Moshe Kam,[1] *Ph.D.; Kishore Gummadidala,*[1] *M.S.; Gabriel Fielding,*[1] *Ph.D.; and Robert Conn,*[2] *Ph.D.*

Signature Authentication by Forensic Document Examiners

REFERENCE: Kam M, Gummadidala K, Fielding G, Conn R. Signature authentication by forensic document examiners. J Forensic Sci 2001;46(4):884–888.

ABSTRACT: We report on the first controlled study comparing the abilities of forensic document examiners (FDEs) and laypersons in the area of signature examination. Laypersons and professional FDEs were given the same signature-authentication/simulation-detection task. They compared six known signatures generated by the same person with six unknown signatures. No *a priori* knowledge of the distribution of genuine and nongenuine signatures in the unknown signature set was available to test-takers. Three different monetary incentive schemes were implemented to motivate the laypersons.

We provide two major findings: (i) the data provided by FDEs and by laypersons in our tests were significantly different (namely, the hypothesis that there is no difference between the assessments provided by FDEs and laypersons about genuineness and nongenuineness of signatures was rejected); and (ii) the error rates exhibited by the FDEs were much smaller than those of the laypersons. In addition, we found no statistically significant differences between the data sets obtained from laypersons who received different monetary incentives.

The most pronounced differences in error rates appeared when nongenuine signatures were declared authentic (Type I error) and when authentic signatures were declared nongenuine (Type II error). Type I error was made by FDEs in 0.49% of the cases, but laypersons made it in 6.47% of the cases. Type II error was made by FDEs in 7.05% of the cases, but laypersons made it in 26.1% of the cases.

KEYWORDS: forensic science, questioned document examination, signature, validation, handwriting

Forensic examination of signatures is performed for authentication of legitimate signatures and for detection of simulations, transfers, and other attempts to manipulate and misrepresent signatures. In adjudicating disputes over signatures, courts often use the testimony of professional forensic document examiners (FDEs). FDEs are called upon due to their reputed expertise in signature examination, deemed to have been developed through "knowledge, skill, experience, training, or education" (Rule 702, Federal Rules of Evidence).

The proficiency of FDEs has become a topic of vigorous debate in the last few years, receiving growing attention in the scientific literature, the courts, the popular press, and law reviews (1–3). Much of this debate stemmed from the scarcity, at least until 1994, of controlled studies that compared the performance of professional FDEs to the performance of laypersons in document-examination tasks. Several controlled studies have been conducted and reported since then (1–3), but they all used *freely and naturally prepared* handwritten texts. In this paper we report on the first controlled study involving *simulated signatures.* We compare the capabilities of FDEs and laypersons in authenticating genuine signatures and detecting simulated ones.

In May 1998 we conducted a comprehensive test of signature authentication, involving 69 FDEs and 50 laypersons (referred to collectively as the "test-takers"). Each test-taker was required to compare two sets of data: (i) the *known* set, comprising six original signatures written by the same person; and (ii) the *unknown* set, comprising six signatures of unknown origin. The number of nongenuine signatures in the *unknown* set could have been any integer from zero to six.

Our study had three objectives:

(i) test the hypothesis that there is no difference between the assessments of FDEs and laypersons about genuineness and nongenuineness of signatures;
(ii) calculate the *error rates* of FDEs and laypersons in the authentication of genuine signatures and the detection of simulated signatures; and
(iii) test the hypothesis that monetary incentives that we offered to the laypersons who took our tests changed the data generated by them.

Organization of the Paper

We provide a summary of the main results, including tables of error rates exhibited by the FDEs and the laypersons. We then provide a detailed description of the test, including data collection procedures and details on the monetary incentives offered to laypersons. The rest of the paper is devoted to statistical tests. We describe the criteria for data comparison, the hypotheses tested, and the results of the statistical tests.

Summary of the Main Results

Comparison of Data

Using standard statistical tests, the hypothesis that there is no difference between the assessments provided by FDEs and laypersons about genuineness and nongenuineness of signatures was *rejected.* We found significant statistical differences between the data generated by FDEs and by laypersons. The laypersons wrongly

[1] Data Fusion Laboratory, Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA.
[2] RABA Technologies, Inc., 10500 Little Patuxent Parkway, Suite 790, Columbia, MD.

Received 10 May 1999; and in revised form 5 August 2000, 25 September 2000; accepted 29 September 2000.

classified nongenuine signatures as "genuine" 13 times more often than FDEs. The laypersons wrongly classified genuine signatures as "nongenuine" four times more often than FDEs. There were no statistically significant differences between the data sets obtained from laypersons who received different monetary incentives in our tests.

Performance

The FDEs made far fewer mistakes than laypersons, as demonstrated in Table 1. The table shows the conditional probabilities Pr (declared signature to be α | signature was β) where *test-taker* $\in$ {FDE, layperson}, $\alpha \in$ {genuine, nongenuine, indeterminable with respect to genuineness}, and $\beta \in$ {genuine, nongenuine}.

A posteriori *Error Probabilities*

The FDEs had much lower *a posteriori* error probabilities compared to laypersons. Table 2 shows these probabilities. They are in the form Pr (Signature was nongenuine | *test-taker* declared signature to be genuine), Pr (Signature was genuine | *test-taker* declared signature to be nongenuine).

Methods

Description of the Test

Data Collection and Data Organization—We recruited 64 individuals for about 3 h of work, requiring the provision of handwritten samples and other tasks. The individuals were graduate and undergraduate students aged 19 to 30 enrolled at the time at Drexel University. No member of this group had past expertise with professional examination of documents, nor had any participated in Drexel University's research on forensic document examination. These recruits were compensated for their services ($25).

In the course of the 3-h period, each participant provided 12 freely and naturally executed examples of his/her normal signature. Each signature was executed on a single, white, unattached sheet of paper (several types of paper of different weights were used). All signatures were written with medium-tip blue or black Bic ballpoint pens, supplied by the authors. Proper care was exercised to avoid leaving any indentation or trace from one signature on a page containing another signature. All pages with the contributed signatures were assigned random numbers for identification purposes.

TABLE 1—*Error distribution in the signature authentication/simulation-detection test.*

	Decision					
	"QS = G"		"QS = ?"		"QS = NG"	
Truth	FDEs	Laypersons	FDEs	Laypersons	FDEs	Laypersons
QS = G	85.89%	70%	7.05%	4.3%	7.05%	26.1%
QS = NG	0.49%	6.47%	3.45%	1.4%	96.06%	92%

"QS = G": Questioned signature is genuine.
"QS = NG": Questioned signature is nongenuine.
"QS = ?": Test-taker could not determine whether the questioned signature was geuine or nongenuine.

TABLE 2—A posteriori *error probabilities.*

	P (NG | "G")	P (G | "NG")
FDEs	0.008	0.08
Laypersons	0.070	0.25

TABLE 3—*Distribution of genuine and nongenuine signatures in the "unknown" package (*i:j *means* i *signatures in the "unknown" package are genuine,* j *are nongenuine).*

6:0	5:1	4:2	3:3	2:4	1:5	0:6
2	7	17	21	12	4	1

The signatures provided by the recruited individuals were all genuine, naturally and freely prepared, and involved no self-tracing or self-copying. All 12 signatures provided by each participant were compared with his/her signature on a check-in form that had been signed before the session began. To the best of our knowledge, all signatures were created in the manner normally used by the signing individuals, using their real names and the normal procedure by which these individuals usually write their signatures.

Each 12-signature set was then divided into two six-signature subsets. Each signed page received a random code number. We used a random assignment of each one of the signed and coded sheets to one of the two subsets. One subset was labeled "known" and other "unknown." Each "unknown" set was then assigned a random number from 0 to 6, indicating how many genuine signatures were to be removed from the set and replaced by simulations. Table 3 shows the distribution of genuine/nongenuine ratios in the resulting packages.

Seven individuals, distinct from the signature providers, were recruited to execute the simulations, using the manual techniques described in a text on forensic document examination (4). To the best of our knowledge, these seven individuals had no prior experience in signature simulation. They used tracing paper, carbon paper, flashlights, and overhead projectors. No computer-generated manipulations were involved, nor were computers used in any other way to create the simulated signatures.

Each simulation was created by a single individual who was provided with the six genuine signatures of the appropriate "known" set and was allowed unlimited time to practice and experiment in the "creation" of a simulation. If two simulations were required for an unknown set, two individuals supplied one simulation each for that set. If three to six simulations were required, up to three individuals supplied simulations, with no more than two simulations provided by any one individual. No auto-forgeries of any kind were requested or executed (as no individual who created original signatures was included in the group of simulators). Nongenuine signatures were created on white sheets of paper of the same types used during the signature-collection sessions. The pages with the simulated signatures were also assigned random numbers for identification purposes.

Random number generation and code management were exercised according to the common procedures for securing codes for one-time-use. Individuals who did not possess our secured master identification list would not be able to separate genuine signatures from nongenuine signatures to any statistically significant degree on the basis of, or through the aid of, the random identification numbers.

Test-Takers—The test was administered four times. On May 9, 1998, in San Diego, California, the test was taken by 44 FDEs attending a meeting of the Southwestern Association of Forensic

Document Examiners. On May 14, 1998, in Rockville, Maryland, the test was taken by 12 FDEs attending the Mid-Atlantic Association of Forensic Scientists. On May 18, 1998, in New York City, the test was taken by 13 FDEs attending a meeting of examiners from the area. On November 17, 1998, in Philadelphia, Pennsylvania, the test was taken by 50 laypersons. The laypersons were students (graduate and undergraduate), staff members, and faculty of Drexel University. The laypersons group was selected to resemble the educational profile of the forensic document examiners. (A detailed questionnaire on education and background was distributed to all FDE test-takers, and almost all completed it in full voluntarily).

The professional FDEs who took our test all met at least one of the following requirements:

(i) certification by the American Board of Forensic Document Examiners (ABFDE);
(ii) membership in the American Society of Questioned Document Examiners (ASQDE);
(iii) membership in the Southwestern Association of Forensic Document Examiners (SWAFDE); or
(iv) membership in the Questioned Document section of the Mid-Atlantic Association of Forensic Scientists (MAAFS).

Test Administration

All test-takers were provided with the following information:

(i) the *known* set contains genuine signatures provided voluntarily by a single individual in the course of a single sitting.
(ii) the *unknown* set contains an unknown number (0 to 6) of genuine signatures written by the person who provided the signatures in the *known* subset. The other signatures in the subset (6 to 0) were written by a simulator or by several different simulators.

All test-takers were allowed to use hand-held magnifiers and a light source, as well as microscopes of the kind used in regular forensic document examination practice. Magnifying glasses, light sources, and microscopes of equal quality were supplied to the laypersons in the Philadelphia test.

Test-takers were requested to state that a signature in the unknown set was written by the person who provided the known signatures (genuine) if they could declare "identification" or "strong probability" per ASTM Standard E1658.

Test-takers were requested to state that a signature in the unknown set was not written by the person who provided the known signatures (nongenuine), if they could declare "elimination" or "strong probability did not write" according to the same standard. These terms were explained at length to the laypersons.

Monetary Incentives

We use the following notation:

A "correct decision" means that a genuine document was declared genuine or nongenuine signature was declared nongenuine.
A "serious error" means that a genuine signature was declared nongenuine or when a nongenuine signature was declared genuine.
"Indecision" means that the document was not declared either genuine or nongenuine.

TABLE 4—*Incentive table.*

Correct Decision	Serious Error	Indecision
\$8	−\$8	\$0
\$8	−\$8	\$4
\$8	−\$8	−\$4

Three types of monetary incentives were offered to the laypersons.

(i) Line 1 in Table 4 shows the first incentive. Correct decisions and serious errors were rewarde-d and penalized, respectively, by \$8 per decision. Conservatism, manifested by indecision, was neither rewarded nor penalized. If the total reward was less than \$24, the test-taker received \$24.
(ii) Line 2 in Table 4 shows the second incentive. Correct decisions and serious errors were rewarded and penalized, respectively, by \$8 per decision. Conservative decisions were rewarded by \$4 per decision. If the total reward was less than \$24, the test-taker received \$24.
(iii) Line 3 in Table 4 shows the third incentive. Correct decisions and serious errors were rewarded and penalized, respectively, by \$8 per decision. Conservative decisions were penalized by \$4 per decision. If the total reward was less than \$24, the test-taker received \$24.

Each layperson was aware of his/her monetary incentive before beginning the test. As we explained previously (3), these monetary incentives are relatively high when compared to the common practice in experimental psychology.

Criteria for Data Comparison

We used three criteria to compare the data provided by the test-takers.

Criterion I: Error Rates

There are four possible errors in our test:

Ia. False authentication: (false ID)	a test-taker is given a nongenuine signature but declares that it is genuine;
Ib. Failure to detect simulation: (missed Elim)	a test-taker is given a nongenuine signature but cannot come to any one of the definitive conclusions (identification/strong probability/ strong probability did not write/elimination);
IIb. Failure to authenticate: (missed ID)	a test-taker is given a genuine signature but cannot come to any one the definitive conclusions (identification/ strong probability/strong probability did not write/elimination);
IIa. False simulation-detection: (false Elim)	a test-taker is given a genuine signature but declares that it is a simulation.

Of the four errors, Ia and IIa are more serious errors than Ib and IIb. The selection of Ib or IIb may reflect conservatism on the part of the test-taker. The four types of errors are *linked* (e.g., one can avoid Type I errors and increase Type II errors by refraining from making any declarations of genuineness).

TABLE 5—*P-rank.*

P-rank	# of Correct Decisions (G | G), (NG | NG)	# of Conservative Errors (? | G), (? | NG)	# of Serious Errors (NG | G), (G | NG)
1	6	0	0
2	5	1	0
3	4	2	0
4	3 or fewer	3 or larger	0
5	more than 3	any	1
6	less than 3	any	1
7	any	any	2
8	any	any	3
9	any	any	more than 3

G = Genuine, NG = Nongenuine, ? = No decision

Criterion II: P-rank

As we have done previously (2,3), we divided our test-takers into nine groups based on performance. The assignment of an individual to a category depends on the difficulty of the test taken by that individual and on his/her capabilities. Therefore the P-rank is useful for comparison of the two data distributions (of the FDEs and the laypersons) while using the same tests, but not for direct proficiency assessment of the individuals who took the test or the groups of individuals. Table 5 shows the assignment.

Statistical Tests

The literature (5–8) 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., FDEs versus laypersons) and data from k ($k \geq 3$) groups (e.g., data from the three groups of laypersons). For each criterion we use a test on *distributions* (of the Kolmogorov-Smirnov type), and a test on *locations* (of the Mann-Whitney type).

Four statistical tests were used: the first two are distribution tests; the other two are location tests. We described the choice and use of these tests in (2).

The Kolmogorov-Smirnov (KS) two-sample test (2) was used to decide whether or not two independent samples have been drawn from the same population (or from populations with the same distribution).

The Birnbaum-Hall (BH) k-sample test (2) was used to decide whether k independent samples have been drawn from populations with the same distribution.

The rank test of Mann and Whitney (MW) (2) was used to test whether populations of two independent samples differ with respect to their means.

The Kruskal-Wallis (KW) one-way analysis of variance by ranks (2) was used to decide whether $k \geq 3$ independent samples are from different populations with respect to means.

Hypotheses Tested

Using the four error rates and the P-rank as the scoring criteria, three hypotheses-tests were conducted. We tested data from: (i) the group of FDEs and the group of all laypersons and (ii) the three sub-groups of laypersons who received different monetary incentives.

Hypothesis-Test 1 (Group of FDEs and Group of Laypersons)—We tested the hypothesis that *there is no significant difference in the scores collected from the group of FDEs and the group of laypersons (H_0)* against the hypothesis that *there is a significant difference in the scores collected from the two groups (H_1).*

Hypothesis-Test 2 (Three Sub-Groups of Laypersons)—We tested the hypothesis that *there is no significant difference in the scores collected from the three sub-groups of laypersons that had different monetary incentives (H_0),* against the hypothesis that *there is a significant difference in the scores collected from the three sub-groups of laypersons that had different monetary incentives (H_1).*

Results/Discussion

Results of Statistical Tests

Results for Professionals—The results of the two hypothesis tests against the three scoring criteria using the Kolmogorov-Smirnov and location tests are given in Tables 6 and 7. The distribution tests provide the following conclusion with respect to all three criteria: the data generated by the FDE group and the layperson group came from populations that are statistically *different.*

Monetary Incentives for Laypersons—We compared the three groups of laypersons who had different monetary incentives. We used the three scoring methods discussed above. Results are shown in Table 8. The tests *do not reject* the hypothesis

TABLE 6—*Hypothesis-test 1 using the Kolmogorov-Smirnov distribution test: should we accept the hypothesis that the samples collected from the FDEs come from the same population as the laypersons?*

H_0: These Samples are from the Same Population Using . . .	Statistic	*p*	Decision
FalseID	0.285	1.20E-2	Reject
FalseElim	0.271	1.97E-2	Reject
P-rank	0.318	3.52E-3	Reject

TABLE 7—*Hypothesis-test 1 using the Mann-Whitney location test: should we accept the hypothesis that the samples collected from the FDEs come from the same population as the laypersons?*

H_0: These Samples are from the Same Population Using . . .	Statistic	*p*	Decision
FalseID	2.50	1.26E-2	Reject
FalseElim	3.36	7.70E-4	Reject
P-rank	3.54	4.01E-4	Reject

TABLE 8—*Hypothesis-test 2 using the Kruskal-Wallis location test: should we accept the hypothesis that the samples collected from the three groups of laypersons who had different monetary incentives came from the same population?*

H_0: These Samples are from the Same Population Using . . .	Statistic	*p*	Decision
FalseID	2.39	0.303	Do not reject
FalseElim	3.87	0.144	Do not reject
P-rank	1.28	0.528	Do not reject

that different incentives do not affect the performance of the laypersons.

Conclusions

We found that in a signature authentication task, data generated by FDEs are statistically different from data generated by laypersons. In addition, error rates of the FDEs were much lower than those of the laypersons. These results point to the superiority of FDEs over laypersons in determination of genuineness of signatures and in detection of simulations. The continued failure of monetary incentives to induce changes in the laypersons data (3) may indicate that money alone—without training and practice—cannot induce laypersons into good performance as forensic document examiners.

Acknowledgments

This study was supported in part by the U.S. Department of the Army (Aberdeen Proving Ground, MD.) under award document DAAD05-98-C-0010 entitled "Proficiency tests for questioned document examiners."

References

1. Kam M, Wetstein J, and Conn R. Proficiency of professional document examiners in writer identification. J Forensic Sci 1994;39:5–14.
2. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. J Forensic Sci 1997;42(5):778–86.
3. Kam M, Fielding G, Conn R. Effects of monetary incentives on performance of nonprofessionals in document-examination proficiency tests. J Forensic Sci 1998;43(5):1000–5.
4. Harrison W. Forgery detection: a practical guide. New York: F. A. Praeger, 1964.
5. Siegel S. Nonparametric statistics for the behavioral sciences. New York: McGraw Hill, 1956.
6. Sachs L. Applied statistics—a handbook of techniques. New York: Springer Verlag, 1984.
7. Conover WJ. Practical non-parametric statistics. New York: John Wiley & Sons, 1980.
8. 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.
Electrical and Computer Engineering Department
Drexel University
3141 Chestnut St.
Philadelphia, PA 19104
E-mail: kam@minerva.ece.drexel.edu