The development of a program for characterizing forensic handwriting examiners' expertise: Signature examination pilot study.

Bryan Found, [1,2] Jodi Sita1 and Doug Rogers[1]

Abstract. *Criticisms levelled at forensic handwriting examination expertise have focused on the clear lack of validation evidence offered to substantiate the claims of its practitioners. In general, expertise can be thought of as a skill that is more developed in the specialist than in the lay person. This paper out lines the shift in the process for delineating, and in time articulating, the nature of the expertise claimed within the Australian and New Zealand government and police document examination communities. A pilot study is presented where we compared the opinions regarding the authorship of one hundred and fifty questioned signatures between seven government trained document examiners and eight lay persons. It was found that the government trained document examiners were statistically better at accurately determining the authorship of questioned signatures than were the lay group.*

Reference: Bryan Found, Jodi Sita, Doug Rogers (1999, Vol. 12 – reformatted and reprinted). The Development of a Program for Characterizing Forensic Handwriting Examiners' Expertise: Signature Examination Pilot Study J. Forensic Document Examination, Vol 29, pp. 53 - 59.

Keywords: Signatures, document examiners' skills, opinion, error rates

1. Introduction

Concerns have been raised both in the literature (Risinger, Denbeaux & Sacs, 1989; Risinger & Sacks, 1996), and in the courts (United States v. Starzecpyzel, 1995) concerning the validity and reliability of document examiners' expertise. In the Starzecpyzel case the court found that the field of document examination "has not convincingly documented the accuracy of its findings," and that there was "no strong statistical validation of handwriting examiners' expertise." Clearly, validation is a cornerstone of scientific endeavour and must appear in a form that is more tangible than simply a belief. Since the publication of the Risinger, Denbeaux and Sacs (1989) article, debate over what the existing tests of expertise showed has been fertile. Galbraith, Galbraith and Galbraith (1995) followed up the criticisms raised in the work by Risinger, et al. (1989), focusing on statistical interpretations of previous work, and in addition, provided new evidence that document examiners significantly outperformed both chance and lay people in their ability to correctly identify the authorship of questioned writings. Risinger and Sacks (1996) discussed these criticisms in light of the statistical treatment of the data and experimental validity issues. Common ground amongst the participants in the debate was the apparent limited number of appropriately designed studies, and the small number of document examiners participating. Kam, Wetstein and Conn (1994) introduced a new phase into document examination validation testing by comparing document examiner and lay opinions on a test based on extended questioned text that they administered to both Federal Bureau of Investigation document examiners and college educated lay persons. The text matching test revealed that the FBI examiners were significantly better in identifying writers than were the lay group. This study was followed up by Kam, Fielding and Conn (1997), again using text based writings. In all, over 100 document examiners and 41 lay persons completed the task. They showed that the opinions expressed by lay persons and docu-

1. Handwriting Analysis and Research Laboratory, School of Human Biosciences, La Trobe University, Bundoora, Victoria, 3083, Australia.
2. Document Examination Team, Victoria Forensic Science Centre, Forensic Drive, Macleod, Victoria.

ment examiners were different. The difference was shown to be in the tendency for lay persons to over-associate writings; that is, erroneously conclude that two samples written by different persons were written by the same hand. The most recent evidence would suggest, therefore, that forensic handwriting experts do exhibit expertise that is real and demonstrable, at least at the tasks used in these studies. It is clear, however, that the depth of the evidence supporting asserted expertise, in conjunction with the limited testing of the breadth of handwriting expertise claimed, challenges statements such as that made by Kam et al (1997) that, "The results of our test lay to rest the debate over whether or not professional document examiners possess writer identification skills absent in the general population. They do." If we compare the limited validation evidence available with the level of case-work activity internationally, the inequality should inspire all practitioners to participate in tests that will provide further evidence that may assist in the characterization of their expertise.

Since 1996, the Australian and New Zealand government and police document examination communities have embraced the criticisms regarding expertise characterizations as articulated in the works discussed above. Informal trials commenced in 1996. In 1997 approval was given by the National Institute of Forensic Science, under the direction of the Senior Managers of Australian and New Zealand Forensic Science Laboratories, to conduct routine trials on document examiners. These trials are designed and administered at La Trobe University and are co-ordinated through the National Institute of Forensic Science, in conjunction with the Special Advisory Group (Document Examination). This paper provides an overview of the nature of the testing administered through the presentation of a limited pilot study, the full version of which is to be submitted for publication in 2000. The first five trials will reach their publication cycle towards the middle of 2000. The delay in publication results from the time taken to move the original documentation around the two countries, and the long analysis and debriefing cycles.

The study presented here is a pilot using seven document examiners from one laboratory, out of the seventeen document examiners and six laboratories that ultimately participated. We specifically focused on signature formations, due to the inherent problems they can pose resulting from a combination of stylized characteristics and limited amount of line trace. Signature comparisons, although forming a large portion of the work carried out by document examiners, appear not to be the medium of choice in large handwriting validation studies to date. This study was designed so that subjects were only given the images themselves on which to draw conclusions regarding authorship. No information regarding the authenticity of each questioned signature was extractable from the document itself from impressions, paper analysis, ink analysis, etc. No information was provided regarding the circumstances under which the signatures were made, other than that no further signature specimens were available. Specifically, the aim of this trial was to determine whether document examiners' opinions as to whether each of 150 questioned signatures were written by the writer of the specimens or were the product of a simulation process, were different from the opinions of lay persons.

2. Method

In this experimental study, document examiners and lay people were asked to form an opinion as to whether one hundred and fifty questioned signatures were either genuine, simulated or inconclusive. The identity of the signature in each case was known to the experimenter but not to the subjects. The performance of each subject was scored, and a between group analysis performed.

3. Subjects

Seven document examiners from one government laboratory participated in the study. Eight individuals with no document examination experience, drawn from academic staff and postgraduate students from La Trobe University, were used as the lay group.

4. Signatures

Thirty signatures, executed on blank sheets of A4 paper, were requested from each of ten volunteers who gave the experimenters permission for their signatures to be simulated and used in this study. For the purpose of this study, the providers of the genuine signatures will be referred to as victims. Simulations were made

on blank sheets of A4 paper by staff members of the School of Human Biosciences. These simulations were made freehand, using three randomly selected genuine signatures from each of the ten victims as the models. Simulators were given an unlimited amount of time to practice, and submitted two simulations each: a *one-off* signature which 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 attempt at forging for each victim. The simulations chosen for inclusion into the validation test exhibited what the experimenters considered to be a wide range of skill.

The test given to subjects was divided into two sections for each of the victims' signatures. The first sections comprised fifteen randomly selected specimen signatures from the victims' thirty genuine signatures. The second sections comprised fifteen questioned signatures, which were a mixture of genuine and simulated signatures. The number of genuine signatures included in this questioned group was determined randomly. Each subject was provided with the same fifteen known and fifteen questioned signatures related to each of the ten victims. All signatures provided to subjects were the original inked images.

5. Instructions to subjects

Document examiners were asked to carry out each examination as though it were part of a normal forensic case. They were provided with an answer booklet, which contained the definition of terms used in the study, along with answer sheets. 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. Document examiners were also asked to fill in an information sheet stating the length of time that they had been examining handwriting.

Subjects were informed that the questioned signatures were written around the same time as the specimen signatures. In addition, they were informed that no further specimens were available. An example was provided of how to fill in the answer booklet. No information was given which would indicate the authorship of the simulated and genuine signatures.

Additional information was given to the lay group in order to allow these individuals to appreciate the implications of any opinions that they reached. They were informed that:

1. If you incorrectly assert that a signature is a simulation when it is in fact genuine, this may result in criminal charges being laid upon an innocent person.

2. If you incorrectly identify a signature as genuine when it is in fact a simulation, this could result in a guilty person being found NOT guilty, or could implicate another innocent person in a criminal act.

3. An inconclusive result would not necessarily have any implications with respect to the guilt or innocence of a particular person.

6. Definition of terms used in the study

The following terms were defined for the subjects:

- Genuine: The questioned signature is, in your opinion, written by the same person who wrote the 'genuine signature' group.

- Simulated: The questioned signature is inconsistent with the 'genuine signature' group and displays features that you consider to be indicative of a 'copying' process. Note that this term does not imply that the 'genuine 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.

For the purposes of anonymity, it was agreed that results of individual document examiners would not be presented. In addition, individual document examiners' results did not undergo quality assurance as would be the normal practice of the laboratory participating in the study.

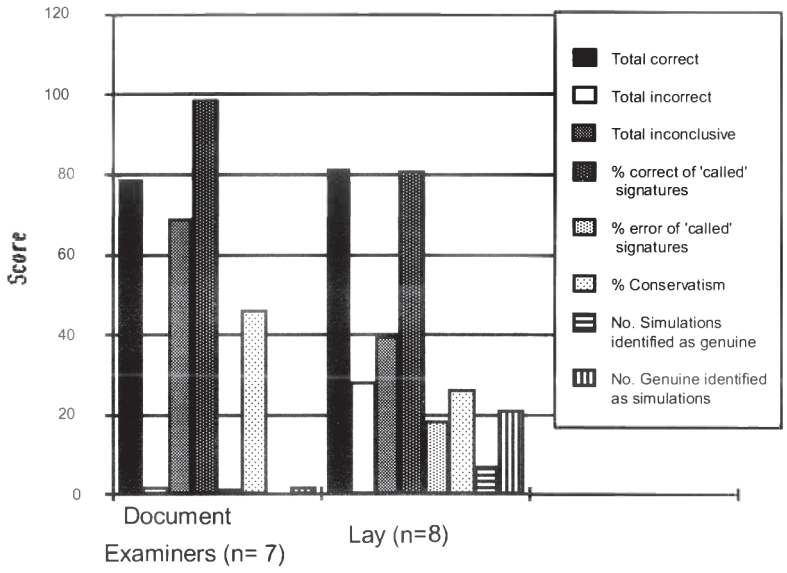

FIGURE 1. Mean results of document examiners and lay persons.

7. Results

Each of the result sheets returned by the participants in the study was marked according to the known answers for each of the comparisons. From this data, eight scores were calculated for each subject in both the document examiner and lay group. These scores were calculated as follows:

- ***Total correct***. This score is the raw number of correct responses from the 150 signature comparisons.
- ***Total incorrect***. This score is the raw number of incorrect responses from the 150 signature comparisons.
- ***Total inconclusive***. This score is the raw number of inconclusive responses from the 150 signature comparisons.
- ***% correct of called signatures***. This score was calculated by dividing the total number of correct·responses by the total number of responses where the subject expressed an opinion (that is where the result was not marked as inconclusive).This score was expressed as a percentage.
- ***% error of called signatures***. This score was calculated by dividing the total number of incorrect responses by the total number of responses where the subject expressed an opinion (that is where the result was not marked as inconclusive). This score was expressed as a percentage.
- ***% conservatism***. This score was calculated by dividing the total number of inconclusive responses by the total number of responses (150). This score was expressed as a percentage.
- ***No. simulations identified as genuine***. This score is the raw number of simulated signatures that were identified as genuine signatures by the subjects.
- ***No. genuine identified as simulation***s. This score is the raw number of genuine signatures that were identified as simulated signatures by the subjects.

Figure 1 represents the mean results of the above scores for the seven document examiners and eight lay persons. The document examiner and lay persons group scores were compared using unpaired, two tailed t tests for the *Total correct*, *Total incorrect* and *Total inconclusive* scores. It was found that there was a significant difference between both the *Total incorrect*

and *Total inconclusive* at $p < .05$ ($p = .0004$ and $.0299$ respectively). No difference was found between the groups for the *Total correct* score.

8. Discussion

Figure 1 provides a summary of the means of the scores for each parameter calculated for both the document examiner and lay group. As can be observed, the profile of these graphs appears quite different with respect to all parameters, excluding the total % correct score. This is confirmed with the non-significant *p* value for the comparison between the groups with respect to the total % correct score. This means that both the document examiner and lay groups, on average, called a similar number of the total questioned signature group correctly. This characteristic was also found by Kam, et al (1997) who stated that the lay group ,'"found as many correct matches as the professionals did - but have declared many non-matching pairs to be matches."

It is the error and conservatism rate that sets these two groups apart statistically. The average error for the document examiner group is approximately 2%, whereas the lay group exhibits approximately a 28% error. Seven percent of the lay subjects' opinions occurred when simulated signatures were erroneously called genuine. No document examiner made such an error. The 2% error associated with the document examiner group was where genuine signatures were erroneously called simulated. This corresponded to an error of 21% for the lay group.

The conservatism rate for document examiners was significantly different from that associated with the lay group. Document examiners 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, further supported by more recent studies by the authors, that the nature of document examiner expertise is best characterized by what they don't say rather than what they do say.

The small number of total errors associated with the document examination group were all signatures that were called simulations when they were, in fact, genuine. An error in this direction could be argued to be the lesser of two evils, as the examiner is not directly expressing an opinion that an individual wrote something when he 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 document examiners and most courts of law. In this study, if the term had meant that the signature was *forged*, then in approximately three of the 150 examinations the experts on average would have produced an erroneous result. The error rate, we would postulate, is the product of the subjective nature of the examination itself and there is no reason why, as with any scientific test, an error rate should not exist. The error rate in this experiment is either the result of a misinterpretation of the indicators of a simulation process that are present in the questioned signature, or simply an 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).

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. The error rate given here cannot necessarily be applied to casework in general due to experimental validity issues. Galbraith, et al. (1995) in their article assessing the treatment of handwriting test data in the article by Risinger, et al. (1989), used the definitions of experimental validity types as articulated in Cook and Campbell (1979). Although it was argued by Risinger and Sax (1996) that the Cook and Campbell (1979) framework was not appropriate to discuss the validity issues associated with the trials under scrutiny, the general ideas behind these validity issues still apply. In this particular study, the sample of document examiners can be rightly criticized as being small. We are hesitant to apply these results across the population of document examiners in general. Inspection of the recently calculated results for the larger group confirm this. In terms of construct validity (did our test measure what we set out to measure), it is always difficult to assess in investigations of this type. The greatest threat to construct validity for tests of this type and proficiency tests in general, is that the test itself may alter the subject's normal approach to the examination which

could produce results which are, to a certain extent, artificial and unlikely to reflect the normal range of results put out for similar examinations. Indeed, similar sources of error can be associated with the lay group whereby the seriousness with which they took the test was unable to be assessed objectively. Threats to the internal validity of this test were reduced by all participants agreeing individually to participate in the study, and by all participants returning their answer sheets. The size of the test was a concern and could be considered, to some extent, to be intrusive. However, all subjects were given an unlimited amount of time to carry out the examinations to reduce the likely effect of this threat.

In terms of external validity, a number of points need to be raised. We have no evidence that the results generated by either our lay group or our document examiner group are able to be generalized across the possible population of these individuals. External validity issues also preclude us from concluding that the accuracy rate exhibited by this group of experts can be taken to approximate the accuracy rate which would be achieved in normal casework. It may be better, worse, or the same. From a single study of this type, this rate cannot be accurately determined.

Accepting the validity issues, we can state that given the sample provided to the document examiners and lay persons used in this study, the document examiners' 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.

One of the more interesting aspects of designing validation tests in this field is that it is unlikely that any one test, regardless of the number of participants, will ultimately provide a conclusive answer as to whether the expertise claimed by the field really exists. This arises on a case by case basis due to the enormous number of variables associated with the available quality and quantity of both questioned and specimen material. For example, document examiners may outperform lay persons when extended text, written in an individual's normal handwriting, is provided for them to match. The reality is, however, that document examiners, in order to express an opinion regarding handwriting, must consider writings that are other than natural, such as writings that are simulated by a person other than the specimen writer, writings that are simulated by the specimen writer, and writings that are disguised. Validation trials that do not incorporate such writings are of little use in characterizing document examiner expertise at the case-work level. In addition, the usefulness of tests would be enhanced by ensuring that all trials are carried out as a structured questioned-to-specimen process as it is done in the forensic setting.

Handwriting comparison remains a product of the subjective processes of cognition and perception. In addition to the variation that we expect from practitioners arising from this reality, is the enormous potential for variation amongst cases that present themselves to handwriting examiners. In spite of the long history of this field, forensic hand writing comparison remains plagued by the lack of accepted theory, the lack of objective comparison techniques, non-uniformity in reporting procedure, and a lack of fundamental guiding research. These different shortfalls can and will be addressed in the medium-to-long term. Given that the evidence continues to be delivered to courts of law, the only short-term measure is to focus on the provision of appropriate evidence as to examiner expertise and possible error rates. The authors believe that once this process begins, as it has in our document community, forensic handwriting examination will irreversibly shift from a culture of faith to one more closely resembling a science.

9. References

Cook, T., & Campbell. (1979). Quasi-Experimentation: Design & Analysis for Field Settings, Chicago: Rand McNally.

Galbraith III, 0., Galbraith, C.S., & Galbraith, N.G. (1995). The principle of the "Drunkard's Search" as a proxy for scientific analysis: The misuse of handwriting test data in a law journal article. *International Journal of Forensic Document Examiners*, 1, 7-17.

Kam, M., Fielding, G., & Conn, R. (1997). Writer identification by professional document examiners. *Journal of Forensic Sciences*, 42, 778-786.

Kam, M., Wetstein, J., & Conn, R. (1994). Proficiency of professional document examiners in writer identification. *Journal of Forensic Sciences*, 39, 5-14.

Risinger, D.M., Denbeaux, M.P., & Saks, M.J. (1989). Exorcism of ignorance as a proxy for rational knowledge: The lessons of hand writing identification "expertise". *University of Pennsylvania Law Review*, 137, 731-792.

Risinger, D.M., & Saks, M.J. (1996). Science and nonscience in the courts: Daubert meets handwriting identification expertise. *Iowa Law Review*, 82, 21-74.

United States v. Starzecpyzel, 880 F.Supp. 1027 (S.D.N.Y. 1995).