–199.

[2] Plamondon, R. and Lorette, G., "Automatic Signature Verification and Writer Identification—The State of the Art," *Pattern Recognition*, Vol. 22, No. 2, 1989, pp. 107–131.

[3] Dinstein I. and Shapira, Y., "Ancient Hebraic Handwriting Identification with Run-Length Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 12, No. 5, 1982, pp. 405–409.

[4] ICDAR 91, *Proceedings of the First International Conference on Document Analysis and Recognition* (September 30–October 2, 1991), Saint-Malo, France: IRISA-INRIA.

[5] Cardot, H., Revenu, M., Victorri, B., and Revillet, M-J., "Coopération de Réseaux Neuronaux pour l'authentification de signatures manuscrites," *Proceedings of Neuro-Nimes 1991.*

[6] Risinger, D. M., Denbeaux, M. P., and Saks, M. J., "Exorcism of Ignorance as a Proxy for Rational Knowledge: the Lessons of Handwriting Identification 'Expertise,' " *University of Pennsylvania Law Review*, Vol. 137, 1989, pp. 731–787.

[*7*] Siegel, S., *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, 1956, pp. 184–194.

[*8*] Inbau, F., ''Lay Witness Identification of Handwriting,'' *Illinois Law Review*, Vol. 34, 1939, p. 433.

[*9*] Todd, I., ''Do Experts Frequently Disagree?,'' *Journal of Forensic Science*, Vol. 18, 1973, p. 455.

[*10*] Peterson, J., Fabricant, E., and Field, K., Crime Laboratory Proficiency Testing Research Program: Final Report, *U.S. Government Report* (*U.S. Department of Justice*), 1978.

[*11*] Collaborative Testing Services, Inc., Crime Laboratory Testing Program, Rep. No. 84-7, 1984, Questioned Document Analysis.

[*12*] Collaborative Testing Services, Inc., Crime Laboratory Testing Program, Rep. No. 85-8, 1985, Questioned Document Analysis.

[*13*] Collaborative Testing Services, Inc., Crime Laboratory Testing Program, Rep. No. 86-5, 1986, Questioned Document Analysis.

[*14*] Collaborative Testing Services, Inc., Crime Laboratory Testing Program, Rep. No. 87-5, 1987, Questioned Document Analysis.

[*15*] Collaborative Testing Services, Inc., Crime Laboratory Testing Program, Rep. No. 89-5, 1989, Questioned Document Analysis.

[*16*] Collaborative Testing Services, Inc., Crime Laboratory Testing Program, Rep. No. 92-6, 1992, Questioned Document Analysis.

Address requests for reprints or additional information to
Moshe Kam, Ph.D.
ECE Department
Drexel University 7-412
Philadelphia, PA 19104