Parsed

professionals were statistically different from the matching decisions made by the nonprofessionals. We have established this result with respect to five scoring criteria for the data and two types of nonparametric statistical procedures: (i) procedures based on *ranks* of scores, and (ii) procedures based on *distributions* of scores.

• The group of professionals incorrectly matched 6.5% of the documents in the *unknown* packages with documents in the *database* packages. The group of nonprofessionals made this mistake for 38.3% of the documents in the unknown packages.

• Nonprofessionals in our test tended to "over-associate" indiscriminately. As a result they found as many correct matches as the professionals did—but have declared many non-matching pairs to be matches.

• There were no significant statistical differences among the samples that came from the three professional groups.

• The sample generated by the trainees was significantly different from the sample generated by the professionals.

• The sample generated by the trainees was significantly different from the sample generated by the nonprofessionals.

The Test-Takers

The professional test-takers were either currently-employed or recently-retired professional questioned-document examiners, employed by law enforcement agencies or in for-profit private practice. The overwhelming majority of professional test-takers (99 out of 105) responded in full to voluntary questionnaires attached to the tests, identifying themselves by name, describing their education and training, and providing detailed accounts of their professional histories and professional duties. Almost all of the professional test-takers that answered our questionnaires were certified by, or were members of, one or more of the following organizations: American Academy of Forensic Sciences—Questioned Documents Section; American Board of Forensic Document Examiners; Southeastern Association of Forensic Document Examiners; Southwestern Association of Forensic Document Examiners; and the American Society of Questioned Document Examiners. Members of the Northeastern group were individually invited to the exam using lists of examiners residing in the vicinity of New York and Washington DC who are members of these professional organizations. Members of the Southwestern and Southeastern groups were attendees of the May 1996 professional meetings of questioned-document examiner regional associations.

The non-professional test-takers were students and educators from the Greater Philadelphia area, 34 of them holders of college degrees (B.A., B.Sc., M. A., M.Sc., M.Ed., MBA, Ph.D.) in Engineering, Education, or Management. The other seven were senior undergraduate students in Engineering. The non-professionals were screened for education level in order to match the educational profile of the professional groups. To induce their best efforts, non-professional test-takers were paid for their time, and were rewarded with financial performance-based incentives. Test-takers received $25 for participation and $25 for each correct match. We subtracted $25 for each incorrect match, and $10 for each missed match. If the resulting payment was less than $25, the participant received $25. The average test that we administered to nonprofessionals would have netted a payment of $102 for a perfect score (plus $25 for participation); some of our tests could yield a payment in excess of $200.

The Documents

Data Collection

We have collected 1800 original handwritten documents generated by a group of 150 writers of ages 20–27, working on wide and well-lit tables in a classroom setting. Each writer generated 12 documents on $8^1/_2$ by 11 in., 20 lb. white pages, copying three given texts four times each. We have supplied the paper and the writing utensils, blue and black medium-tip Bic pens. We instructed the writers to switch pens every 2–3 documents so that both 'blue' and 'black' documents be created.

The following texts were used:

1. The claim of their lawyers was simple. No one got hurt, the police framed the alleged stick up perpetrators, the dye packs stopped the defendants, and the alarms went off before any dollar bills changed hands.
2. No, these were not the people who would commit a robbery, use bombs or guns. But in their lust for money these administrators have done even worse. They have betrayed the public trust.
3. His progress was slow, but at the end his persistence paid off. Many of his classmates were deemed brighter, more promising. But at the end of the day, he has surpassed them all, using the most potent weapons—dogged pursuit; eyes always on the prize; nobody, nothing ever capable of throwing him off the track or dampening his spirit.

Data Organization

We have created five 360-document sets—each was generated by 30 distinct writers. Each set was labeled by a letter *A* through *E*.

The 360 documents of set *A* were assigned random numbers; a key, associating the random numbers with their writers, was created and secured. Twelve packages of six documents were created and marked *unknown* (thus creating *unknown A1*, *unknown A2*, ... , *unknown A12*). The *unknown* documents in each package were selected randomly from set *A*, under the condition that six distinct writers be represented in each *unknown* package. The remaining 288 documents were randomly grouped into twelve *database* packages with 24 documents each (thus creating *database A1*, *database A2*, ... , *database A12*).

This procedure was repeated for sets *B*, *C*, *D*, and *E*. Every test-taker was provided with an *unknown* package and a *database* package, both from the same 360-document set (e.g., *unknown D3—database* D12).

Security

In order to frustrate hypothetical (and highly-unlikely) individual attempts to use recorded answers from one test in a later test, we have changed the pairing of tests in every session. Thus if the pairing *unknown A1—database A1* was used in New York, it was never used in Washington, Atlanta, or Reno. Every time *unknown A1* was used again, it was paired with another *database*-set from the *A* group. Even if the correct results of an early test were fully known to all test-takers in a later session, this information was practically useless. We do not believe that any attempt was made to record or share results from our tests between test-takers. However, even if such attempts were made, they could not affect the results in a meaningful way.

TABLE 1—*In* p *percent of the tests,* m *documents in the* unknown *package had matches in the* database *package.*

m Unknown Documents	p Percentage
0	0.00
1	3.47
2	15.97
3	36.11
4	27.09
5	15.97
6	1.39

TABLE 3—*The percentage* p *of tests that had* m *matching pairs.*

m Matching Pairs	p Percentage
0	0.0
1	2.8
2	6.9
3	18.0
4	21.5
5	22.9
6	16.7
7	4.9
8	3.5
9	1.4
10	1.4

The processing and preparation of the document packages and the solution key were executed by a team headed by the second author. This team did not include the first author who administered the tests in New York, Washington DC, Reno, and Atlanta. The solution key (that associated the document identification number with its writer) was secured by the second author in Philadelphia. It was not shipped or communicated to any other individual or group, including the first author, until the tests were completed. The key was not available physically or electronically to anyone in New York, Washington DC, Reno, or Atlanta, including the first author who conducted the tests there.

Statistics on the Documents

Every test-taker received an *unknown* package of six documents and a *database* package of 24 documents. We have compiled a set of statistics about all possible pairings of tests. These statistics are: the percentage of tests for which *m unknown* documents had matches, the percentage of *unknown* documents with *m* matches, and the total number of matches per *unknown* package.

Table 1 shows the percentage *p* of tests where *m unknown* documents had matches (here *m* = 0, 1, 2, 3, 4, 5, 6). The third row of the table reads: "in 15.97% of the tests exactly two of the six documents in the *unknown* packages had a match or matches in the corresponding *database* packages".

The table shows that each test-taker was expected to find matches for at least one document. The most likely number of *unknown* documents in a test for which matches existed was three.

Table 2 shows the percentage *p* of *unknown* documents in our tests that had *m* matching documents in the *database* packages (where *m* in our tests ranged from 0 to 4). The third row of this table reads: "13.3% of the documents in the *unknown* packages had exactly two matches in the corresponding *database* packages".

Finally, Table 3 shows the percentage *p* of tests that had *m* matching pairs of documents in the *unknown* and *database* packages. The third row of this table reads: "in 6.9% of the tests, the test-taker should have found exactly two matching pairs".

The Task

All the tests were administered personally by the first author. He provided written instructions for each test-taker and repeated them orally before the test was taken. Test-takers received a description of the two document packages, including the statement: "these documents were obtained from willing participants who wrote naturally in a well-lit environment. To the best of our knowledge, no forgeries or disguised writings are included".

The task was described as follows: "Decide for each of the *unknown* documents whether or not it has a match in the *database* document package, and list the matches by their code numbers. By "match" we mean that you can state that the two documents were generated by the same writer using the term "identification" or "strong probability". (We use the nine-point ASTM standard E1658 scale of identification/strong probability/probable/indications/no conclusion/indications did not/probably did not/strong probability did not/elimination.) We ask you to tell us whether you made an "identification" or "strong probability" conclusion about each match". The test was designed for a period of 3 hour but no test-taker required that long. Table 4 shows the format of the answer sheet.

A copy of the ASTM standard was available in the room where the exam was given. Instructions for professional and nonprofessional test-takers were the same, except that the first author has explicitly explained orally to each nonprofessional participant the ASTM scale and the terms "identification" and "strong probability".

The format of the test was selected to resemble a multi-suspect case in which extensive examination of documents was required. The test was designed to require a large number of comparisons

TABLE 2—*The percentage* p *of* unknown *documents in our tests that had* m *matching documents in the* database *packages.*

m Matches	p Percentage
0	43.3
1	40.3
2	13.3
3	2.8
4	0.3
5	0.0

TABLE 4—*Answer sheet format for the test.*

Code # of Unknown Document:	Code Number of Documents that match the sample document, please indicate whether strong probability (S.P.) or identification (I.D.):		
1: 111-AAA	234-BBB (S.P.)	____	( )
	567-C1A (I.D.)	____	( )
	____ ( )	____	( )
	____ ( )	____	( )
	____ ( )	____	( )
2: 789-A29	no matches ( )	____	( )
	____ ( )	____	( )

on relatively short texts. In about $^1/_3$ of the cases, test-takers had to study document pairs that carried the same text, and in the remaining $^2/_3$ of the cases, document pairs that carried different texts.

Performance Results

In this section we present statistics on the results from the tests. In the subsequent sections, we compare the data collected from professionals, nonprofessionals, and trainees to determine whether or not they came from genuinely different populations. We present three performance indices. The first performance index is a pair of detection/false alarm probabilities of the {Type I; Type II} structure used in hypothesis testing (14,19). The second is the average number of wrong matching declarations. The third is a Bayesian cost, representing financial reward for performance.

Performance Index 1: Hit Rate and Wrong Association Rate

This criterion is a *pair* of probabilities:

• probability that a match was declared given that a match existed, (this is $P(D_1|H_1)$—the probability of match detection) and,
• probability that an *unknown* document was wrongly matched to a *database* document (wrong association rate, *w.a.r.*; the probability of document-association *false alarm*).

These two probabilities are closely linked and are presented in Table 5 together.[3]
We show $P(D_1|H_1)$, the probability of match detection in two ways:

(i) Group hit rate $= \dfrac{m}{n}$,

where

m = number of correct matches declared by the tested group, and
n = number of matches that existed in the group's tests;

(ii) Average hit rate $= \dfrac{1}{N} \sum_{i=1}^{N} \dfrac{m_i}{n_i}$,

where

N = total number of test-takers,
m_i = number of correct matches declared by test-taker i, and
n_i = number of matches that existed in test-taker i's test.

We calculate the *w.a.r.* as follows:

$w.a.r. = \dfrac{1}{N} \sum_{i=1}^{N} \dfrac{m_i}{n_i}$,

where

[3]The higher the probability of detection (which is desirable), the higher the probability of false alarm (which is undesirable). It is easy to achieve perfect performance in either one of these probabilities-but not in both (by declaring that all *unknown* documents match all the *database* documents one gets $P(D_1|H_1) = 1$, but then *w.a.r.* = 1; by declaring no matches at all, *w.a.r.* = 0, but then $P(D_1|H_1) = 0$ as well.

TABLE 5—*Hit rate and wrong association rate.*

Group	Hit Rate Group	Average	*w.a.r.*
Professionals-Northeast	0.883	0.891	0.064
Professionals-Southeast	0.876	0.873	0.059
Professionals-Southwest	0.878	0.852	0.072
All professionals	0.879	0.871	0.065
Non-professionals	0.877	0.875	0.383
Trainees	0.878	0.852	0.083

N = total number of test-takers,
m_i = number of unknown documents wrongly associated by test-taker i, and
n_i = number of unknown documents examined by test-taker i.

The *group hit rate* and the *average hit rate* should be very close in groups that are homogeneous with respect to detecting a match. The ideal value for both is 1.00. The *wrong association ratio* should be 0.00 ideally. In the context of our test, reluctance to match (manifested by hit rates less than 1.00) is less serious than engagement in wrong matchings (manifested by *w.a.r.*s. greater than zero). Failure to detect a match (type II error) can be described as overcautious. Declaration of false matches (type-I error) is much more dangerous because it may link a person to an incriminating document that s/he did not create. Table 5 shows the group hit rate, average hit rate, and wrong association rate for the tested groups.

These are very interesting results. All groups have roughly the same rates of finding a match when one exists. However, the nonprofessionals are grossly over matching. They erroneously match *unknown* documents with *database* documents at very high rates.

In fact, we have created here a Neyman-Pearson test with constant detection rate (14), and found that nonprofessionals are 6 times more likely than professionals to match two documents that were created by different writers.

Performance Index 2: Absolute Number of Wrong Associations

The first performance index counted how many *unknown* documents were mismatched (i.e., were paired at least once with a document not created by the same author). Thus, if one *unknown* document was mismatched twice, we said there that *one unknown* document was mismatched. In this section we count wrong matching declarations. If one *unknown* document was mismatched twice, we count here *two* wrong matching declarations. We show the average number of wrong matching declarations made by test-takers in Table 6.

The nonprofessionals are very distinct, making as many as 10 times more wrong matching declarations than the professionals.

TABLE 6—*The average number of wrong matching declarations per test-taker.*

Group	Average Number of Wrong Associations
Professionals-Northeast	0.59
Professionals-Southeast	0.35
Professionals-Southwest	0.57
All professionals	0.50
Non-professionals	5.85
Trainees	0.50

TABLE 7—*Performance-based payment (Bayesian cost).*

Group	Payment		Earning Ratio
	Actual	Perfect	
Professionals-Northeast	74.41	106.62	0.70
Professionals-Southeast	79.26	106.62	0.74
Professionals-Southwest	87.30	122.30	0.71
All professionals	80.52	112.14	0.72
Non-professionals	−61.88	101.88	−0.61
Trainees	93.75	128.13	0.73

Performance Index 3: Financial Reward for Performance (Bayesian Cost)

In the actual test nonprofessionals were paid $25 for participation, and $25 for each correct match. The sum of $25 was subtracted for each incorrect match, and $10 was subtracted for each match that was missed. If the final payment was less than $25, the participant received $25.

Table 7 shows what would have been paid to the participants if the guaranteed $25 minimum payment were eliminated. This index is a particular realization of a Bayesian cost (14) with $C_{10} = -25$, $C_{01} = -10$, $C_{11} = +25$, and $C_{00} = 0$. The first column shows the average payment for actual performance. The second column shows the average payment for perfect performance. The third column shows the ratio of the two.

As the table demonstrates, the professionals have "earning ratios" exceeding 0.7. The nonprofessionals, under this scheme, would have to *pay* a penalty.

Time Spent on the Test

Table 8 shows the average time period spent on the test and the standard deviation of this period. Nonprofessionals spent the least amount of time, followed by trainees and professionals. All participants in our test had strong incentive to test at peak performance, and all were told that they can take up to 3 hours to complete the test. No one required that long. We conclude that the test-takers saw no reason to use more time than they have actually used under the conditions of the test. We applied standard tests to seek correlations within groups between time spent on the test and performance, but no significant correlations were found.

Comparison Criteria

The dictionary's definition of an expert is "a person who has special skill or knowledge in some particular field" (15). Skill or knowledge in experts is "special" only if it does not exist in the general population. In this context, it is important to compare

TABLE 8—*Average time spent on the test and the standard deviation (hours:minutes).*

Group	Time Spent	
	Average	Std. Dev.
Professionals-Northeast	1:23	0:32
Professionals-Southeast	1:34	0:36
Professionals-Southwest	1:38	0:20
All professionals	1:32	0:30
Non-professionals	0:58	0:24
Trainees	1:16	0:43

samples created by the professionals and nonprofessionals, and determine whether or not they came from different populations with respect to handwriter-identification abilities.

Statistical Tests

The literature (16–21) offers a number of statistical tests for comparing samples, each relying on its own set of assumptions regarding sample size and statistical distributions of the data. Our study requires tests that compare data from two groups (e.g., experts vs. non-experts) and data from $k(k \geq 3)$ groups (e.g., data from the three groups of experts). In each case we use a test on *distributions* (of the Kolmogorov-Smirnov type), and a test on *locations* (of the Mann-Whitney type). Examination of the data that we compare tends to favor the distribution tests. We include location tests for completeness because some relevant past studies have used them. Four statistical tests were used; the first two are distribution tests, the other two are location tests.

Distribution Tests of the Kolmogorov-Smirnov Type

The Kolmogorov-Smirnov (KS) two-sample test (16,17) was used to decide whether or not two independent samples have been drawn from the same population (or from populations with the same distribution). This two-tailed test is "sensitive to any kind of difference in the distributions from which the two samples were drawn" (16). It uses the statistic

$$D = \max_{x} |S_1(X) - S_2(X)|,$$

where

$S_1(X)$ = cumulative step function for the first sample which has n_1 points. It takes a value of K/n_1, where K is the number of scores equal to or less than X, and

$S_2(X)$ = cumulative step function for the second sample which has n_2 points. It takes a value of K/n_2, where K is the number of scores equal to or less than X.

For values of n_1 and n_2 such that $n_1 + n_2 > 35$, and a significance level of 0.05, the critical value of D is given by $D_c = 1.36\sqrt{(n_1 + n_2)/(n_1 n_2)}$ (17). If D is greater than D_c then we reject the hypothesis H_0 that the two samples were drawn from the same population. Otherwise, we accept this hypothesis. Alternatively, we may calculate the probability p of obtaining, under hypothesis H_0, a more extreme value of D than the observed value of D (18).

The Birnbaum-Hall (BH) k-sample test (20,21) was used to decide whether k independent samples have been drawn from populations with the same distribution. It uses the statistic

$$T = \max_{i,j,X} |S_i(X) - S_j(X)|,$$

where

$S_i(X)$ = cumulative step function for the sample i (calculated as described in the Kolmogorov-Smirnov test), and

$S_j(X)$ = cumulative step function for the sample j.

This statistic is the maximum distance between any pair of the k distribution functions. The probability p of obtaining, under

hypothesis H_0, a larger value for the statistic than the observed value is calculated by the iterative scheme described in (21).

Location Tests Based on Ranks

The rank test of Mann and Whitney (MW) (17) was used to test whether the populations of two independent samples differ with respect to their means. The test uses the statistic

$$U = \min(U_1, U_2),$$

$$U_1 = mn + \frac{m(m + 1)}{2} - R_1,$$

$$U_2 = mn + \frac{n(n + 1)}{2} - R_2,$$

where

m, n = number of elements in samples 1 and 2 respectively,
R_1 = sum of the ranks in group 1, and
R_2 = sum of the ranks in group 2.

The following equation is used to adjust for ties.

$$\hat{U} = \frac{\left| U - \frac{mn}{2} \right|}{\sqrt{\left[\frac{mn}{N(N - 1)} \right] \cdot \left[\frac{N^3 - N}{12} - \sum_{i=1}^{T} \frac{(t_i^3 - t_i)}{12} \right]}}$$

N = total number of elements, $N = m + n$,
T = the total number of ties observed, and
t_i = number of equal ranks in the ith tie.

For large sample sizes, $\hat{U}$ is approximately normally distributed. For a significance level of 0.05, we accept H_0 if $\hat{U} < 1.96$, otherwise we reject H_0. We may calculate the probability p of obtaining, under hypothesis H_0, a more extreme value of $\hat{U}$ by using tables of the normal distribution.

The Kruskal-Wallis (KW) one-way analysis of variance by ranks (16,17) was used to decide whether $k \geq 3$ independent samples are from different populations with respect to means. "The Kruskal-Wallis technique tests the null hypothesis that the k samples come from the same population or from identical populations with respect to averages" (16). The Kruskal-Wallis statistic is

$$H = \frac{\frac{12}{N(N + 1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N + 1)}{1 - \frac{\sum_{i=1}^{T} (t_i^3 - t_i)}{N^3 - N}},$$

where

k = number of the groups being compared,
n_j = number of members in the jth group,
N = the sum of the members in all groups. $\Sigma_{j=1}^{k} n_j$,
R_j = sum of ranks in the jth group,
T = the total number of ties observed, and
t_i = number of equal ranks in the ith tie.

When the sample sizes are sufficiently large $(n_j > 5)$, H is distributed as chi square with $df = k - 1$ degrees of freedom. At a significance level of 0.05, the critical value H_C for $k = 3$ is 5.99 and for $k = 4$, $H_C = 7.81$. If $H > H_C$, we reject the hypothesis that the samples are from the same population. Alternatively, we may calculate the probability p, under hypothesis H_0, of obtaining a larger value for the statistic than the observed value of H by using chi-square distribution tables.

Scoring Criteria

We have used five scoring criteria to compare the data obtained from the tested populations. These criteria contain and expand upon the performance indices described earlier.

Criterion 1: Number of Wrongly Associated Documents—We have assigned to each test-taker a score based on the number of *unknown* documents that s/he has associated wrongly with *database* documents. The score 0 was given to examiners that associated no *unknown* document wrongly with a *database* document. The score 6 was given to examiners that wrongly associated all six *unknown* documents with *database* documents. This scoring method is in essence the *w.a.r.*

Criterion 2: Probability that an Association Declaration is Correct—We have scored all test-takers in terms of the ratio of correct-matching decisions to the total number of matching decisions $(P(D_1|H_1))$.

Criterion 3: Hit Rate—We scored individuals by hit rate = m/n, where m = number of matches declared by the individual, and n = number of matches that existed in the individual's test.

Criterion 4: Earning Ratio (Based on Payment as the Bayesian Cost)—This criterion uses the Bayesian cost. A correct match was rewarded by \$25, an incorrect match cost \$25, and a missed match cost \$10. We have scored test-takers by the ratio of the payment they have received to the payment that they would have received had they been perfectly correct ("earning ratio").

Criterion 5: P-Rank—We developed a grading scheme, called a *performance rank* (or P-Rank) based on the different types of errors observed in our test. This scheme divides the test-takers into ten different sub-groups, based on the severity and number of errors observed. The grading scheme is described in Table 9.

TABLE 9—*The P-rank method for assigning grades for performance.*

Incorrrect Matches	Missed Detections	Grade
0	0	1
0	1	2
0	2	3
0	>2	4
1	0	5
1	1 or more	6
2	0	7
2	1 or more	8
3	0	9
All other combinations of errors		10

TABLE 10—*Hypothesis-test 1 using the Birnbaum-Hall distribution test: should we accept the hypothesis that the samples collected from the three professional groups come from the same population?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	0.0541	~1.0000*	Accept
hit rate	0.1471	0.9543	Accept
$P(H_1|D_1)$	0.1176	0.9990	Accept
payment	0.1113	0.9993	Accept
P-rank	0.0644	~1.0000†	Accept

*The actual value of p is 1–7.25E-17.
†The actual value of p is 1–6.11E-10.

Hypothesis Tested

Using these five criteria, three hypothesis-tests were conducted. We have tested data from: (i) the three professional groups, (ii) four groups—the group of nonprofessionals and the three groups of professionals, and (iii) the group of all professionals and the group of nonprofessionals.

Hypothesis-Test 1 (Three Professional Groups)—We tested the hypothesis that *there is no difference in the scores collected from the three professional groups* (H_0) against the hypothesis that *there is a difference in the scores collected from the three professional groups* (H_1).

Hypothesis-Test 2 (Four Groups, Three Professional, One Nonprofessional)—We tested the hypothesis that *there is no difference in the scores collected from the three professional groups and the group of nonprofessionals* (H_0), against the hypothesis that *there is a difference in the scores collected from the three professional groups and the group of nonprofessionals* (H_1).

Hypothesis-Test 3 (Two Groups, All Professionals and the Nonprofessionals)—We tested the hypothesis that *there is no difference in the scores collected from the group of all professionals and the group of nonprofessionals* (H_0), against the hypothesis that *there is a difference in the scores collected from the group of all professionals and the group of nonprofessionals* (H_1).

Results for Professionals

The results of the three hypothesis tests against the five scoring criteria using the Kolmogorov-Smirnov and location tests are given in Tables 10 through 15.

The distribution tests provide the following conclusions with respect to all five criteria: (i) The data generated by the three professional groups came from populations that are statistically

TABLE 11—*Hypothesis-test 1 using the Kruskal-Wallis location test: should we accept the hypothesis that the samples collected from the three professional groups come from the same population?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	0.0659	0.9676	Accept
Hit rate	0.2197	0.8960	Accept
$P(H_1|D_1)$	0.1254	0.9392	Accept
Payment	0.3998	0.8188	Accept
P-rank	0.7053	0.7028	Accept

TABLE 12—*Hypothesis-test 2 using the Birnbaum-Hall distribution test: should we accept the hypothesis that the samples collected from the three professional groups and the nonprofessional group come from the same population?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	0.5165	3.72E-04	Reject
Hit rate	0.5022	5.47E-04	Reject
$P(H_1|D_1)$	0.4964	9.29E-04	Reject
Payment	0.4534	3.71E-03	Reject
P-rank	0.5072	4.71E-04	Reject

TABLE 13—*Hypothesis-test 2 using the Kruskal-Wallis location test: should we accept the hypothesis that the samples collected from the three professional groups and the nonprofessional group come from the same population?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	37.3169	3.94E-08	Reject
Hit rate	0.3248	9.55E-01	Accept
$P(H_1|D_1)$	32.4363	4.23E-07	Reject
Payment	20.9509	1.08E-04	Reject
P-rank	22.9556	4.13E-05	Reject

TABLE 14—*Hypothesis-test 3 using the Kolmogorov-Smirnov distribution test: should we accept the hypothesis that the samples collected from the professionals come from the same population as the nonprofessionals?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	0.6098	7.23E-03	Reject
Hit rate	0.5366	2.57E-02	Reject
$P(H_1|D_1)$	0.5610	1.71E-02	Reject
Payment	0.5122	3.78E-02	Reject
P-rank	0.5610	1.71E-02	Reject

TABLE 15—*Hypothesis-test 3 using the Mann-Whitney location test: should we accept the hypothesis that the samples collected from the professionals come from the same population as the nonprofessionals?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	6.1057	5.14E-10	Reject
Hit rate	0.2893	3.86E-01	Accept
$P(H_1|D_1)$	5.6862	6.51E-09	Reject
Payment	4.5412	2.80E-06	Reject
P-rank	4.7352	1.10E-06	Reject

similar; (ii) the data generated by the three professional groups and the nonprofessional group came from populations that are statistically *different*; (iii) the data generated by the professional group and nonprofessional group came from populations that are statistically *different*.

The same conclusions are borne out by the location tests with the exception of the hit rate criterion. This is hardly surprising, as we already know that the average hit rate is approximately the

TABLE 16—*The Kolmogorov-Smirnov distribution test: should we accept the hypothesis that the samples collected from the trainees come from the same population as the professionals?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	0.488	7.51E-07	Reject
Hit rate	0.362	5.90E-04	Reject
$P(H_1|D_1)$	0.648	9.94E-12	Reject
Payment	0.486	6.06E-06	Reject
P-rank	0.469	2.33E-06	Reject

TABLE 17—*The Kolmogorov-Smirnov distribution test: should we accept the hypothesis that the samples collected from the trainees come from the same population as the nonprofessionals?*

H_0: these samples are from the same population using	Statistic	p	Decision
w.a.r	0.610	0.007	Reject
Hit rate	0.537	0.026	Reject
$P(H_1|D_1)$	0.561	0.017	Reject
Payment	0.512	0.038	Reject
P-rank	0.561	0.017	Reject

same in all tested populations. The Mann-Whitney and Kruskal-Wallis tests, which examine ranks by averages, therefore accept the hypothesis that hit rate data of the professionals and nonprofessionals have the same averages. The Kolmogorov-Smirnov type tests, however, are sensitive to differences in *distributions*, and recognize that although the averages are the same (Tables 13 and 15), the *distributions are different* (Tables 12 and 14). Indeed the distribution tests distinguish between the hit-rate data generated by the professionals and the hit-rate data generated by nonprofessionals even though "on average" they are the same.

Notes—When test-takers who did not fill out the voluntary questionnaire were dropped from the professional pool, there was no meaningful change in the group performance measures. When nonprofessional test-takers with no college degree were dropped from the nonprofessionals pool, there was no meaningful change in the group performance measures.

Results for Trainees

We compared the trainees against the group of all professional document examiners, using the five scoring methods discussed previously. Results are shown in Table 16. We also compared the trainees against the group of all nonprofessionals. These results are shown in Table 17. Because these values were obtained with a much smaller number of individuals (8 trainees), they are somewhat less statistically significant than the values reported in the previous section.

It is apparent that the score distributions of the trainees' differ from both the distributions of the professionals and the nonprofessionals. At this stage the trainees' data, as judged by values of p in Tables 16 and 17, are somewhat "closer" to that of the nonprofessionals than to that of the professionals. We indeed expect trainees to be "between" nonprofessionals and professionals. It would be interesting to test the same trainee group in a year or two and see if its performance "moved" toward that of the professional group.

Concluding Remarks

• Test results lay to rest the debate over whether or not professional document examiners possess writer-identification skills absent in the general population. They do.

• Professional document examiners were capable of excellent performance in our tests—especially in avoiding wrong matching of documents to each other—even without the benefit of laboratory equipment or consultation with colleagues (that are available in practice).

• Lay persons are deficient as document examiners because they tend to over-match documents by declaring that documents generated by different people have come from the same hand.

• The data collected may provide answers to several additional interesting questions relating to proficiency. Is there a common property that characterizes the few errors made by the professional document examiners? Is there correlation between length of professional experience (or other factors) and performance? Are there sub-groups of examiners whose training, professional affiliations, or court experience make them more proficient than other groups? How important were the similarity/dissimilarity of texts and the lengths of texts in affecting performance?

• During the preparation of the test, we created a large corpus of original documents that could be used as a database for training, testing, machine intelligence studies, and further development of standards in writer identification.

References

1. Risinger DM, Denbeaux MP, Saks MJ. Exorcism of ignorance as a proxy for rational knowledge: the lessons of handwriting identification 'expertise'. University of Pennsylvania Law Review 1989;137:731–787.
2. Kam M, Wetstein J, and Conn R. Proficiency of professional document examiners in writer identification. J Forensic Sci 1994;39:5–14.
3. Galbraith O, Galbraith CS, Galbraith NG. The principle of the "drunkard's search" as a proxy for scientific analysis: the misuse of handwriting test data in a law journal article. Intl J Forensic Doc Exam 1995;1(1):7–17.
4. Beck J. Sources of errors in forensic handwriting evaluation. J Forensic Sci 1995;40(1):78–82.
5. United States of America against Roberta Starzecpyzel and Eileen Starzecpyzel, United States District Court, Southern District of New York, Case No. S1:93 CR.49.
6. United States of America against Kathleen Kresner Jones, United States Federal District Court, Eastern District of Tennessee, Case No. 3:95 CR.24.
7. United States vs. Specialist Jeffery A. Ruth, US Army Court of Criminal Appeals, Case No. 9400093.
8. United States of America; Government of the Virgin Islands vs. Edwin Velasquez, United States Court of Appeals, Third Circuit, Case No. 93-7236.
9. United States of America vs. Ruben Renteria Sr. and Ruben Renteria Jr., United States District Court for the District of New Mexico, Case No. 95-320-JP.
10. United States of America vs. Joseph Pravato and Syed Arrhad Mahmood, United States District Court, Eastern District of New York, Case No. 95 CR 981.
11. United States of America vs. Muhammed Ijaz Chohan, United States District Court, Eastern District of New York, Case No. 95 CR 876.
12. The People of the State of California vs. Victor Kiet Diep and Thuan Wu, Superior Court of the State of California, San Francisco, Court No. 157436-01/02.
13. American Society for Testing and Materials, ASTM Standard E 1658: Terminology for Expressing Conclusions of Forensic Document Examiners.

14. Van Trees HL. Detection, estimation, and modulation theory, Part 1. New York: Wiley, 1968.

15. The Random House Dictionary of the English Language, New York: Random House, 1987.

16. Siegel S. Nonparametric statistics for the behavioral sciences. New York: McGraw Hill, 1956;127–136,184–94.

17. Sachs L. Applied statistics—A Handbook of Techniques. New York: Springer Verlag, 1984.

18. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C. New York: Cambridge University Press, 1992; 623–28.

19. Gibbons JD. Nonparametric methods for quantitative analysis. New York: Holt, Rinehart and Winston, 1976.

20. Conover WJ. Practical non-parametric statistics. New York: John Wiley and Sons, 1980.

21. Birnbaum ZW, Hall RA. Small sample distributions for multi-sample statistics of the Smirnov type. Ann Math Stat 1960;31:710–20.

Additional information and reprint requests:
Moshe Kam, Ph.D.
Data Fusion Laboratory
Electrical and Computer Engineering Department
Drexel University
32nd and Chestnut Sts.
Philadelphia, PA 19104