

Jodi Sita,¹ *B.HSc (Hons)*; Bryan Found,^{1,2} *Ph.D.*; and Douglas K. Rogers,¹ *Ph.D.*

Forensic Handwriting Examiners' Expertise for Signature Comparison*

ABSTRACT: This paper reports on the performance of forensic document examiners (FDEs) in a signature comparison task that was designed to address the issue of expertise. The opinions of FDEs regarding 150 genuine and simulated questioned signatures were compared with a control group of non-examiners' opinions.

On the question of expertise, results showed that FDEs were statistically better than the control group at accurately determining the genuineness or non-genuineness of questioned signatures. The FDE group made errors (by calling a genuine signature simulated or by calling a simulated signature genuine) in 3.4% of their opinions while 19.3% of the control group's opinions were erroneous. The FDE group gave significantly more inconclusive opinions than the control group.

Analysis of FDEs' responses showed that more correct opinions were expressed regarding simulated signatures and more inconclusive opinions were made on genuine signatures. Further, when the complexity of a signature was taken into account, FDEs made more correct opinions on high complexity signatures than on signatures of lower complexity. There was a wide range of skill amongst FDEs and no significant relationship was found between the number of years FDEs had been practicing and their correct, inconclusive and error rates.

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

The debate substantially initiated by Risinger, Denboux, and Saks (1) concerning forensic document examiner (FDE) expertise in handwriting identification continues in the literature (2–10) and the courts (reviewed in 11) and has emerged as an issue in a recently published document examination textbook (12). This paper attempts to contribute further evidence pertinent to the debate.

Generally expertise in a particular area may be thought of as ability that is superior to that of the average person. To our knowledge, there are only a limited number of published studies (2,4,6,9,10) that compare FDEs' ability in handwriting identification with that of the average person. Empirical studies on text-based handwriting identification have shown evidence of expertise in trained FDEs (2,6). These studies by Kam and colleagues compared FDEs' opinions to those of a control group of non-examiners in text-matching exercises. They found that FDEs made significantly fewer errors than the control groups and were significantly more cautious in making decisions, however, the control groups were able to correctly identify similar amounts of writing.

The same pattern of results was found in a signature comparison task (9). In this study the opinions of a small group of seven FDEs was compared with that of a control group of non-examiners on each of 150 questioned (unknown) signatures that were a mixture of genuine and simulated signatures. Again FDEs made significantly fewer errors than the control group, had a significantly higher inconclusive rate and there was no difference between groups in the raw number of correct opinions. In a more recent sig-

nature comparison study (10), 69 FDEs and 50 control subjects provided opinions on a set of six unknown signatures that were a variable mixture of genuine and simulated signatures. The study found a significant difference between the group of FDEs and the control group with FDEs making substantially fewer errors.

The evidence from controlled studies to date supports the existence of FDE expertise. However, further studies involving different FDEs and different handwriting material are required to increase the generality of the findings. In addition, evidence is required in relation to the spread of skill amongst FDEs and how FDE performance relates to the complexity of the questioned writings which was not described in the two previous signature comparison studies.

Here, we report on the findings of a test involving multiple signature comparisons that was administered to both Australian and New Zealand government FDEs and a lay group of non-examiners. The primary aim of the study was to provide evidence in relation to the existence or non-existence of expertise for this group of government practitioners in signature examination. Further aims were to determine for the FDE group:

- Error rates for this test of signature examination
- The existence and magnitude of any variation in identifying genuine signatures compared with simulated signatures, and for signatures of different complexity
- The variation in group performance for different signature cases
- The variation in performance between individuals
- Whether a relationship between professional experience and performance could be demonstrated

Methods

In this study, practicing FDEs and control subjects (non-FDEs) were asked to examine 150 questioned signatures and provide

¹ Forensic Expertise Profiling Laboratory, School of Human Biosciences, La Trobe University, Victoria, 3086, Australia.

² Document Examination Team, Victoria Forensic Science Centre, Forensic Drive, Macleod, Victoria, 3085, Australia.

*Presented at The Australian & New Zealand Forensic Science Society, 15th International Symposium on the Forensic Sciences, Gold Coast, Queensland, Australia, 5–10 March 2000.

Received 14 Jan. 2002; accepted 20 Feb. 2002; published 31 July 2002.

opinions as to whether each signature was *genuine*, *simulated* or (in cases where they considered that it was not possible to differentiate between the two categories) provide an *inconclusive* opinion. Each questioned signature and each subject's opinions were recorded and analyzed. Subjects' results were recorded as numerical scores for the number of opinions correct, wrong, and inconclusive. A between group analysis was used to address the question of FDE expertise and further analyses of subjects' performance were undertaken as described below in the "Procedures" and "Results" sections.

Subjects

Seventeen FDEs from five Australian and New Zealand government forensic laboratories participated in the study. FDEs were aged between 20–50 years, 14 were fully qualified and had been practicing handwriting examination between 3.5 and 20 years and three were trainees with 1–3 years pre-qualified training. The mean training period prior to qualification for the group was 3.6 years with a range of three to five years. Thirteen individuals with no document examination experience or any prior professional association with handwriting examination were used as the control group. The control subjects were aged between 20–49 years and were all tertiary qualified staff or postgraduate students drawn from La Trobe University, Melbourne, Australia.

Materials

In total the test consisted of ten mock cases (one for each signature provider, *primary author*, see below) made with 15 "Exemplar" signatures and 15 "Questioned" signatures. Because of the convention in Australia and New Zealand at the time of the study, these Exemplar signatures were referred to as "Standards".

Signature Collection—Ten staff of the Victoria Forensic Science Centre, who were not FDEs, volunteered to provide their signatures and consented to their signatures being simulated for this experiment. Each of the ten volunteers executed 30 signatures on blank sheets of A4 paper using a variety of different pens. The signatures collected were freehand, natural signatures sampled within a 12-month period. A random selection of these signatures was used to make up the genuine "Exemplar" packages and the genuine signatures of the "Questioned" packages. For the purpose of this study, the providers of these signatures will be referred to as *primary authors*.

Different volunteers were recruited to attempt to simulate the *primary authors'* signatures. For the purpose of this study, the providers of these signatures will be referred to as the *simulating authors*. Twenty-five volunteers, both academic and general staff from La Trobe University partook in attempting to simulate each of the *primary authors'* signatures. Simulations were made freehand on blank sheets of A4 paper using three randomly selected genuine signatures from each of the ten primary authors as the models. Simulators were given an unlimited amount of time to practice, and each submitted two simulations of each of the ten primary authors' signatures. The two simulations consisted of a "one-off" signature that was executed on a specifically marked sheet of paper, and a "best-try" signature which was the signature that the simulators perceived to be their best forgery. The simulations chosen for inclusion into the validation test exhibited what the experimenters considered to be a wide range of skill. This pool of simulations, whose authors were known to the researchers, was used to complete the "Questioned" packages.

Test Packages—Each A4 page from each *primary author* or *simulating author* contained one signature, which was cut into a 6 × 21 cm section. Four of these were attached to the left side of an A4 sized card, with a piece of clear plastic covering the page. This permitted the test subjects to inspect both sides of the document on which the signature appeared without damaging or removing the document. Both FDEs and non-FDEs were provided with the originals of the test packages.

"Exemplar" Packages containing only genuine signatures were assembled for the ten *primary authors*. For each package 15 genuine signatures were randomly chosen from the pool of genuine signatures. The signatures used as models by the simulating authors were not included in the packages. The ten packages were numbered (1–10) with the printed initials of the *primary author* and case number heading each page.

"Questioned" Packages containing a combination of genuine and simulated signatures were assembled for each of the ten cases. The ratio of simulated to genuine signatures included in each package was randomly determined. In addition, a random number between 1–1000 was allocated to each questioned signature. This number was used by the experimenters to identify the actual author of the signature.

Procedures

The assembled test packages were circulated in turn amongst the participants. In addition to the test packages, each participant received a tick-box answer booklet, instructions and a description of relevant terms. Subjects were requested to complete the test individually. They were requested to compare each "Questioned" signature to the relevant set of "Exemplar" signatures as though it were part of a normal forensic case and to provide an opinion as to whether each questioned signature was genuine, simulated or that they were unable to say (inconclusive opinion). For each signature, which was coded randomly, subjects were required to tick a box indicating whether, in their opinion: a) the signature was genuine, b) the signature was simulated, or c) the examination was inconclusive. They recorded their opinion by ticking the appropriate column of the answer booklet. Subjects were provided with the following definition of terms:

Genuine—The questioned signature is in your opinion written by the same person who wrote the standard signature group.

Simulated—The questioned signature is inconsistent with the standard signature group and displays features that you consider indicative of a copying process. Note that this term does not imply that the standard signature group writer did not write it.

Inconclusive—You are not prepared to express an opinion as to whether the questioned signature is genuine or simulated.

Additional information was given to the control group to ensure that every effort was made to allow them to appreciate any implications their opinions would have in a real case scenario. Control subjects were informed that if these were real cases, and they wrongly identified a genuine signature as a simulation, that could result in criminal charges being laid upon an innocent person; and if they wrongly identified a simulated signature when in fact it was a genuine, it could result in a guilty person being found not guilty, or could implicate another innocent person in a criminal act. They were also informed that reaching an inconclusive opinion, with respect to the guilt or innocence of a particular person, would not necessarily have any criminal implications.

Each FDE was requested to complete an information sheet stating the length of time they had been qualified, the length of time they had been examining handwriting and the length of time spent in pre-qualification training.

Ethics Approval

Approval for this study was obtained from the La Trobe University Human Ethics Committee on the basis that the *primary authors* providing their signatures for simulation did so with full consent and that images of their signature would not be presented in published material. In addition, it was agreed that information regarding the results of individual subjects would remain strictly confidential. Informed consent was obtained from all participants.

Data Analysis

Each subject's opinions were recorded numerically. Opinions given as genuine were coded as "1", simulated opinions as "2" and inconclusive opinions as "3". This data was recorded, stored and analyzed using the Microsoft Excel database spreadsheet program. Numbers of correct, wrong, and inconclusive opinions for each subject and each questioned signature were calculated.

Statistical analyses were performed using the Statistical Package for the Social Sciences (Version 9). Planned comparisons of the numbers of correct, wrong, and inconclusive opinions were performed using unpaired, 2-tailed student *t*-tests between:

- FDE and control groups
- Genuine and simulated signatures
- High complexity and medium complexity signatures
- High complexity and medium complexity genuine signatures
- High complexity and medium complexity simulated signatures

The complexity score for each signature set was calculated using the method previously described (13). In this study, we calculated each case's complexity score based on the number of turning points and line intersections in the set of primary authors' signatures.

To evaluate the correlation between FDEs' performance and years of experience a number of Pearson's correlation coefficients were calculated.

Results of Comparison of FDEs and Controls

Subjects provided opinions on whether each of the 150 questioned signatures were simulated or genuine. Each opinion was considered to be an opinion unit. For the FDE group the total number of opinion units was 2550 (17 × 150) and the control group was 1950 (13 × 150). Analysis of opinions as correct, wrong, and inconclusive showed that the FDE group had 1397 correct opinion units (54.8% of opinions), 1067 inconclusive opinion units (41.8%) and 86 wrong opinion units (3.4%). The control group had 1113 correct opinion units (57.1%), 461 inconclusive opinion units (23.6%) and 376 wrong opinion units (19.3%). Thus the error rate for the FDE group was 3.4% and for the control group was 19.3%.

Table 1 shows the mean number of opinions per subject for each group. It can be seen that the number of correct opinions was similar for both groups. The mean number of correct opinions for the FDE group (mean = 81.9) and the control group (mean = 85.6) were not different ($t = 0.58, p = 0.568$). The difference found between the two groups lies in the number of wrong and inconclusive opinions expressed. Significantly more errors ($t = 7.33, p < 0.001$)

TABLE 1—Mean number of correct, wrong, and inconclusive opinions per subject for each group.

Group	Correct Opinions	Wrong Opinions	Inconclusive Opinions
FDEs	81.9	5.0	63.1
Lay persons	85.6	29.0	35.4

TABLE 2—Mean number of FDE's correct, wrong, and inconclusive opinions per questioned signature for genuine and simulated signatures separately.

Signature Category	Correct Opinions	Wrong Opinions	Inconclusive Opinions
Genuine	6.48	0.90	9.62
Simulated	12.0	0.26	4.74

were made by the control group (mean = 29) than were made by the FDE group (mean = 5), however the FDE group expressed significantly higher ($t = 3.48, p = 0.002$) numbers of inconclusive opinions (mean = 63.1) than the control group (mean = 35.4).

The "Called" Opinion

An opinion that was given for any questioned signature as either *simulated* or *genuine* was referred to as a *called* opinion; referring to the fact that the subject decided to *call* it rather than giving an inconclusive opinion. The number of "called opinions" is derived from the total number of opinions given minus the inconclusive opinions. The FDE group called 94.2% of opinions correctly with an error rate of 5.8%. The control group called 74.7% of opinions correctly with an error rate of 25.3%.

Further Results for FDEs

Results of further analyses of FDEs' performance are considered below. For the sake of brevity and clarity, we have only included results of the control group when they appear to provide substantially extra evidence in characterising FDE expertise relative to lay subjects.

Opinions Given for Genuine Questioned Signatures Versus Simulated Questioned Signatures

The questioned signature set consisted of 73 genuine signatures and 77 simulated signatures randomly allocated over the ten cases. FDEs' responses in relation to whether the questioned signatures were genuine or simulated have been analyzed separately in this section. The number of correct, wrong, and inconclusive opinions, for genuine and simulated questioned signatures were determined for each subject and for each signature.

Table 2 shows a summary for the FDE group expressed as the mean number of opinions per signature. FDEs performed significantly better for simulated signatures than for genuine signatures, giving more correct opinions ($t = 6.72, p < 0.001$), with fewer errors ($t = 3.93, p < 0.001$) and lower inconclusive rates ($t = 6.57, p < 0.001$).

The raw error rates for the group was 5.31% for genuine signatures and 1.52% for simulations. When considering only the *called* opinions, the group error rate for the genuine signatures was 12.2% (539 called opinions of which 66 were wrong) where the error rate

for simulations was 2.1% (944 called opinions of which 20 were wrong).

Relationship Between Signature Complexity and Opinion Rates

The results in relation to each mock case's signature complexity level were investigated. A complexity score (13) was calculated for the "Standard" signatures for each case. Cases were grouped according to their complexity score into a high complexity group or a medium complexity group. There were no cases with low complexity.

The high complexity group comprised of six cases with 90 questioned signatures (46 genuine, 44 simulated) and the medium complexity group contained four cases with 60 questioned signatures (27 genuine, 33 simulated). For each complexity group, the number of correct, wrong, and inconclusive opinions were calculated and compared using independent sample, two tailed *t*-tests.

Table 3 shows that FDEs provided more correct opinions per signature ($t = 4.43, p < 0.001$) for high complexity group signatures than medium complexity group signatures. The mean number of inconclusive opinions per signature for high complexity group signatures was significantly less ($t = 4.83, p < 0.001$) than for medium complexity group signatures, however, there was no difference in the number of errors made between the complexity groups ($t = 0.728, p = 0.468$).

Analysis of the responses for the control group found that they also gave more correct opinions ($t = 2.03, p = 0.044$) for high complexity group signatures (mean = 7.9) than medium complexity group signatures (mean = 6.6) although the magnitude of the difference was substantially less than for the FDEs. Unlike the FDEs, the control group showed no difference ($p = 0.292$) between conditions in relation to the number of inconclusive opinions made (high complexity; mean = 2.9, medium complexity; mean = 3.3). As with the FDE group there was no difference ($p = 0.110$) between conditions for the numbers of wrong opinions made (high complexity; mean = 2.2, medium complexity; mean = 3.1).

Signature Complexity and FDEs' Opinions on Genuine and Simulated Signatures

Further analysis of the complexity group's effect on opinions was investigated by dividing the high complexity group's signa-

tures into simulated and genuine signatures and the medium complexity group's signatures into simulated and genuine signatures. A comparison of the correct, wrong, and inconclusive opinions between genuine signatures of both complexity groups and between simulated signatures of both complexity groups was completed. Table 4 shows the mean number of opinions per signature.

The complexity score of genuine signatures affected the number of correct and inconclusive opinions made by FDEs between high and medium complexity signatures. There was no difference found in the number of wrong opinions between complexity groups ($t = .859, p = 0.393$). FDEs achieved significantly more correct opinions ($t = 4.39, p < 0.001$) for high complexity genuine signatures than for medium complexity genuine signatures and expressed more inconclusive opinions ($t = 4.8, p < 0.001$) on medium complexity group signatures than on high complexity group signatures. A similar pattern was found relative to the complexity score of simulated signatures. FDEs made more correct opinions ($t = 3.68, p < 0.001$) for high complexity simulated signatures than simulated signatures of medium complexity. Again, FDEs' inconclusive opinions were greater ($t = 3.91, p < 0.001$) for simulated signatures of medium complexity compared with numbers of inconclusive opinions on high complexity group signatures. There was no statistical difference between the number of wrong opinions made on simulated signatures of high or medium complexity ($t = .532, p = 0.596$).

Summary of Called Error Rates

The called error rates for the data sets discussed above are summarized in Table 5.

FDE Group Scores for the Different Cases

An analysis of the total number of correct, wrong, and inconclusive opinions made by the FDE group for each case was completed. The responses for each case were quite different; with some cases exhibiting high numbers of correct opinions with small numbers of

TABLE 3—Mean number of FDEs' correct, wrong and inconclusive opinions per signature categorized by complexity rating.

Signature Category	Correct Opinions	Wrong Opinions	Inconclusive Opinions
High Complexity	10.9	0.5	5.6
Medium Complexity	6.9	0.7	9.4

TABLE 5—The called error rate for FDEs' opinions on the different categories of questioned signatures.

Signature Category	Called Error Rate (%)
All simulated signatures	2.1
All genuine signatures	12.2
All high complexity signatures	2.0
All medium complexity signatures	16.5
High complexity, simulated signatures	1.6
High complexity, genuine signatures	8.6
Medium complexity, simulated signatures	2.9
Medium complexity, genuine signatures	25.6

TABLE 4—Mean number of FDEs' opinions per signature categorized by complexity rating for genuine and simulated signatures separately.

Signature Type	Correct		Wrong		Inconclusive	
	High Complexity	Medium Complexity	High Complexity	Medium Complexity	High Complexity	Medium Complexity
Genuine	8.5	3.1	0.8	1.1	7.7	12.8
Simulated	13.5	10.0	0.2	0.3	3.3	6.7

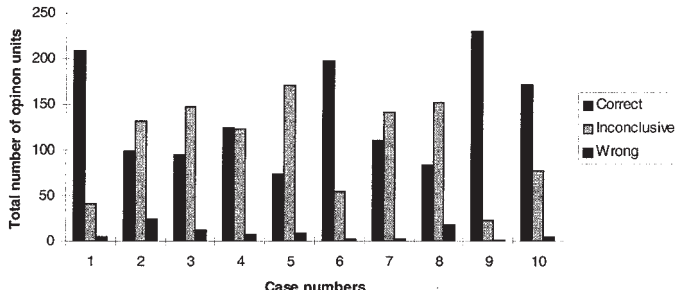


FIG. 1—The total number of correct, inconclusive and wrong opinions for the FDE group for each case.

errors, while others had high numbers of inconclusive opinions and larger numbers of errors.

Figure 1 shows that only four cases have high numbers of correct opinions and relatively low numbers of wrong opinions (Cases: 1, 6, 9, 10). Three cases have high numbers of errors (Cases: 2, 3 and 8; having 24, 12, and 18 errors respectively) and six out of the ten cases had moderate to high numbers of inconclusive opinions (Cases 2–5, 7, 8). The called error rates ranged from approximately 20% (Case 2) to 0.4% (Case 9).

Cases 1, 2, 5, 6, 9 and 10 are high complexity group cases. Of these, Cases 1, 6, 9 and 10 show high numbers of correct opinions with relatively low rates of inconclusive and erroneous opinions. The other two high complexity cases, numbers 2 and 5, do not show this trend; however, these were cases where the primary author's signature was either quite varied (Case 2) or had relatively poor line quality (Case 5). Cases 3, 4, 7, and 8 were cases consisting of medium complexity signatures. All of these cases show higher amounts of inconclusive opinions and errors and much lower rates of correct opinions.

The group results (reported in an earlier section) showed that there were higher correct rates and lower error rates for simulated signatures than for genuine signatures. Hence, the difference between cases could potentially just reflect the different proportions of simulated and genuine questioned signatures per case. To eliminate the chance that the proportion of different types of questioned signatures in each case was responsible for the results, we correlated the number of simulated signatures in each case to the total number of correct opinion units for each case and found no relationship ($r = 0.3, p = 0.34$).

Error Scores for Individual FDEs

Investigation of individual FDE's errors revealed that the amount of error observed for the group of FDEs was influenced by a few of the individual examiner's results. Figure 2 shows the variation in the group of FDEs. Twelve of the 17 FDEs made errors; five did not make any errors. Of the FDEs who made errors, three made only one error, five made between two–five errors and four examiners made between 13–27 errors each. Five FDEs mistakenly called a simulated signature genuine. Of these two FDEs made one error, one examiner made two errors of this type, one made six errors and one FDE mistakenly called 11 simulated signatures genuine.

FDE Experience

Calculation of Pearson's correlation coefficient was used to determine whether the number of years of experience as a qualified

FDE was related to the number of correct, wrong, and inconclusive opinions made for each FDE. No relationship was found for experience and number of correct opinions made ($r = 0.1, p = 0.674$), experience and number of wrong opinions made ($r = 0.07, p = 0.799$), or for experience and number of inconclusive opinions made ($r = 0.1, p = 0.648$).

Discussion

Expertise

This study is in agreement with previous studies on signature comparison (9,10) and supports the existence of expertise for this group of FDEs whose performance was significantly superior to that of the control group of average, well-educated people. Although the control group correctly identified similar numbers of questioned signatures to the FDE group this was at the expense of making significantly and substantially (six times) more errors than the FDE group. The difference between groups is also apparent when the called opinions are considered. Called opinion rates represent correct or error rates when subjects were prepared to express an opinion other than inconclusive and are arguably the rates with most significance for legal determinations. FDEs obtained a correct of called rate of 94.2% compared with only 74.4% for control subjects with the error rate for the control group (25.3%) being substantially more than the FDEs' called error rate of 5.8%.

The finding of this study, where both the FDE and non-FDE groups identified similar amounts correctly, but where the experts achieved significantly fewer errors is consistent with the findings of the earlier text matching studies (2,4,6). In each study, the control subjects were similar to the FDEs in the number of correct opinions made, but consistently made more errors.

In the present study, it is the error and inconclusive rate that sets the FDE and control groups apart statistically. The inconclusive rate for FDEs was significantly different from that associated with the lay group. FDEs clearly were far more conservative in calling these signatures than were the lay people in this study, in spite of the warnings given to lay people regarding the implications of expressing the wrong opinion. This provides some evidence that a component of FDE expertise is characterized by what they don't say rather than what they do say which may reflect their greater knowledge of the limitations of the comparison process and what conclusions can be drawn from observations made. Related to this was the observation that, in this study, there was a difference between FDEs and controls for the differential response rates for high complexity compared to medium complexity signatures. This may be a reflection of the superior skill of FDEs and may contribute to their lower error rate.

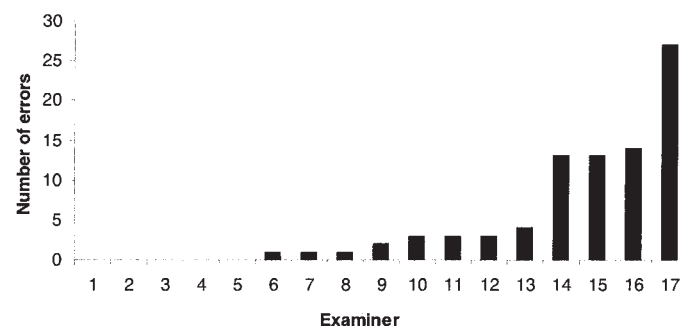


FIG. 2—The total number of errors made by each FDE.

Variation Between FDEs

It was found in this study that FDEs displayed a wide variety of skill. Some FDEs performed exceptionally well while a few performed relatively poorly. As handwriting comparison remains a product of the subjective processes of perception and cognition, some variation between FDEs is to be expected. The normal quality assurance procedures performed in the laboratories concerned may act to reduce the consequences of this. However, the results highlight the need for thorough testing of individual performance with appropriate revision and mechanisms for corrective action when required. It is important for continued judicial acceptability that the variation between FDEs is reduced to a minimum so that the inter-examiner reliability remains as high as possible.

Variation of Results Between Cases for the Group of FDEs

In addition to the variation that we expect from practitioners due to the process of handwriting examination, there is the enormous potential for variation amongst cases that present themselves to handwriting examiners. This case by case variation is due to the large number of variables associated with the available quality and quantity of both questioned and specimen material. Such material includes signatures and extended text. These writings may include an individual's normal writing (either the specimen writer or someone other than the specimen writer), writings that are simulated by a person other than the specimen writer, writings that are simulated by the specimen writer, or writings that are disguised. As can be seen from the results of the current study, where even for the same group of practitioners, examining only signatures where there were no auto-simulations or disguise, there is a large variation in performance for the different cases. The difference in the group results for the different cases within this test adds weight to the position held by a number of authors (5,8,9,14) that no one test will provide a reliable indication of FDE expertise. Factors that seem to have influenced performance and resulted in the variation between cases include signature complexity, line quality, and marked variation in the normal signature of a primary author.

Complexity Analysis

It has been argued that the ease or difficulty with which a signature can be simulated is related to its complexity (12,13,15,16) which is associated with the signature's feature characteristics such as, number of turning points and intersections. Thus the more of these two characteristics there are in a signature, the more complex it is and the more likely it will be that a simulator will have difficulty reproducing it. This, in turn, makes it more likely that an FDE can identify the signature as genuine or simulated. Hence opinions can be made more confidently on high complexity signatures while more caution should be afforded to opinions made on signatures of lower complexity. The results of this study are in accordance with this proposition. Analysis showed that when the signatures were classified as high or medium complexity on the basis of two spatial characteristics (number of turning points and number of intersections of the line trace), FDEs identified significantly more high complexity questioned signatures correctly with less inconclusive opinions. The relationship of signature complexity to correct and inconclusive rates held for genuine and simulated questioned signatures considered independently. It was interesting to note that with all of the complexity analyses, there was no significant difference between the number of wrong opinions made. Although this may be due to the relatively small number of errors for statistical

comparison, maintenance of error rate across complexity ratings is predicted from complexity theory (16). The theory predicts that as complexity decreases, inconclusive rate should increase to maintain an error rate that ideally approaches zero. According to the theory, the called error rate should also remain relatively constant. This was not evident for the group although it was observed for most individuals. We suggest that this may be a consequence of some FDEs making relatively large numbers of errors in regard to genuine questioned signatures that may have arisen due to them confusing natural variation in features with features normally associated with simulation.

While the spatial complexity rating can account for some aspects of the observed results, other factors contribute to the findings. The complexity rating used does not consider line quality (fluency of the signature trace) which is an important factor for consideration in determining authorship. Case 5 comprised high complexity signatures but had the highest inconclusive rate. The signatures of the primary author for this case were of poor line quality, as subsequently independently assessed by two FDEs who did not participate in the trial. It is likely that the poor line quality resulted in the conservative response by the FDEs in this case despite the spatial complexity of the signatures. Case 2 also comprised high complexity signatures but had the highest error rate. In this case the signatures of the primary author were in two forms where both forms shared the same features but one form contained extra features. Due to the randomization process in allotting signatures to the standard and questioned signatures, only one form was included in the Standard set for this case. The difference in features for some of the questioned genuine signatures appeared to result in some FDEs concluding that they were simulations.

Identification of Simulated Compared with Genuine Signatures

FDEs were better at determining whether signatures were simulated than they were at identifying genuine signatures. They gave more correct opinions and less inconclusive opinions for simulated signatures and errors were more likely to be made by calling a genuine signature a simulation than by calling a simulated signature genuine. Kam et al. (10) found that the *a posteriori* error probability was ten times greater for genuine signatures than for simulated signatures and although not statistically analyzed, the same trends are apparent in their results regarding correct and inconclusive opinions. Erroneously expressing the opinion that a genuine signature was produced using a simulation process clearly arises from FDEs incorrectly interpreting features as indicating simulation when in reality those features form part of the natural variation of the specimen signature. Since not all FDEs made these types of errors presumably they may be corrected through appropriate revision and training regimes.

An error where signatures were called simulations when they were, in fact, genuine could be argued to be the lesser of two evils, as the FDE is not directly expressing an opinion that an individual wrote something when they actually did not. According to the definition of terms used in this study, this particular opinion did not exclude the specimen writer as having written the questioned signature. The term "simulation" was, and still largely remains, a confusing term with reference to forensic handwriting examination. This term appears to imply "forgery" to many FDEs and most courts of law. In this study we are not able to tell if, when FDEs incorrectly opined that the signature was simulated, they considered that the signature was not written by the specimen writer. The ability of FDEs to exclude the specimen writer as having simulated their own signature is the subject of future research. It can be said,

however, that the error observed in the present study reflects FDEs misinterpreting features present in the genuine questioned signature as indicators of a simulation process.

We can not exclude the possibility that experimental error caused by the exhaustive task of examining such a large quantity of material (300 signatures in total, with up to 2250 comparisons overall) may have contributed to the error rate observed.

Experience

Our results indicate that in this situation, there was no relationship between the number of years of experience and the degree of accuracy or error rate achieved by individuals. This is despite the fact that one might have expected that the inclusion of three trainees in the sample of participants would bias the results toward showing a positive correlation of skill with experience. We did not show the scattergrams for these data in the "Results" section of the paper as to do so could compromise the anonymity of the participants' results. However, we can say that the range of experience was reasonably wide (1 to 20 years) and the distribution within this range was fairly even. If there was a meaningful relationship between experience and performance for this group, it should have been demonstrable with our sample.

Having said this, due to the small sample size of this study, one must be cautious in extrapolating this finding to the wider population of FDEs. However, the lack of correlation found in this study is consistent with the findings of earlier proficiency tests (The 1987 Proficiency Advisory Committee Comments quoted in 5) and we are unaware of the existence of any evidence that supports the proposition that experience, once an FDE is deemed qualified, is a major factor in determining the level of skill in forensic handwriting examination.

FDEs' Error Rate

The notion that there may be an error associated with a forensic identification discipline is one that may invoke great concern amongst both practitioners and the judiciary. The elucidation of error associated with forensic tasks is further made alarming in view of statements such as "unique," "absolute certainty" and other such phrases in the forensic literature. It could be argued that an error rate greater than zero is unacceptable in the case-work setting. However, it is reasonable to suspect that, as with any other scientific test, an error rate could exist. Kam et al. (6) reported a 6.5% wrong association rate even within a trial structure that only contained individuals' natural writings. In our earlier study on signature comparisons (9) with a smaller sample of FDEs, the error rate was 2%. For the signature comparison study of Kam et al. (10), FDEs' error rate was 7.5% for genuine signatures and 0.49% for simulations. The 3.4% error rate reported here further supports that opinions can be erroneous and the precise value will vary depending on the nature of the questioned signature and the FDEs involved. The interest, for practitioners, should now lie in determining what the most appropriate index is to indicate the extent to which an opinion could be erroneous. The current 3.4% error was calculated on group opinions, using multiple signature types and on varying proportions of complex and moderately complex genuine and simulated signatures. We do know that there was substantial variation within the group of FDEs in terms of correct, incorrect, and inconclusive rates. In addition, the error rates between different signatures, of differing complexity, were found to vary. Presumably if a different set of signatures was provided, it would not be surprising to find that the error rate was markedly different to the

3.4% reported here. An individual's called error rate is potentially a far more useful reference of actual potential error than any score derived from group data. In addition the called error rate may be made more specific by calculating it for different classes of writing. Ultimately an estimate of error will rest on continued application of handwriting and signature trials to FDEs that are composed of random numbers of genuine, disguised, and simulated writings from a large number of individuals. For each class of writings, an individual's called error rate could be cumulatively calculated that could result in an estimate of uncertainty for the technique that could be offered within the judicial setting.

It must be noted that the results presented here have not been through the participating laboratories' peer review system. The error rates, should this have occurred, may be different to those that are presented.

Methodological Limitations

As with any trial such as that described here, there are almost always criticisms that can be raised as to the validity of the trial itself. Limitations of this study which can effect aspects of validity include the relatively small sample size of FDEs, the possible difference in FDEs' performance in the test compared with real case-work situations, and the limited breadth of tasks undertaken. It could also be argued that the incentive for FDEs to perform well was much higher than for the control groups. Financial incentives for the control group were considered by the researchers, however, the possible benefits could not be guaranteed or ruled out as biasing the subjects and studies indicate that reward schedules have no significant effect on control subjects in similar studies (10,17).

Accepting the validity issues, we can state that given the sample provided to the FDEs and lay persons used in this study, the FDEs' opinions concerning the authorship of the signatures were significantly better than the lay group. This provides additional support to previous studies for the existence of real expertise in this forensic discipline.

Conclusion

This study examined the notion of FDE expertise and found support for the proposition that FDEs possess abilities superior to the average person in signature comparison tasks. This expertise was reflected by FDEs significantly lower error rates relative to non-FDE subjects. Additionally, it was demonstrated that FDEs performed better when determining the process of simulation than they were at determining authorship of genuine signatures. FDEs were able to make more accurate determinations on high complexity signatures in comparison to signatures of medium complexity. In spite of the differences found between the FDE and lay subjects, striking features of the results were the variation amongst FDEs (where the performance of a few FDEs adversely affected the error rate for the group) and variation between cases for the FDE group.

References

1. Risinger DM, Denbeaux MP, Saks MJ. Exorcism of ignorance as a proxy for rational knowledge: the lessons of handwriting identification "expertise". *U Penn L Rev* 1989;137:731-92.
2. Kam M, Wetstein J, Conn R. Proficiency of professional document examiners in writer identification. *J Forensic Sci* 1994;39:5-14.
3. Found B, Rogers, D. Contemporary issues in forensic handwriting examination. A discussion of key issues in the wake of the Starzecpyzel decision. *J Forensic Doc Exam* 1995;8:1-31.
4. Galbraith O, Galbraith CS, Galbraith NG. The principle of the "Drunkard's Search" as a proxy for scientific analysis: The misuse of handwriting

- ing test data in a law journal article. *Intl J Forensic Doc Exam* 1995; 1:7–17.
5. Risinger DM, Saks MJ. Science and nonscience in the courts: Daubert meets handwriting identification expertise. *Iowa L Rev* 1996;82:21–74.
 6. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. *J Forensic Sci* 1997;42:778–86.
 7. Moenssens AA. Handwriting identification evidence in the post-Daubert world. *UMKC L Rev* 1997;66:251–332.
 8. Risinger DM, Denbeaux MP, Saks MJ. Brave new “post-Daubert world”—A reply to Professor Moenssens. *Seton Hall L Rev* 1998;29: 405–85.
 9. Found B, Sita J, Rogers D. The development of a program for characterising forensic handwriting examiners’ expertise: signature examination pilot study. *J Forensic Doc Exam* 1999;12:69–80.
 10. Kam M, Gummadidala K, Fielding G, Conn R. Signature authentication by forensic document examiners. *J Forensic Sci* 2001;46:884–8.
 11. Zlotnick J, Lin JR. Handwriting evidence in federal courts—From *Frye* to *Kumho*. *Forensic Sci Rev* 2001;13:87–99.
 12. Huber RA, Headrick AM. *Handwriting identification: facts and fundamentals*. Boca Raton: CRC Press, 1999.
 13. Found B, Rogers D, Rowe V, Dick D. Statistical modelling of experts’ perceptions of the ease of signature simulation. *J Forensic Doc Exam* 1998;11:73–99.
 14. Found B, Rogers D, Herkt A. The skill of a group of forensic document examiners in expressing handwriting and signature authorship and production process opinions. *J Forensic Doc Exam*. In press.
 15. Brault J, Plamondon R. A complexity measure of handwritten curves: Modeling of dynamic signature forgery. *IEEE Trans Sys Man and Cyber* 1993;23:400–12.
 16. Found B, Rogers D. A consideration of the theoretical basis of forensic handwriting examination. The application of “Complexity Theory” to understanding the basis of handwriting identification. *Intl J Forensic Doc Exam* 1998;4:109–18.
 17. Kam M, Fielding G, Conn R. Effects of monetary incentives on performance of non-professionals in document examination proficiency tests. *J Forensic Sci* 1998;43:1000–5.

Additional information and reprint requests:

Doug Rogers, Ph.D.
 Forensic Expertise Profiling Laboratory
 School of Human Biosciences
 La Trobe University
 Bundoora, Victoria, 3086
 Australia
 E-mail: D.Rogers@latrobe.edu.au