JOURNAL OF FORENSIC DOCUMENT EXAMINATION VOL 29 2019

Contents

Journal of forensic document examination (Online)
ISSN 0895-0849

Journal of forensic document examination (Online)
ISSN 0895-0849

FROM THE EDITOR...

Standing on the shoulders of giants, we gaze forwards.

This issue is very important for the **Journal of Forensic Document Examination** (**JFDE**) community, since it signifies the next step in the evolution of the journal. Already, the 2018 issue was published online by PKP Press. This started a series of interconnected actions that culminated in the current format and method of peer reviewing system. As you might have already seen, the website was fully overhauled with special focus on the detailed and up to date Guidelines for the Authors. The current format of those Guidelines is fully compatible with the online scientific data bases and aims at a scientific homogeneity of the included papers. Also, the Editorial Team has evolved, focusing on a multidisciplinary approach on the discipline of Forensic Document Examination. Currently, amongst our ranks, are forensic document examiners, academics, neuroscientists, pattern recognition and machine learning experts as well as legal scientists. Furthermore, striving for scientific and academic worldwide interconnectedness, we have recruited experts from both the New and the Old World, working in universities, laboratories and institutes worldwide.

Moving on, we have introduced the use of OJS platform, through which we now operate in a double-blind peer review system for each contribution, with neither the authors nor the reviewers knowing the identity of others. Already the platform and the system have been tested and from the issue of 2020 and on, all papers will be reviewed and published in this manner.

Having ushered the Journal into a new threshold, we are fully aware that neither the **JFDE** nor the Forensic Document Examination as a discipline would not be in the current state of art without the valuable contribution of the late Bryan Found, PhD.

Since the late 1980s, Dr. Bryan Found had accomplished more than any other researcher in the world to develop the science of handwriting identification. He had been an unrelenting advocate for not permitting biasing or context irrelevant information to enter into forensic handwriting examinations. Dr. Found had been invited to over 20 countries to present workshops on the science of handwriting individualization and on human factors. Most recently he was invited to be a speaker for a plenary session at the International Symposium on Forensic Science Error Management sponsored by the National Institute of Standards and Technology (NIST) in July 2015. He, along with his colleagues, had published over 40 peer reviewed forensic scientific journal articles, including in the **JFDE**, 44 conference abstracts, and three invited book and encyclopedia chapters. Dr. Found was during the end of his life the Chief Forensic Scientist at the Victoria Police Forensic Services Department in Australia, one of the world's largest multi-disciplinary laboratories, where he strived to maintain the highest standards for forensic laboratories. These standards include educating practitioners, staff members, investigators, and attorneys about cognitive factors that include the potential impact of exposing practitioners to domain irrelevant context information. One could only wonder what further contributions would he add to science, if he was still alive today.

This issue is a compendium of several very important papers by Dr. Found and his colleagues - most often Doug Rogers - at LaTrobe University in Melbourne, Australia, that we believed made a significant impact to the scientific development of handwriting identification as we know it today. These publications, along with the *Modular Forensic Handwriting Method* (JFDE, Vol 26), and the interview titled, *A Discussion of Issues Around Human Factors And Bias In Forensic Handwriting Examinations: The Present And Future For Practitioners* (JFDE, Vol. 25), encapsulate his importance for our discipline.

A main purpose of this compendium is to educate the researchers, field practitioners and students about Dr. Found's critical contribution on the research that has led to where we are today and culminated in the NIST report, scheduled for publication in 2020, as well as create a chronological perspective of his work. However, the reader should not think that the collected papers have only a historical value. On the contrary, the analyzed subjects are today as important as they were the time they were authored.

The first paper published in 1995 titled, *Contemporary Issues in Forensic Handwriting*

Examination. A Discussion of Key Issues in the Wake Starzecpyzel Case, is perhaps the most influential paper in this issue, let alone heretical at the time it was published. In this paper, Bryan urged Forensic Document Examiners to accept the criticism of their field, mainly focusing on the Southern District of New York Federal Court's Judge Lawrence W. McKenna's decision in U.S.A. v. Starzeckpyzel that the handwriting identification was not a science, but a technical skill. Dr. Found encouraged the use the court's decision as a springboard for further scientific evolution to revisit and to reinvent Forensic Document Examination as a more robust identification science.

It must be noted that Found, himself, stated that the initial response of the forensic world to this paper was *mostly suspicion.* For all practical purposes, the lack of serious scientific research in the United States on handwriting identification at that time, coupled with the lack of awareness of the majority of FDEs in the U.S. regarding research that was going on in Australia, New Zealand, and in the Netherlands, proved exactly his points. It was the Association of Forensic Document Examiners and the **JFDE** that welcomed Dr. Found's and Huub Hardy's (Netherlands) more scientific approach to document examination. This is one of the reasons, Found and his colleagues were frequent contributors to the **JFDE** that published the first Modular system in the 1999 issue and the latest in the 2016 issue.

In his work, Dr. Found focused much effort on the subject of cognitive bias. According to him, bias is the biggest source of errors, where humans are involved. Characteristically, he notes, *There is no shame in making errors, the only shame is not understanding the systems that caused them, not learning from them and not having mitigation strategies in place to avoid them in the future* (**JFDE,** Vol 25). Part of his and his colleagues' approach towards evolving strategies to avoid bias is highlighted in the papers *Matrix Analysis: a Technique to Investigate the Spatial Properties of Handwritten Images*, where the authors' research on objective measurement strategies to assist experts to make judgements about spatial consistency is described, and the paper, *The Objective Static Analysis of Spatial Errors in Simulations,* which deals with the objective spatial error scores resulting from measurement of forged and genuine signatures.

Another major contribution of Dr. Found in the field is his research on the assessment of the complexity of handwritten images that culminated in the Module 5 of his magnum opus, the *Modular Forensic Handwriting Method* (**JFDE**, Vol. 26), which must be noted is one more procedure to reduce bias and error in the case work of examiners. His insights regarding assessing complexity are analyzed in the paper, *Statistical Modelling of Experts' Perceptions of the Ease of Signature Simulation.*

But above all, Bryan Found was a stout defender of the scientific value of Forensic Document Examination. He stated many times that there is real expertise associated with being a handwriting specialist in a forensic environment, as it is demonstrated in the paper, *The Skill of a Group of Forensic Document Examiners in Expressing Handwriting and Signature Authorship and Production Process opinions.* Furthermore, his research has proven that when testing the abilities and claims of the FDEs and comparing them to laypeople, it is evident that the skill of the handwriting examiners is real and - most importantly - this skill can be demonstrated. This important subject is thoroughly discussed in the paper, *The Development of a Program for Characterizing Forensic Handwriting Examiner's Expertise: Signature Examination Pilot Study.*

Finally, no compendium would be complete without including Dr. Found's lynchpin paper, *Comparison of Document Examiners' Opinions on Photocopied Signatures* originally published in the JFDE in 2001. This paper is one of the more widely referenced papers in the field of forensic document examination.

Michael Pertsinakis, LL.B., Ph.D., MCSFS
Editor

Contemporary issues in forensic handwriting examination. A discussion of key issues in the wake of the Starzecpyzel decision

Bryan Found[1], Doug Rogers[1]

Abstract: *A considerable amount of attention has been focused on the field of forensic handwriting examination as a result of a recent Daubert hearing regarding the admissibility of forensic handwriting testimony (United States of America v. Roberta and Eileen Starzecpyzel, 1995). The findings of the hearing provide us with an opportunity to reflect on some of the basic postulates and practices associated with the field, particularly as they are perceived by individuals working within mainstream scientific paradigms. It appears that there are some postulates that are still mounted as underpinning forensic handwriting examination that defy even basic logic when seen in the environment of normal behavioural sciences. Rather than dwell on the possible reasons for this phenomenon, a few basic alternatives to the current explanation of theory and practice will be overviewed. Although what is presented here is largely 'theoretical' in nature, it does provide a framework which currently forms the focus of our research. Ultimately, the question as to whether what we do can be regarded as science or a practical skill falls within the frame of reference of those who choose to define those terms. What is important is not that we waste time and effort arguing over the details of which group we belong to, but rather that we concentrate on improving the paradigm within which we all work. The first step in this process is defining what the paradigm is.*

Reference: Bryan Found, Doug Rogers (1995, Vol. 8 – reformatted and reprinted). Contemporary issues in forensic handwriting examination. A discussion of Key Issues in the Wake of the Starzecpyzel Decision. J. Forensic Document Examination, Vol. 29, pp. 5 - 22.

Keywords: Daubert Hearing, feature detection, expressing opinions, complexity theory, similarities, differences

1. Introduction

"The Daubert hearing established that forensic document examination, which clothes itself in the trappings of science, does not rest on carefully articulated postulates, does not employ rigorous methodology, and has not convincingly documented the accuracy of its determinations" (US v. Starzecpyzel, 880 F. Supp. 1027, [SDNY.1995]). This statement highlights major problems associated with the field of forensic handwriting examination. It includes criticizms that have been made in other articles (Huber & Headrick, 1990; Risinger, Denbeaux & Saks, 1989) concerning the science of forensic handwriting examination and associated issues of method and validation. There is yet to be a standard text from which we have been able to extract clear statements of what can reasonably be said about handwriting, together with an accompanying theoretical basis and a study of validation. Nevertheless, in the 'Memorandum and Order' Judge McKenna stated that "Saks' testimony established that there is no strong statistical evidence supporting or disproving the 'two fundamental principles' or the reliability of forensic document examination". There is not, therefore, a suggestion that the practices of

1.National Forensic Handwriting Consultancy, Handwriting Analysis and Research Laboratory, School of Human Biosciences, La Trobe University, Locked Bag 12, Carlton South, Victoria, 3053, Australia

the field can't be done, but that there is a lack of an accepted theoretical basis on which we conduct our work and an absence of proof of our reliability. If we are to be recognized as adhering to the process of science, this theoretical basis must be supported by appropriately designed research, and the application of the resulting theory must then be validated. In the scientific environment, validation studies do not refer to case examples or even the features associated with known forgeries, for example, but rather to extensive and realistic tests of examiners to produce the correct result when the true answer is not known to them. There is no question that there has been a significant lack of these classically-designed validation trials. As a profession we are responsible for this shortfall and should heed the criticism, regardless of its source, in a professional manner.

Primarily it appears that what is currently most inappropriate is the image of what is done in the field under the banner of science. This is reflected in the statement that, "The problem arises from the likely perception by jurors that FDEs are scientists, which would suggest far greater precision and reliability than was established by the Daubert hearing. This perception might arise from several sources, such as the appearance of the words s*cientific* and *laboratory* in much of the relevant literature, and the overly precise manner in which FDEs describe their level of confidence in their opinions as to whether questioned writings are genuine." Unfortunately, there is an underlying assumption that all document examiners conduct the work on the same basis that was suggested in this hearing. We do not, and certainly do not suggest that all others do.

The creation of a science of handwriting analysis as was suggested by Judge McKenna is, although young in forensic terms, already having an impact. An example of this type of approach comes from the joint conference of the International Graphonomics Society and the Association of Forensic Document Examiners held in Canada in 1995. In addition, measurement techniques and criteria developments specific to forensic handwriting examination have been reported on (Cheung & Leung, 1989; Baier, 1995; Found & Rogers, 1995; Found, Rogers & Schmittat, 1994; Found, Rogers, Metz & Schmittat, 1994). Part of our treatment of handwriting examinations has been to attempt to standardize and document handwriting methodology (Found & Dick, 1992; Found, Dick & Rogers, 1994; Metz, Found, Dick & Rogers, 1995). The primary change to existing technique has been reporting procedures, which necessarily have been made to reflect both the considerable limitations associated with the type of material being examined, and the need for clarity of meaning of opinion in the court environment. This process is, of course, very slow due to the normal resistance to change, lack of research time and money and suitably qualified individuals devoted to the field. As was noted, "...this discipline has no counterpart in industry or academia with an economic incentive to study and refine its scientific basis"). The handwriting examination component of document examination has largely drifted and not developed at the rate that normal science would have expected. The field of forensic handwriting examination, for these and other essentially theoretical reasons, falls well short of the *identification science* that it has commonly been perceived to be. Indeed, evidence based on the outcomes of human movement cannot and should not in any way be paralleled to forensic fields such as DNA and fingerprints. It could be argued that the severity of criticism that we have been subjected to is probably related to the power that this branch of forensic science has claimed. The claim is simply not supported in theory, nor have we supplied the evidence in practice. We as a group are responsible for this reality. We are, however, like those before us, only transient in this process. We have a choice to either participate in reconstructing and validating the discipline such that its value, if we find it to have value, is maintained for those who follow us.

This paper aims to deal, in a general way, with some of the issues raised in the Daubert hearing which impinge upon the above major concerns or criticisms. To review in any exhaustive fashion what was said in that hearing, as well as the conclusion of the court, would be too extensive a task to explore here in any meaningful way. Indeed, we found it an impossible task as handwriting specialists, given that much of the questioning was based on statements of underlying beliefs and reporting formats that, although they appear to have gained general acceptance in the forensic community, we do not agree with.

2. Identifying key issues

There were a number of issues brought out during the hearing that we feel should be developed. These are the notion of individual and class characteristics, issues regarding similarities and differences and reporting procedures. Our aim is not to analyze the question and answer process associated with the hearing, but rather to discuss in general the criticisms that were raised in the context of the rationale for the methodology of the forensic comparison conducted in our laboratory.

3. Class and individual characteristics as a basis for handwriting opinion

The belief in the notion of class and individual characteristics has remained a pillar in forensic handwriting examination and appears to be used as the fundamental basis by which handwriting examiners claim they can identify an individual (Conway, 1959; Harrison, 1958; Hilton, 1982; Osborn, 1929). The following passage, extracted prosecutions handwriting expert, also indicates that it underpins the evidence she was giving: " ...you have a familiarity with the copy book standards that are being taught and you can evaluate the letter forms on how much they diverge from the standard to get an idea of how unique that is. You also understand the uniqueness of different letter forms or a particular quality of a writing based on the study you have done of the literature and of the treatises and once again drawing on your own experience in previous cases that you have also examined." The fundamentals of the class/individual theory are represented here by the notions of copybook form, divergence from the form and the assessment of uniqueness of characteristics based on experience. We do not intend to exhaustively restate the theory here, as it can be found in various forms in most of the standard texts in the field. However, basically it is claimed that the validity of a document examiner's opinion is based on his or her ability to distinguish between what are class and what are individual characteristics. There is some sort of assessment of the uniqueness of the features based on an individual's knowledge of character manifestations and combinations in the population. It has been argued, however, that although it appears to make sense superficially, there is limited theoretical and/or practical support for it (Lacey and Dick, 1992). Some of the problems with the theory as we perceive it are outlined below.

1. No evidence has been provided that experience from doing forensic casework increases the examiners ability to differentiate between class and individual characteristics.
2. No evidence has been provided that experience increases the validity of findings.
3. Even given this theory, some handwriting specialists believe it is possible to examine and express opinions as to the authorship of foreign writings.
4. Given this theory for handwriting, signatures are somehow included, even though they may exhibit no class characteristics whatsoever and the uniqueness of the features in the image have no way of being assessed according to the theory.

So how is it that such a theory has survived? It appears to make sense when explained to the layperson and it provides a platform on which expertize can be claimed and on which one's position within the field can be improved. In addition, it is not directly falsifiable, as we have no database on which an individual's judgment of uniqueness can be validated. The theory can, however, be indirectly tested. The simple test for any person claiming to have the knowledge base to construct an analysis on the basis of this theory is as follows: 1.) Select two equally experienced examiners from a forensic laboratory and provide them with the identical sample of handwriting of a number of individuals where the class system is known. 2.) Ask them to individually determine each of the class systems and then list and rank the individual characteristics according to their degree of uniqueness in the population. The results could then be compared and correlated. We think that the results would not justify the apparent enthusiasm for the theory. This type of validation trial has been discussed with numerous document examiners and yet there has been no race to conduct the experiment.

There is, however, a place in the profession for class/individual theory. We apply the theory for the purpose of explaining to the lay person the process by which inter-and intra-writer variation emerges. There can be no reasonable grounds on which to doubt that handwriting is normally learnt in the first instance by reproducing a copy book system. It is common knowledge that individuals introduce into their writing, either consciously or subconsciously, additional features or modifications on that copy book form for a whole range of reasons such as increased speed production, incorrect adherence to the system, changing the aesthetic, qualities etc. The problem arises as a result of the belief by some handwriting experts that they can retrospectively determine the source of the components of the graphemes and then claim that divergent characteristics can be subjectively weighted as to their respective uniqueness and individualising power. Since the theory is not supported logically, is able to be tested but has not been, and the basis of opinion relies on information that is individual specific and not falsifiable, it does not sit easily within a scientific paradigm.

We suggest that we can do no less than either modify or abandon this theory. But is there an alternative theory which can be validated and on which opinions can be mounted that makes sense? The reality is that there may be a variety of theories that could be proposed. We choose to rationalize the examination process and the underlying logic, not according to the determination of significant individualizing characteristics, but rather to the determination of overall similarity or difference associated with observable features and basic relationships which are thought to exist between the underlying physiological mechanism responsible for the image, the variation that is observed in image production in the population, and the observed difficulty that individuals have in copying complex movements.

4. Modifying the theory

We can propose a basic model of the forensic comparison method that we conduct in our laboratory, a simplified version of which is represented in Figure 1. Fundamentally it is a comparison where the resultant first stage of the opinion, similar to traditional approaches, deals with the notion of similarity or difference. It is fairly straightforward to advance a plausible explanation once we have made a decision about this if one is able to be made. The decision arrived at should be understandable, logical and illustratable to any impartial person. The legal system rightly tends to focus on this stage of the examination because of its subjectivity and the resultant implications to the conclusions regarding the dispute. The notion of significant similarity or difference will be elaborated on later. There are a number of aspects of this particular process that we feel should be discussed.

One of the most difficult aspects when reflecting on visual comparison processes is to explain exactly how it is that our brains are processing the information that we are providing it with. There is great difficulty in verbally describing what our brains judge to be similar or different. Since we are dealing with a visual phenomenon, sense can only be made of the concept according to visual illustration. Semantic gymnastics on this point, of the type observable in the Daubert hearing, result directly from this phenomenon. There did appear to be some confusion at this point regarding distinguishing inter-writer differences from natural variation. However, in terms of the approach outlined here, this is not the stage where that distinction is considered. Decisions as to overall similarity or difference are about all of the elements of the image, from line details, character constructions, character combination constructions, word constructions and features associated with the entire text. No significance as such is attached to this opinion. Judge McKenna did not dispute the ability of document examiners to express an opinion regarding this stage of the examination: "Although Ms. Kelly was unable to explain to the Court's satisfaction precisely how significant similarities or differences were identified, the Court has no doubt that such identifications can be performed, in some cases by cursory examination." Attempting to verbally describe this process is analogous to describing the difference and similarities between two paintings of an identical scene, but where specific paintings have not been provided to the audience.

Confusing the issue of image comparison is the usual tendency of both document examiners and the legal fraternity to focus discussions on the process in terms of character formations. We talk about *g*

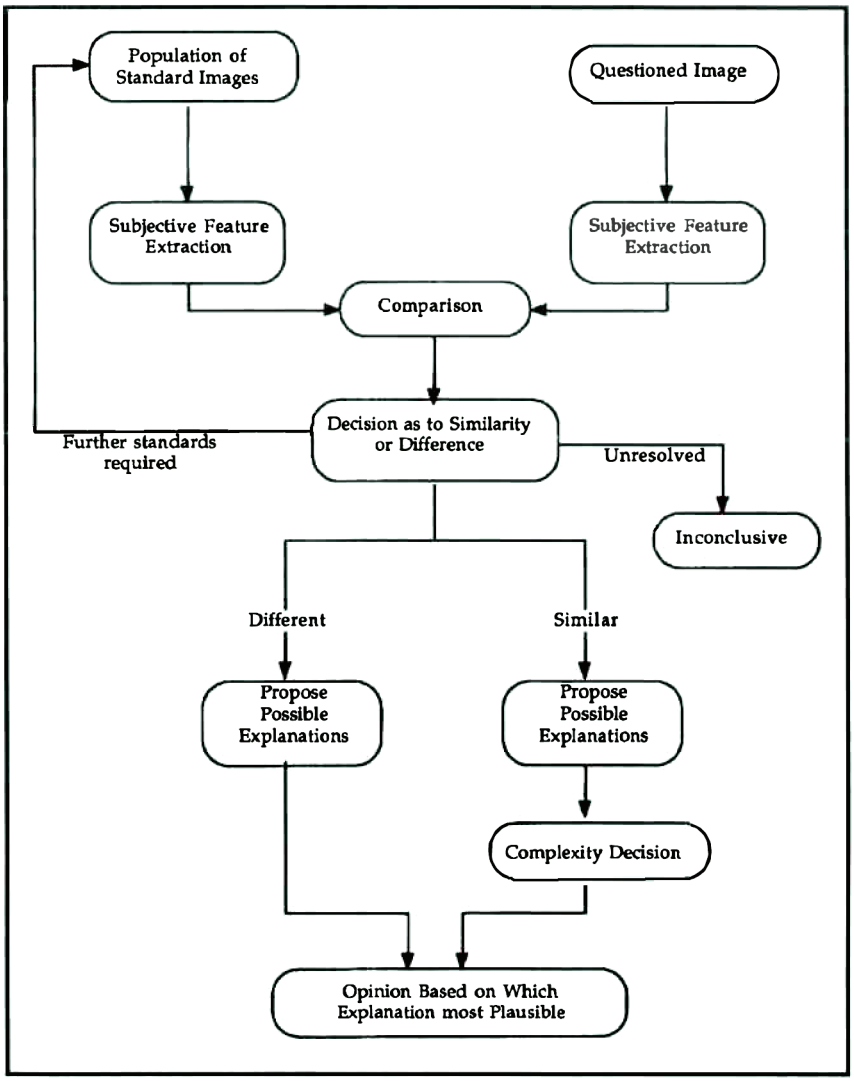


Figure 1. Flow diagram of the stages in the forensic examination of handwriting.

formations, *e* formations, proportions of staffs on different characters, etc. Writing is thought about in this way because of the relationship between the static image and its purpose, which in most instances is for the writing to be read.

It is most convenient, of course, to structure the analysis process according to these visual/linguistic cues. However, there is the hazard that this emphasis could be misconstrued to mean that these characters form the fundamental basis of the examination and opinion process. Characters can be considered the middle ground of the overall comparison and provide us with a reference point to make the comparison process manageable, particularly when we have extended text. Clearly, we are making inferences at the first stage of the examination process as to whether the images, the artifacts of the human movement system, are the product of similar or dissimilar movement commands. The characters themselves are simply fabrications of the movement system, given significance only in light of their value in communication. This is, of course, not what forensic handwriting examination is about. We are attempting to determine whether any meaningful statements can be made purely on properties associated with the movement outcome itself. The purpose of mentioning this at this stage in the discussion is that the misunderstandings associated with this concept continue to support the enthusiasm for the construction of handwriting characteristic databases. Databases, when appropriately constructed and used, can be very powerful, particularly in systems where the construct characteristics of the file are easy

to isolate, where the substance being filed has relatively invariant properties, and where the population being sampled is relatively static and invariant itself (such that the database can be considered a reasonably representative sample). DNA and fingerprints, to differing extents, suit a paradigm revolving around significance determined from databases. Handwriting, however, does not. The nature of handwriting as a changeable outcome of learnt human movement violates each of the requirements for databases if we choose to use them to make statements concerned with statistical individualizing power. Databases such as the FISH system, which have been in use for many years as recently described by Baier (1995) and Hecker (1995), have not been reported to be used for this purpose and should not be erroneously included in this debate. We can only guess at the court's response to quoting frequencies of characteristics and significance in view of these limitations.

5. Feature detection

Our method is underpinned by an approach to the examination of handwriting which we have coined *feature detection.* Feature detection is based on the rationale that, under normal conditions, given a sufficient amount of writing, no two skilled writers are likely to produce handwritten images that are exactly the same in terms of the combination of construction, line quality, formation variation and text structure features. This statement is different from that offered as one of the two basic principles in the Daubert hearing that, "no two people write exactly the same way." The underlying principle associated with this theory is quite appropriately heavily qualified, and the limitations which impose this qualification should be expressed along with any findings. Basically, however, we would argue that if we were to select at random any number of extended handwriting samples from the general population, the incidence of samples that share exactly all combinations of features should be low. There is evidence for this, although criticisms regarding this notion not having been proved in a scientific way are quite valid. The basis of the working hypothesis of inter-writer difference comes from a variety of sources:

1. That handwriting is a learned behaviour involving very complex manipulations of muscles by the nervous system. As with any skilled movements, people are observed carrying them out in different ways to achieve what are often very similar goals; e.g., playing a sport, talking, playing a musical instrument, painting, etc. The reality is that it is accepted that the outcome of these movements differs from person to person and in the skilled 'mover' may result in a movement style that is to some extent characteristic. It is no different with handwriting. We commonly see evidence of this through the course of our lives, through recognizing the writing of our wife or husband or workmate. The problem is that we do not have significant numerical support for this notion.

2. There has been no report of extended writings that are exactly the same, even though the field of forensic handwriting examination has been operating in an organized way for decades, and databases of handwriting samples are kept in some form at many government laboratories. Databases associated with anonymous letter files are routinely searched on some basis. Those who have had to carry out this process indicate that it is not difficult because of the vast array of ways that writing presents itself.

3. If the handwriting of individuals was commonly similar and the pictorial results of the movements were easy to reproduce, the commercial world should have experienced anarchy by now as a result of the ease with which funds could be fraudulently withdrawn.

4. Instruments such as the FISH system would be of no value, as the search strategy which relies on healthy inter-writer variation would invariably throw back at the operator an unmanageable sample of potential hits. Yet the FISH

system has survived and research using the databases is still being reported (Baier, 1995; Hecker, 1995). Other handwriting classification schemes have also been developed (Hardcastle & Kemmenoe, 1990; Hardcastle, Thornton & Totty, 1986), the latter of which contains further references to these schemes.

5. We would not see the level of research being carried out on optical character recognition devices. Examples of this type of research appears in the Proceedings of the Third International Conference on Document Analysis and Recognition, 1995. One of the major problems for these devices is the vast number of ways that handwriting presents itself, both inter- and intra-writer.

6. There have been a number of studies carried out that, although they focus on only a small quantity of writing and only a limited number of characteristics, still provide evidence of this variation (Eldridge, Nimmo-Smith & Wing, 1985; Livingston, 1963; Muehlberger, Newman, Regent & Wichmann, 1977; Franks, Davis, Totty, Hardcastle & Grove, 1985; Wing & Nimmo-Smith, 1987).

7. Perhaps one line of support for this notion of inter-writer variability is the inability to support the alternative hypothesis that most people write the same as each other. One can only wonder what a court's response would be if we stood up and claimed that most people write exactly the same as each other. This flies in the face of common knowledge.

8. There have been reports of large-scale handwriting searches that have been successful in isolating individuals, even on the basis of limited comparison strategies (Baxendale & Renshaw, 1979; Harvey & Mitchell, 1973). Although there are only a small number of published reports, strategies of this type are not uncommonly put to use.

It is not unreasonable to accept inter-writer variation as a working hypothesis, even though it has not been delineated mathematically. In addition, the second principle stated in the Daubert hearing that, "no two people will write exactly the same when repeating," although it should not have been stated in such absolute terms, is able to be observed and reasonably explained. It results from a combination of an individual's motor output varying to different extents due to the non-muscle specific nature of the movement's representation in the brain (Van Galen, 1980), personal tolerances of motor output, the relative position of the movement system when the entry is to be executed, changes associated with particular character combinations, or conscious changes to the movement process.

The inter-and intra-writer variation can be thought of as a product of these factors. Given this breakdown, it is not surprising that persons in the general population recognize easily familiar writings and routinely conduct their own handwriting examinations. We could argue that it is elements of the *picture* of the writing that their brains are comparing to a given number of known *writing pictures* stored in memory. These writing pictures are laid down by constant exposure to the handwriting of others. For this recognition to be achieved, the brain must be making a decision based on *pictorial features*, or more probably a range of them, within the writing. In this situation the brain may be excluding alternate pictorial memories where the features do not match, in favour of those that do. It is plausible, therefore, that the brain is making decisions based on those features that pictorially characterise the writing.

This process is relatively straightforward for a member of the general population, as only a limited number of pictorial memories are referred to and an incorrect judgment may have no implications. The writing is then either judged as *known* or *unknown*. Handwriting examiners are faced with a different situation in that every sample of writing submitted is unknown. Collected or requested handwriting standards are then used to form a working knowledge of the writer suspected of writing the questioned entries. The gathering of handwriting standards is covered adequately in the texts and will not be further discussed here. The question is, 'On what basis is the

handwriting being compared and what is the nature of the expertise that is claimed?'

Feature detection rationalizes that, given an adequate quantity of skilled standard and questioned writing, the brain can perform an analysis of the standard writing and determine either visually or using magnification, spatial features or line quality features which contribute to the writing's pictorial character. It is these features that are being compared to the questioned writing. It is on the basis of these features that the primary opinion is formed as to whether the body of questioned material is similar to or different from the body of standard material. There is no speculating as to whether characteristics are class, individual or a combination of both.

6. Similarity or difference

Perhaps one of the most confusing of concepts in our field is the explanation of what in writing constitutes a similarity or a difference, particularly in light of the variation phenomena. In terms of our model, we define the terms generally; similarities are pictorial or structural features that appear consistent between the populations of questioned and standard images. The similarities can be observed in terms of the way the strokes are concatenated into letter, letter combinations and word formations, the features that can be described, and the relative placement of images. Differences are pictorial or structural features that appear dissimilar between the populations of questioned and standard images. The dissimilarities can be observed in terms of one, or combinations of the way the strokes are concatenated into letter, letter combinations, word formations and the features that can be described. The criteria for features to be described as different is that they are fundamental to the pictorial or structural character of the writing and are not shared between the bodies of questioned and standard writings. Examples of differences would be a character which is consistently constructed in a different way between the questioned and standard images, or where the line quality is visually dissimilar between the questioned and standard images etc.

Clearly, these definitions do not address issues of authorship. What they do, however, is to focus the examination on the appropriate set of hypotheses. What is important is that a decision at this stage in the methodology is illustratable. In many instances, the comparison process stalls at this point and a reasonable opinion cannot be formed as to difference or similarity. This results in an inconclusive result.

At this stage there is simply no numerical answer as to what is an adequate amount of known or questioned handwriting. This remains another limitation of the examination technique which must be respected. We can, however, show in specific examples why in our opinion there is an insufficient amount and why in another example there is sufficient.

7. Expressing opinions based on observations of similarity or difference

Given that we have subjectively formed an opinion as to whether the questioned material is similar to or different from the standard material, we can now propose explanations that could account for that primary observation. The ultimate aim is to express an opinion as to which of the alternative explanations is the most plausible. This process should always be carried out in an environment where no other peripheral information is taken into account. Peripheral information belongs to the investigators and to the courts. We can certainly be asked in the courtroom how certain factors may effect handwriting, but this should not contaminate our perception of what we can reasonably accomplish dealing solely with the handwritten images. Let us consider a typical example.

Imagine that we have performed an analysis of a questioned signature. The opinion of the examiner is that there are no differences in the line quality, construction or spatial characteristics when compared to the population of standard material. We could conclude from this that the image is the product of the same or similar movement commands or different movement commands that produced what would appear to be an artifact consistent with the population of standard signatures. These statements are not about authorship. We can develop on these statements to propose three explanations or hypotheses to explain these similarities in terms of authorship. This section of the method was referred to by Judge McKenna as "...the second stage of their analysis where FDEs combine their first stage results and draw inferences as to the genuineness of questioned signatures".

The three explanations that we propose are:

1. The questioned signature was written by the writer of the standard material.
2. The questioned signature was simulated by a writer other than the standard writer such that no evidence of the simulation process remains.
3. We have a chance match between the questioned signature and another person's signature.

We could, if we choose, stop at that point and let the court make a ruling as to which of these explanations is acceptable beyond reasonable doubt or in the balance of probabilities. Of course the court in many instances will have a great advantage over the document examiner, as other evidence can be brought in which may change the plausibility of any one of these explanations. We would argue, however, that the expertize of the document examiner can still be applied at this stage. The expertize required to do this, however, is not based on properties such as the determination of uniqueness or individual characteristics, but rather is derived from a number of fundamental relationships that we propose exist and beg further investigation. What follows is an explanation of these relationships.

8. Complexity theory

Skilled handwriting is thought to be manufactured by a series of concatenated single curvilinear strokes. The function of the motor system, summarized by Thomassen and van Galen (1992), although relevant to the underlying theory, will not be detailed here. What is important is that in skilled writers, underlying kinematic order is observed amongst individuals. In the absence of this order we would be unable to carry out any sort of examination based on theories such as are being proposed. Obviously there is a relationship between characters, the concatenation of strokes and the underlying physiological system. When handwriting examiners draw out features, what they are doing is describing the relationship between the participating strokes in the resultant character which may describe the shape or construction of a complete character, sections within it, or relationships between them. There theoretically is, with a skilled writer, a relationship between the number of these stroke concatenations, the resultant features, and the complexity of a given sample of handwriting. It is the notion of *complexity* that is central to our method, enabling opinions to be expressed regarding authorship.

Complexity of handwriting can theoretically be related either singularly or jointly to a whole range of characteristics resulting from differing orientations of concatenating strokes. Examples of these resultant characeristics may be the total length of the line, the number of points where the line exhibits feathering, the degree that the line is superimposed on itself etc. We propose that there are a number of basic relationships that exist which enable opinions to be expressed about any nexus that may exist between questioned and standard writings, once the decision that the questioned writing is similar to or consistent with the standard writing has been made. These theoretical relationships can be investigated using normal scientific validation protocols. These relationships are described in figures 2 to 4. For clarity, the general logic underlying these relationships will be described.

8.1 The number of concatenated strokes versus the complexity

The first relationship is the number and relative orientation of concatenating strokes, or a measure of this parameter such as the number of curvature maxima, as a predictor of complexity. That is, in the skilled writer, the greater number of times the pen was required to change direction without a penlift, the more visually complex the image appears.

8.2 The complexity versus the likelihood of a chance match

This relationship follows from that stated above in that, given that all writings share common components such as concatenating strokes, and given that the number of concatenating strokes contribute to the complexity, then if we were to choose random samples exhibiting identical text, as we proceed through an analysis of the concatenations, the complexity increases and so does the likelihood that the samples will diverge in some way from each other.

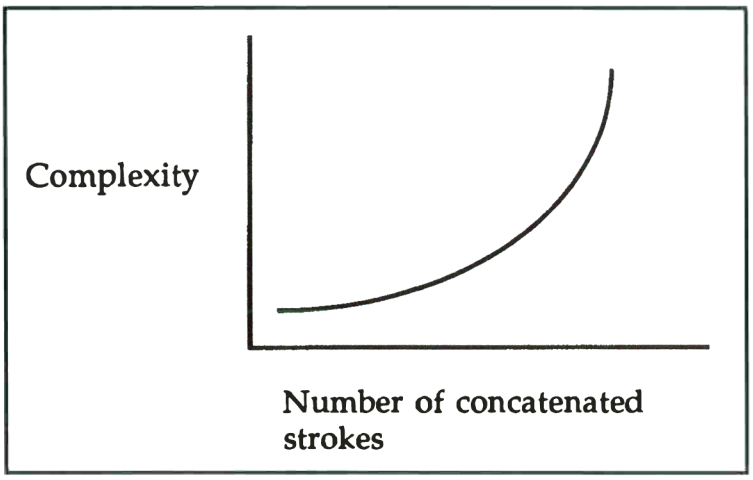


Figure 2. The theoretical relationship proposed between the number of concatenated strokes (or a measure thereof) and a handwritten image's complexity.

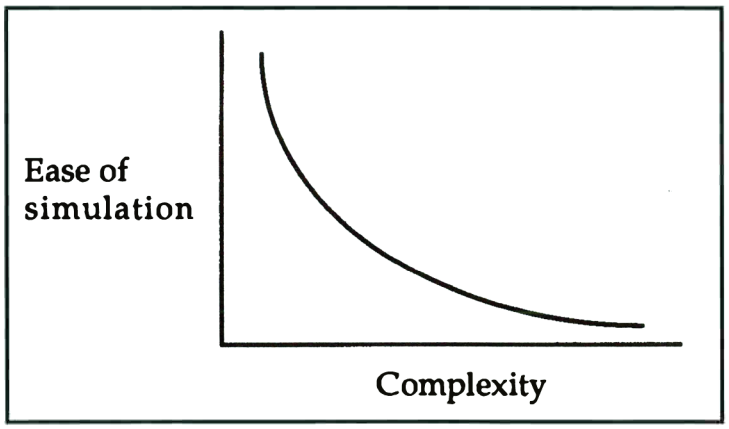


Figure 3. The theoretical relationship proposed between the complexity of a handwritten image and the ease with which it could be simulated successfully.

8.3 The complexity versus the ease of simulation

Given the above, as the image becomes more complex, it would make sense that it would be more difficult to simulate. An example of this would be copying a straight line in comparison to copying an extended section of text.

The issue becomes not one associated with the frequency of feature formations in the population or the subjective assessment thereof, but rather, if we accept that most individuals write differently from one another, the complexity of the static image. The research direction is therefore investigating the questions: What evidence do we have which supports the proposed relationships? What features best predict a written image's complexity? How can we objectively measure complexity predictors? There are a number of ways to investigate this phenomena. One way is to get handwriting experts to group images according to their perceptions of complexity and then to analyze the image according to characteristics that can be counted or measured objectively. This approach has been reported on, but not for the reasons as stated here (Found & Rogers, 1995). Another method is to correlate parameters measured for handwritten formations with a measure of success in actually forging these characteristics. The latter has not as yet been attempted by the authors but is feasible if undertaken in an objective manner.

The complexity theory also enables us to explain the common ground between handwriting text base

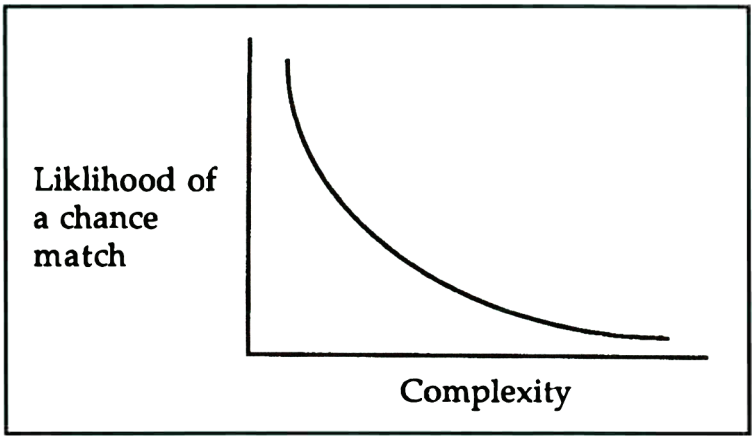


Figure 4. The theoretical relationship proposed between the complexity of a handwritten image and the likelihood of a chance match between the handwritten features of any two individuals

examinations and signature examinations because we have diminished the importance of the *linguistic* cues. The issue of whether we should attempt to examine foreign writings, however, remains questionable because of the difficulty in constructing the general sense of the examination through letter, letter combinations and word cues. This issue is associated with method and will not be further discussed here. Perhaps this problem will only be overcome when the human factor is completely removed from the examination equation.

Complexity, we believe, may also be the key to defining how much text is needed to express a valid opinion. Objective means to quantify the best predictors of complexity may also provide us with better definitions of what constitutes similarities or differences.

Given the underlying order of handwriting production in terms of concatenating strokes, the proposed relationship between complexity, and the likelihood of chance match and ability to be simulated, we can now address the explanations given for the similarity of the questioned and standard signature given in our example in terms of the theory. An example of the rationale for expressing an opinion that there exists a nexus between the questioned and standard writing is as follows:

1. Explanation number 3 is in our opinion implausible on the basis that it is not believable that a signature falling within the range of variation in the standard material has been incorporated into the document by accident, or that an individual attempting to mark the document fraudulently has by chance produced a signature appearing to be genuine, even though that was not the intention. We do not, therefore, support this explanation.

2. Explanation number 2 is in our opinion implausible on the basis that the signature appears to be complex, fluently written and not bearing any indicators of a simulation process. The opinion of the examiner is that the signature exhibits sufficient features that would bestow upon it a measure of difficulty if it were attempted to be simulated. We do not, therefore, support this explanation.

3. Explanation number 1 is, therefore, the only remaining explanation. Given the complexity of the image and the absence of differences, it is considered the most plausible hypothesis. We therefore support this hypothesis.

Therefore, it is the opinion of the examiner that the most plausible explanation accounting for the similarities observed is that the writer of the standard signatures also wrote the questioned signature.

In this way we can still express an opinion as to

the authorship of the questioned entry. However, it is made quite clear that there are alternatives that have to be recognized. These alternatives can never be absolutely excluded due to a combination of factors including the nature of behavioural artifacts, the lack of objective techniques available to analyze them, and the inability to meaningfully support or exclude them statistically. We do not express the results in probabilistic terms, but only on our beliefs according to this method and expertise in applying it. Of course there will be many instances where there is not a clearly most plausible hypothesis. These cases are inconclusive. It may be possible to suggest limited support for one explanation over the others on the grounds that, for example, there is limited standard material such that all of the features could not be accounted for.

The process outlined above is relatively straightforward when dealing with a case where the questioned entry is similar and complex. There is a tendency to believe that these sorts of processes should work equally well in both directions. They simply do not. This illogical belief is reflected in some reporting procedures where opinions are stated to range from a total identification of a given person to a questioned document, to the total elimination of that person as having authored something. This can pose some difficulty in court, particularly when it is commonly touted and believed that forensic handwriting examination is equally as good at excluding individuals from having written a questioned document. Let us proceed through the same hypothetical situation as before. Let us imagine that we have performed an analysis of questioned handwriting. The opinion of the examiner is that there are differences in the line quality, construction and spatial characteristics associated with the questioned handwriting when compared to the population of standard material. Given this situation, there exists only one general explanation that could be advanced to explain the differences: the questioned signature was unlikely to have been produced using the same neuromusculature commands as were used to form the standard writing.

As can be observed, this is not a statement about beliefs as to authorship. We now must look at possibilities that could account for this primary finding. Examples of these explanations are:

1. The standard writer did not write the entries.
2. The population of writings submitted as standard is not representative of the standard writer's normal handwriting and the standard writer was responsible for the entries.
3. The standard writer is capable of producing more than one writing style.
4. The standard writer has purposefully changed his or her writing.
5. The writing of the standard writer has been affected by unknown internal or environmental factors. Examples of these factors are age, illness and intoxication, references to which can be found in Ellen 1989(p.45).

The greatest problem that we have in this situation is providing support for one of these plausible explanations over all of the others. It is very difficult to justify the opinion that the standard writer did not write the entries, as to do this one must be able to illustrate that that writer was incapable of having written the entries. In addition, we must also provide meaningful research that would justify not supporting the alternatives. Although through research (e.g. effects of alcohol on writing) we may be able to state general trends, there are real threats to external validity on applying these results to any specific case example. With a great number of standards and questioned material taken from around the same time, it may be possible to reasonably provide limited support for hypothesis 1, but given the nature of the alternatives, exclusion would be inappropriate.

This should not, of course, be seen to detract from the evidential power of handwriting examinations. The opinion that the standards and questioned entries are different in a major and illustratable way does provide the court with information that may be of use given the other lines of evidence.

The relationship between handwriting analysis and exclusionary opinions becomes more distant as we move down the scale of complexity as illustrated in figures 3 and 4. At the lower end of this scale we

have not only spurious signatures and small amounts of questioned writing, but also larger samples of writing that are not considered to be skilled; that is, writings with poor line quality or writings that are simply or variably constructed. If we look at the same plausible alternatives to explain differences with signature formations, then we are faced with the situation that it is almost impossible to support any one of the alternatives in a meaningful way. The misunderstanding of this concept is seemingly illustrated in an article by Beck (1995) who stated that, "The principle of elimination is as simple as basic scientific method: no matter how much evidence exists for a theory, it must be rejected if even a single significant contradiction is encountered." This discussion then proceeds to support this statement on the basis of other statements by Harrison (1958), Osborn (1929) and Hilton (1982). In this case, we consider the logic is being applied to the wrong section of the methodology. The *theory* in this instance is at the level of whether the questioned material is similar to the standard material. If a 'single significant contradiction' is encountered, then we would agree that the opinion would be that the bodies of writing are different, not that the writers of the bodies of writing are different as is stated.

9. Reporting procedures

Having established the subjective nature of the examination process and the limitations imposed by the underlying theoretical framework, we now must consider how best to express the results of our analysis. The interface between what we do and what the perception is of what we do and mean is conveyed primarily at this stage. Reporting procedure is diverse in the field. However, it appears that in America at least, the probability scale is popularly accepted (McAlexander, 1991). The problem with this scale is that it implies a level of exactness that is not supportable by any studies, nor by the theory underlying it. This was reflected in Judge McKenna's comment that "No showing has been made, however, that FDEs can combine their first stage observations into such accurate conclusions as would justify a nine level scale." In addition, there is predictable confusion between the probability terms used and the mathematics that usually underlie them in traditional scientific paradigms.

Arguably the most flawed aspect of its use on scientific grounds is the top and bottom two levels of opinion, where we have both highly probable and certainty. We would argue that, even given this system, highly probable would be the highest opinion that could reasonably be expressed because of the inability of the examiner to absolutely exclude alternative hypotheses to account for the differences and or similarities observed. Indeed, the use of the word *certainty* in the court room is most inappropriate, particularly when the subjective nature of the analysis may not have been made clear, and where the perception of the study may have been coloured by terms such as, scientific, identification, individual characteristics, experience, etc.

Fortunately, there are alternatives to expressing results according to the scales described above. The first suggestion is to make clear in written reports the limitations associated with the type of evidence that is being presented:

1. The images are the artifacts of human movement and do not in themselves define the process by which they were carried out. Indeed, the image that we examine can at best be considered a sample of the overall movement outcome. Dynamic information, although it can in some ways be inferred, is not available.
2. The written image from the same person can manifest differently, primarily as a result of the underlying neuromuscular system which is responsible for its execution. In addition, environmental factors associated with the writing implement, the writing medium, and body position may alter the artifact.
3. Handwriting, as with any learnt motor behaviour, can be modified (either consciously or subconsciously) or mimicked.
4. Although handwriting features are focused upon when making comparisons, the absolute significance of these features are not able to be determined.

The results section should contain a statement as to similarity or difference, a list of the plausible explanations that could account for this primary opinion, and a discussion as to why alternative explanations were excluded in favour of the one that the examiner is supporting.

10. Theory and forgery

It appears that the dispute over forensic handwriting examination in US v. Starzecpyzel was related to a signature case where the conclusion of a forensic document examiner was that the signatures were "not genuine." Given that this was the starting point of the dispute, it may be appropriate to discuss the examination of static signature formations in light of the theories proposed in this paper. The example used by Judge McKenna will be used to investigate this point. These signatures are drawn from Harrison (1958). Judge McKenna states that the illustration "...shows two signatures with many identifiable differences such as the ornamentation of each "B" and the curvature of the initial stroke of each "M." Given no other exemplars, the lay examiner might correctly conclude that one of the signatures was a forgery. While an FDE might come to the same conclusion, he or she would first have considered the possibility that both signatures were genuine, the differences arising from such sources as natural variation, the passage of time, purposeful alteration (e.g., elaborate signatures used when signing checks), illness, or intoxication. As Ms. Kelly repeatedly stated throughout her testimony, FDEs are aware that forgery detection requires an adequate quantity of genuine writings to eliminate such possibilities.

By way of clarification, let us use the definition of *forgery* or *fraudulent* signature as stated by Hilton (1982):"A forged signature. It involves the writing of a name as a signature by someone other than the person himself, without his permission, often with some degree of imitation." This term, therefore, is not only a statement regarding non-authorship, but also intent. This approach simply does not fit within the model that has been proposed, nor the method that we use. Referring back to Harrison's example, we can state that there are identifiable differences. We can illustrate these differences and if we chose to, could objectively measure them using specific software (Found, Rogers, Metz & Schmittat, 1994); Found, Rogers & Schmittat, 1994; Found, Rogers & Schmittat, 1995). Having expressed the opinion of difference, we can then state that in our opinion the questioned signature was not produced using the same neuromusculature commands as was used to form the standard writing. The plausible explanations to explain this opinion can then be stated as has been discussed previously above. We are left with a set of explanations where we cannot reasonably, nor scientifically, exclude each of the alternatives in favour of only one. Fundamentally, in the example used, if we relate the questioned image to the complexity relationships, the complexity level is low and so the number of individuals that could perform the signature is high. There is no reasonable basis on which to exclude the standard writer as one of these individuals that could have performed this particular signature. This logic is mirrored in the more recent text by Ellen (1989) who states "When significant differences typical of those found when signatures or other writings are copied or discovered in a questioned signature, and are not present in any adequate number of those known to be genuine, it can be safely concluded that the signature is not the normal signature of the suspect. If it also shows a clear overall similarity to the genuine signatures, too close to have arisen by a chance match, it can be reported as a simulation, and that there is no evidence that it was made by the writer of the genuine signature. In such cases it is usually unwise to report that because it is a simulation it was not made by the person whose writing has been simulated."

The support for the notion that *forgeries* can be identified comes from observation of known forgeries. Reports of these are found all through document examination literature. The reality is, however, that because the differences noted in the questioned signature are similar to those noted in known forgeries, it does not mean that we can instantly conclude this is a forgery and exclude the standard writer. Of course there is a body of research that indicates how it is that individuals *forge* their own signature, what happens to signatures in various states of ill health, etc. Let us not mistake this type of research as providing validation of our ability to absolutely exclude alternative plausible explanations to account for observed differences in signatures. We can and do use this research to

answer questions in court regarding general trends. However, we do not use it to exclude the standard writer from having authored the simulation, if that is the conclusion that we come to. We would argue that the most important role for the handwriting specialist in this case is to illustrate to the court that the questioned signature is different and explain what the possibilities are that could account for these differences. If asked, "Are the different features that you observed typical of a forgery process?" we can answer that they are, but that does not mean that the standard writer did not perform the entry and that there are other explanations that could be proposed. The court has the great advantage that they can have other relevant information such that they could, under certain circumstances, support the hypothesis that the signature was forged. This is, of course, the role of the court and not the document examiner.

What has then been discussed here is a limited example. Obviously there is a difference between examining the limited line trace associated with signatures and examining extended amounts of text, although the plausible explanations accounting for similarities or differences remain similar. Overall, it is the subjective nature of the entire process, coupled with the variable nature of writing traces, that impose the limitations on any inferences that can be made regarding the authorship of questioned handwriting. So where is the research headed to validate the model that is being proposed here?

11. Research directions

The decision as to similarity or difference is a primary candidate for research into *objective static analysis techniques* to aid in the decision process. As with the above-mentioned rationale, this research is not focused on issues of authorship, but on providing examiners with objective criteria to supplement the subjective assessment of whether a population of images is similar to or different from another population of images. Signature formations have been the initial subjects of this kind of research (Found, Metz, Rogers, Schmittat, Black & Ganas, 1994.) Signatures are straightforward to investigate in this environment because we are making inferences, mounted on its consistency and complexity, about the plausibility of a single questioned image being the product of the same neuromuscular processes as was used to form the standard images. We can, therefore, construct at least spatial criteria that have to be met in order to proceed to the stage of proposing hypotheses about the explanations as to why an image is similar or different. An example of this kind of approach would be that in order to express the opinion that the questioned image was written by the standard writer, the signature would be required to reach a spatial criteria consistent with the population of standard images and fulfill other criteria such as those associated with complexity and subjective line quality assessments. Although this approach is theoretically and practically achievable, the research is still in its infancy. There are problems, however, in translating research on common images to examinations of extended text. The limiting factors are that we observe a phenomena thought similar to *context specific variation* for speech (discussed in Schmidt, 1988, p.238.) That is, we observe structural variation within and between characters according to their placement within word formations and/or the surrounding characters. This, coupled with a lack of objective analysis techniques that can make the required measurements efficiently, poses a challenge for the application of measurement techniques in this area.

12. Conclusion

As with any opinion expressed on the outcome of human movements there is a fundamental requirement to be familiar with the normal range and variation of movement outcomes in the population from which routine examination material is drawn. For handwriting examiners, this experience comes mainly from the exposure we have to handwrite throughout the course of our life, the majority of which normally would occur before specializing in forensic handwriting examination. Forensic training serves to focus our approach to the comparison process according to the method. It should not be seen to be isolated from the real basis on which our opinions are formed which is a general exposure to the population of writing images, coupled with a knowledge of the limitations of the technique and the relationship between neural representations, artifacts of movement, complexity of images, and what can reasonably be said regarding

authorship of entries based on these elements.

Handwriting examination has traditionally been a study that has developed in relative scientific isolation. The field is small and the emphasis, as we would expect, has been on application, as this is why forensic handwriting examination came about. Research and validation have suffered as a result. It has become clear that as practitioners dealing with the artifacts of human movement, we share a great amount of common ground with scientists working in mainstream paradigms. It is unlikely, however, that forensic handwriting examination will ever be considered as a science similar to these traditional scientific paradigms. The results of the Daubert hearing, given the type of information that they were provided with, appears reasonable almost to the point of generosity. The future for our profession is based on learning from the types of criticisms that have been raised and recognizing that some of the traditional beliefs in the field must be abandoned.

Only a small number of the points raised during the Daubert hearing have been discussed here. It is not suggested that the approach outlined in this paper provides a quick fix to the problems that our field is experiencing. Indeed, what has been presented requires a great deal of work to validate in the terms that were suggested by the scientists giving evidence in the hearing. That the expertize of document examiners is properly characterized as "practical in character" rather than scientific we do not consider to be inaccurate or inappropriate. However, what is important is that in common with scientific practice we present results in a way that reflects the type of information that we deal with, and respects the limitations of the assumptions and techniques we. use to reach those results. Furthermore, the future of the field will ride on the back of scientific research and the criticisms raised can only aid us in attracting suitably qualified individuals and funding to carry out the required work.

13. References

Baier, P.E. (1995). Image processing of forensic documents. Proceedings of the Third International Conference on document Analysis and Recognition (pp .1-4). Montreal, Canada: IEEE Computer Society Press.

Baxendale, D., & Renshaw, I.D. (1979). The large scale searching of handwriting samples. *Journal of the Forensic Science Society*, 19, 245-251.

Beck, J. (1995). Sources of error in forensic handwriting evaluation. *Journal of Forensic Sciences*, 40, 78-82.

Cheung, Y.I., & Leung, S.C. (1989). A comparative approach to examination of Chinese handwriting. Part 4-Identification by statistical classification techniques. *Journal of the Forensic Science Society*, 29, 77-78.

Conway, J.V.P. (1959). *Evidential Documents.* Illinois: Charles C. Thomas.

Eldridge, M.A., Nimmo-Smith, I., Wing, A.M., & Totty, R.N. (1984). The variability of selected features in cursive handwriting: categorical measures. *Journal of the Forensic Science Society*, 24, 179-219.

Ellen, D. (1989). *The Scientific Examination of Documents: Methods and Techniques.* West Sussex: Ellis Horwood Limited.

Found, B., & Dick, D. (1992). The structure of handwriting comparisons. An overview of the limitations within the currentmethodology and a discussion of an alternative method. Paper presented at the meeting of the 11th Australian and NewZealand International Symposium of the Forensic Sciences, Hobart, Tasmania.

Found, B., Dick, D., & Rogers, D. (1994). The structure of forensic handwriting and signature comparisons. Forensic Linguistics. *The International Journal of Speech Language and the Law. 1*, 183-196.

Found, B., Metz, H., Rogers, D., Schmittat, R., Black, D., & Ganas, J. (1994, November). An analysis of Spatial Errors in freehand simulations. Poster presented at the meeting of the 12th Australian and New Zealand International Symposium of the Forensic Sciences, Auckland, New Zealand.

Found, B., Rogers, D., & Schmittat, R(l994). A computer program designed to compare the spatial elements of handwriting. *Forensic Science International*, 68, 195-203.

Found, B., & Rogers, D. (1995). Investigation of signature complexity for forensic purposes. Proceedings of the Annual Symposium of the Association of Forensic Document Examiners and the Seventh Biennial Conference of the International Graphonomics Society (pp. 52-53). Ontario, Canada: Phylmar Associates.

Found, B., Rogers, D., & Schmittat, R(l995). Techniques for overcoming difficulties taking objective spatial measurements from handwriting. Proceedings of the Annual Symposium of the Association of Forensic Document Examiners and the Seventh Biennial Conference of the

International Graphonomics Society (pp. 54-55). Ontario, Canada: Phylmar Associates.

Found, B., Rogers, D., Schmittat, R., & Metz, H. (1994, November).A computer technique for objectively selecting measurement points from handwriting. Paper presented at the meeting of the 12th Australian and New Zealand International Symposium of the Forensic Sciences, Auckland, New Zealand.

Franks, J.E., Davis, T.R, Totty, RN., Hardcastle, RA., & Grove, D.M. (1985). Variability of stroke direction between left and right handed writers. *Journal of the Forensic Science Society*, 25, 353-370.

Hardcastle, RA., & Kemmenoe, D.(1990). A computer-based system for the classification of handwriting on cheques. *Journal of the Forensic Science Society*, 30, 97-103.

Hardcastle, RA., Thornton, D., & Totty, RN. (1986). A computer based system for the classification of handwriting on cheques. *Journal of the Forensic Science Society*, 26, 383-392.

Harrison, W. R (1958). *Suspect Documents, their Scientific Examination*. New York : Praeger.

Harvey, R, & Mitchell, RM. (1973). The Nicola Brazier Murder: The role of handwriting in a large-scale investigation. J*ournal of the Forensic Science Society*, 13, 157-168.

Hecker, M.A. (1995). Gender identification through handwriting. Proceedings of the Annual Symposium of the Association of Forensic Document Examiners and the Seventh Biennial Conference of the International Graphonomics Society (pp. 74- 75). Ontario, Canada : Phylmar Associates.

Hilton, 0. (1982). *Scientific Examination of Questioned Documents*. New York : Elsevier Science Publishing Co., Inc..

Huber, RA., & Headrick, A.M. (1990). Let's do it by numbers. *Forensic Science International*, 46, 209-218.

Lacey, K., & Dick, D. (1992). Class and individual characteristics. Do they have a place in handwriting identification. Poster presented at the meeting of the 11th Australian and New Zealand International Symposium of the Forensic Sciences, Hobart, Tasmania.

Livingstone, 0.B. (1962). Frequency of certain characteristics in handwriting, pen-printing of two hundred people. *Journal of Forensic Sciences 8*, 250-259.

McAlexander, T., (1991). The standardization of handwriting opinion terminology. *Journal of Forensic Sciences*. 36, 311- 319.

Metz, H., Found, B., Dick, D., & Rogers, D., (1994, November). The common forensic handwriting and signature method. Paper presented at the meeting of the Australian Society of Forensic Document Examiners, Wellington, New Zealand.

Muehlberger, R.J., Newman, K.W., Regent, J., & Wichmann, J.G. (1976). A statistical examination of selected handwriting characteristics. *Journal of Forensic Sciences*, 22, 206-215.

Osborn, A. S. (1929). *Questioned Documents* (2nd ed.). Chicago: Nelson-Hall Co..

Proceedings of the Annual Symposium of the Association of Forensic Document Examiners and the Seventh Biennial Conference of the International Graphonomics Society. Ontario, Canada : Phylmar Associates.

Proceedings of the Third International Conference on Document Analysis and Recognition. Montreal, Canada : IEEE Computer Society Press.

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

Schmidt, R.A. (1988). *Motor Control and Learning* (2nd ed.). Illinois : Human Kinetics Books.

Thomassen, A.J.W.M., & van Galen, G.P.(1992). Handwriting as a motor task: Experimentation, modeling, and simulation. In J.J. Summers (Ed.), *Approaches to the Study of Motor Control and Learning* (pp. 113-144). North Holland : Elsevier Science Publishers B.V..

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

van Galen, G.P. (1980). Handwriting and drawing: a two-stage model of complex motor behaviour. In G.E. Stelmach & J. Requin (Eds.), Tutorials in motor behaviour. Amsterdam: North Holland.

Wing, AM., & Nimmo-Smith, I. (1987). The variability of cursive handwriting measures defined along a continuum: Letter specificity. *Journal of the Forensic Science Society*, 27, 297- 306.

Journal of forensic document examination (Online)
ISSN 0895-0849

Matrix analysis: a technique to investigate the spatial properties of handwritten images

Bryan Found [1] , Doug Rogers [2] and Robert Schmittat[3]

Abstract: *Research on objective measurement strategies to assist forensic handwriting experts to make judgements about spatial consistency are providing novel techniques that exhibit considerable potential. We have developed the 'Matrix Analysis' computer program based on the PEAT system philosophy that allows the operator to objectively select measurement points from handwriting, edit these points according to the accepted relationship between curvature maxima and velocity minima, and calculate automatically the measurement range between all the combinations of points selected. This provides the examiner with a semi-automated objective score of the spatial consistency of the questioned image when compared to the range of variation in the standard images. On a typical signature the Matrix Analysis technique compares between 25,000 and 100,000 measurements to generate a spatial consistency score. It is at the stage of determining whether a questioned image is consistent or inconsistent with the range of variation in a standard image group that techniques of this type offer great potential. Examiners of handwriting can then use this information to explore hypotheses from which opinions regarding authorship can be mounted.*

Reference: Bryan Found, Doug Rogers, Robert Schmittat (1998, Vol 11 – reprinted and reformatted). *Matrix Analysis*: A Technique to Investigate the Spatial Properties of Handwritten Images. J. Forensic Document Examination, Vol 29, pp. 23 - 34.

Keywords: Computer based handwriting analysis systems, spatial properties of handwritten images, Matrix Analysis techniques

1. Introduction

A fundamental of forensic handwriting theory relates to the difficulty which individuals experience when attempting to copy the handwriting traces produced by others. The simulator is required to produce what are often a complex series of movements with the aim of generating an image that captures the appropriate combination of space, construction and line quality characteristics. The approach to the investigation of this phenomena in the forensic environment has centered around literature and personal experiences of attempts to simulate writings (Osborn, 1929; Harrison, 1958; Conway, 1959; Hilton, 1982; Ellen, 1989). Much of this information and research has necessarily relied on purely subjective comparison processes. The almost total absence of objective measurement approaches in forensics can be attributed to a number of factors, the most significant being the difficulty in taking and comparing measurements from images that are directional, non-linear and where one portion of the line may intersect and overlap with previously formed sections. An example of such a signature is shown in Figure 5. In addition, the objective assessment of line quality, a measure of fluency or dysfluency of movement, is difficult to achieve on static images. Compounding the philosophy of the importance of objectivity in comparisons is that the data generated does not necessarily provide the examiner with information that may be relevant to issues of authorship. Found and Rogers (1998) have argued that objective tests of

1. Handwriting Analysis and Research Laboratory, School of Human Biosciences, La Trobe University, Bundoora, Victoria, 3083, Australia.
2. Document Examination Team, Victoria Forensic
3. Journal of Forensic Document Examination. Vol. 11, Fall, 1998, pp. 51-71·

the type discussed in this paper are important in the first phase of opinion formation. The primary opinion at this stage concerns whether or not the examiner believes that a questioned image is consistent with the known image in terms of line quality, construction and features associated with space. Computer techniques clearly can provide detailed information regarding spatial consistency which would be difficult to extract using visual processes. According to the traditional texts, observational approaches have been shown to be both efficient and effective. The difficulty with these types of approaches, however, is that visual information may be treated differently, depending on the observer. In an expert system the opinions of one expert may be contrary to the opinions of another, and this has clear implications for the social justice system. Without objective techniques or indices, it may be extremely difficult to isolate the basis of the difference in opinion.

Computer-based handwriting analysis systems are being reported from a variety of fields including handwriting recognition, signature verification, signature identification, database searching, forensic comparison and administrative areas. Signature identification systems aim to identify a questioned signature from a database of known signatures. Han and Sethi (1996) described a signature identification system that utilised geometric (horizontal and vertical bars, loops etc.) and topological features (end points, branch points etc.). The extracted comparison features were then normalized to control for translation, rotation and scaling. A number of search and match strategies were then applied to attempt to optimize the identification of the questioned signature from the reference set. Murshed, Bortolozzi & Sabourin (1996) described an off-line signature verification system which, it was argued, compared images in a similar way to forensic experts. The technique involved preparing the image to remove background information, dividing the image into a number of smaller regions and comparing features within each of the regions to regions within images in the database. A decision is then made regarding the identity of the signature based on a training technique with genuine signatures. Although there is great potential to apply such techniques in the forensic environment, there are some limitations that must be noted. Forensics necessarily deal exclusively with static signatures so verification based on dynamic data, even though some of this data can be inferred (Found, Rogers & Schmittat, 1997), cannot be utilized. The validity of applying algorithms that normalize or distort the image may be a cause for concern in the court environment where such changes made to images extracted directly from items of evidence is subject to criticism. Off-line systems still do not appear to be extracting line fluency information.

Research on objective measurement strategies which may assist forensic handwriting experts to make judgements about spatial consistency are providing novel techniques that exhibit considerable potential. Phillipp (1996) provides a summary of some of the systems which are more relevant to forensic handwriting examination with a view to assessing their applicability as a supplement to existing subjective comparison approaches. These systems have shown to be capable of detecting 100% of random and simple forgeries and over 90% of skilled forgeries (Ammar, 1995). The 'Forensic Information System for Handwriting (FISH)' is a well known system within the forensic sciences and uses a combination of manually entered descriptive features, automatically calculated non-textual features, textual features and features which are measured with the assistance of the operator, to define a sample of questioned handwriting and compare it to a large number of both known and unknown writing samples in a database (Hecker, 1996; Philipp, M, 1992). Sagar and Leedam (1996) noted the limited research effort in the field of forensic document examination with regard to automating the comparison process. These authors describe a number of collaborative projects dealing with computer-aided examinations. The Forensic Document Examination System (FODES) is a software package that enables the examiner to extract characters from digital images and generate charts using these characters and overlay images. It has been reported that this technique provides a significant time saving in the routine construction of display charts (Holcombe, Leedham & Sagar, 1996). The Writer Identification System (WIS) combines information extracted from the context of the document content in combination with global features (character slant, size, ascender and descender heights etc.), interactive determined

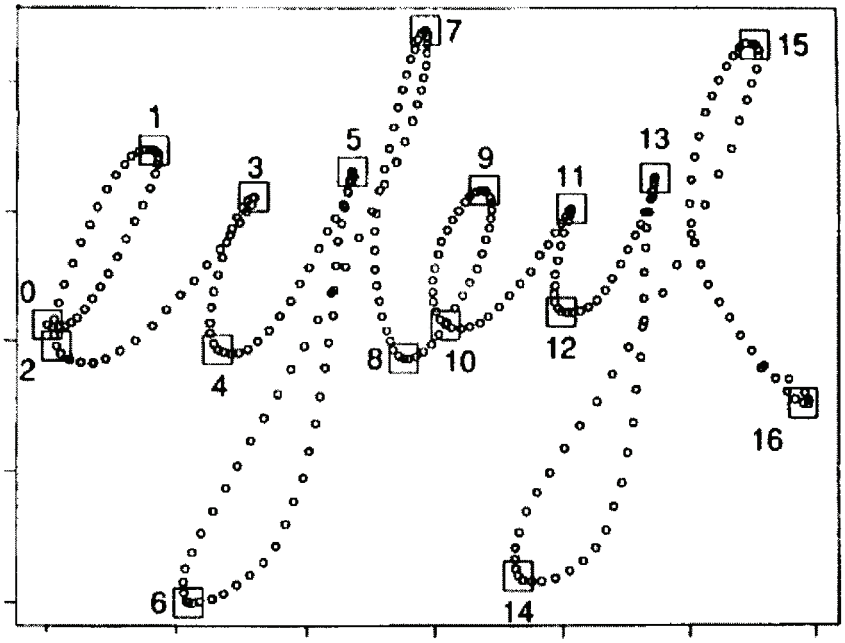


FIGURE I. A sample of handwriting with marks indicating stroke based segmentation. Each dot in the trajectory represents a sample equally based in time. The tangential velocity of the pen along the curve is proportional to the radius of the curve at any point. (Wright, 1993, *Acta Psychologica*, 82, 5-52.)

local features (character classifications according to temporal construction) and texture features which can be generally thought of as characteristics of the image independent of their linguistic meaning. This information is planned for use in the comparison process but is reported to still be in its research phase. The *Forensic Document Examination Tools* software is reported to be a similar system to WIS but is designed to extract global and local features automatically. This system is also reported to still be in its development phase.

The Pattern Evidence Analysis Toolbox (PEAT) (Found, Rogers & Schmittat, 1994) and the related Angular Differential software (Found, Rogers & Schmittat, 1997) were developed along a similar philosophy to those systems described above. The approach adopted for our systems was based on overcoming the traditional measurement difficulties by designing specific tools to follow the path of the line (Smartline), interact with the operator (PIG Grid) and rationalize spatial measurement points according to motor control theory and the observation that maxima in line curvature correspond to velocity minima. Evidence of this is provided in Figure 1. The systems developed by the authors are specific to the comparison of like images; that is, comparing a questioned signature to a group of known signatures with a similar identity to derive a measure of the consistency of the questioned image in terms of space. The approach rationalizes the comparison using the relationship between the dynamics of signature production and extractable measurement points from the static image. Once the temporal sequence of the measurement points is determined, the spatial characteristics of the signature image can be thought of as a series of temporally ordered points in space. These points are therefore analyzed and are thought to represent all of the characteristics that document examiners subjectively assess in terms of spatial features (eg. ascender and descender heights, internal proportions, etc).

An obvious criticism that can be levelled at objective measurement techniques is the choice of what is to be measured on a particular image in combination with questions as to what is to be done with the measured data. Techniques such as the PEAT and other measurement strategies can, to different extents, be susceptible to criticisms on these grounds. The *matrix analysis* technique described here overcomes criticisms of this type as the operator is provided with a method of objectively picking the measurement points, using the angular

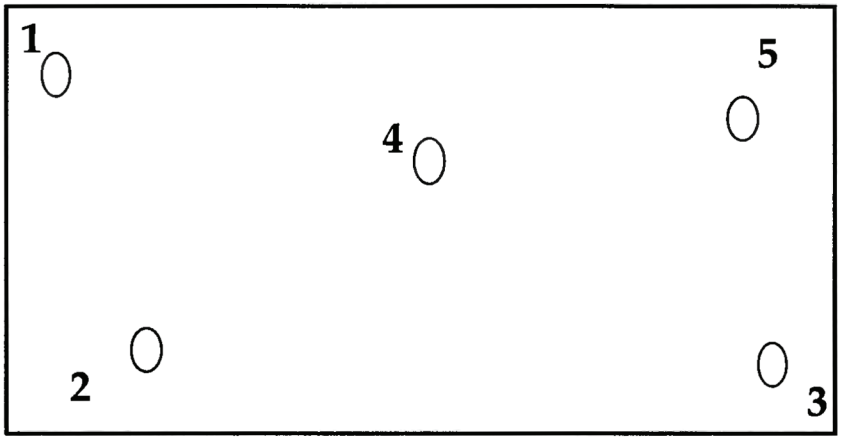


FIGURE 2. Points of velocity minima, corresponding to curvature maxima, isolated from a static signature. It is the relative position of each of these points that is determined using the matrix analysis program.

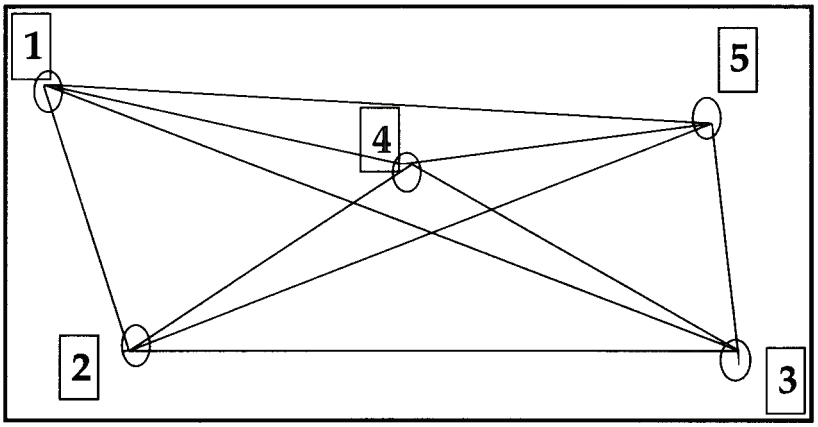


FIGURE 3. Raw distance measures calculated by the matrix analysis program eg. measurement in millimeters of 1-2, 1-3, 1-4, 1-5, 2-3, 2-4 etc. From this data the ratio of distance measurements are determined eg. 1-2/1-3, 1-2/1-4 etc.

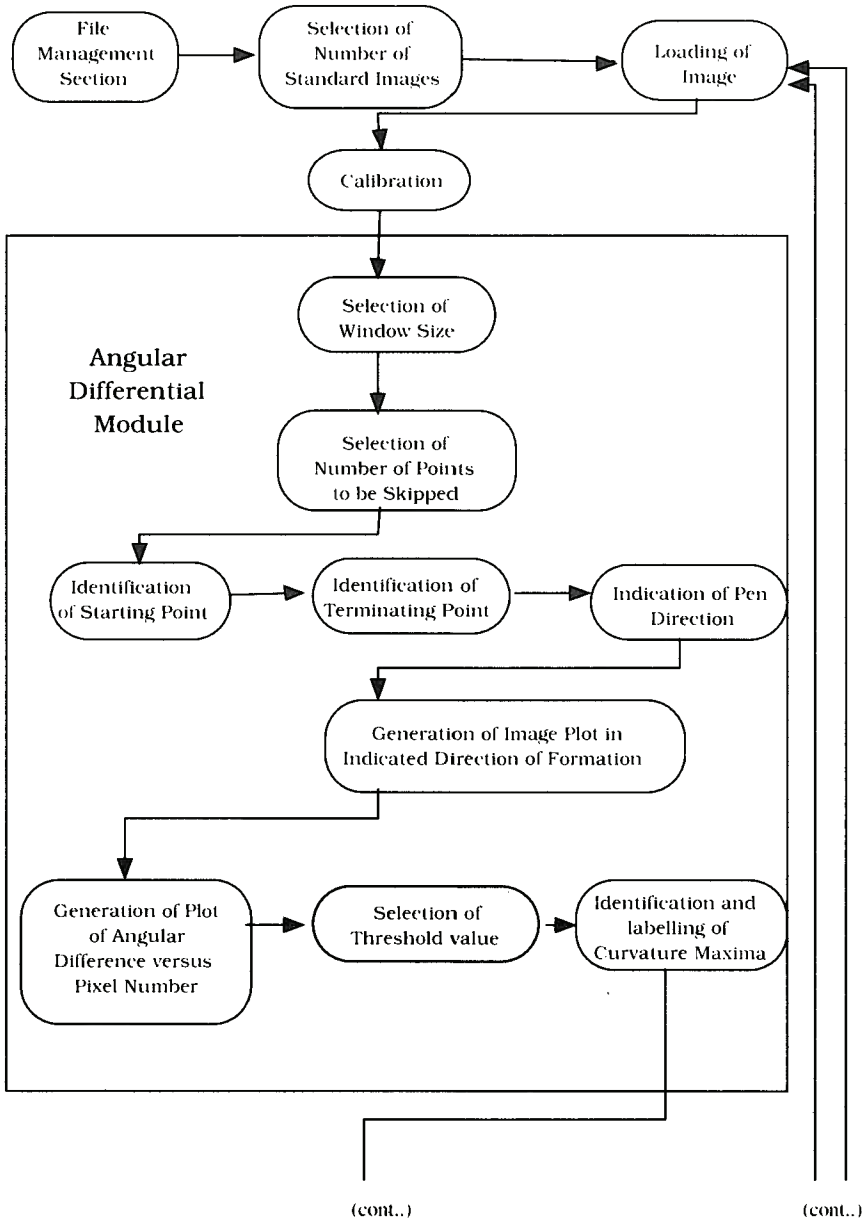
differential module, and measuring the distances and the relationship between the distance between all the measurement points identified. In this way a total spatial consistency score can be measured for each of the questioned images and compared to the standard image group.

2. Equipment

This technique requires a Macintosh series II computer or above, a scanner, the *Matrix Analysis* software, an image processing package (NIH Image 1.41-1.57) and a spreadsheet package (ClarisWorks).

3. Method

The basic technique is to scan both the questioned and standard signatures into the computer. These images are then reduced to a line thickness of one pixel using a skeletonisation technique such as that provided within the NIH image software. Images are then sequentially opened into the *Matrix Analysis* software where the points of maxima curvature are identified with the assistance of the Angular Differential software. Once the measurement points are identified and their temporal order entered, the software calculates all combinations of distance



measurements between the points and compares the range of variation in these measures in the standard material with the corresponding measurements in the questioned material. A spatial consistency score is calculated and provided to the examiner in spreadsheet form. In simplified format Figure 2 represents an array of temporally ordered velocity minima points associated with calculated curvature maxima after the line trace itself has been removed. Figures 3 and 4 show the matrix measurement strategy employed to generate a final spatial score. It is the values of each of these measurements calculated from the questioned signature that is compared to the range of variation in the values of the same measures in the standard group that is used to generate a spatial consistency score. The details of each of the techniques are given below and are represented in summary form in Figure 4.

4. Image Preparation

Since handwritten images are relatively small it may be necessary to enlarge them before the scanning process. This can be achieved using an enlarging photocopier. Images requiring analysis are enlarged

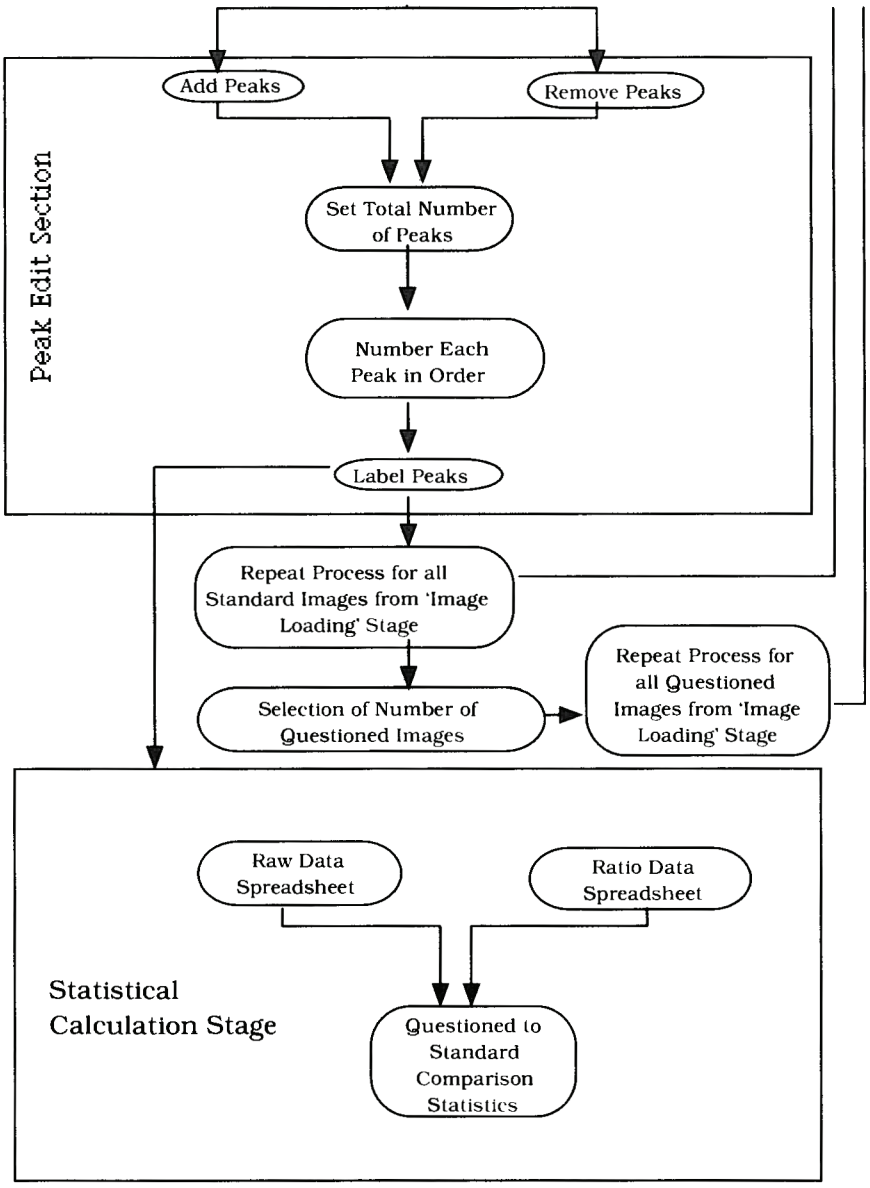


FIGURE 4. An overview of the Matrix Analysis technique for objectively comparing the spatial consistency of a questioned image in comparison to the range of variation in a standard im

to approximately fit across an A4 sheet of paper. A calibration grid accompanies each image through the enlargement process.

5. Scanning

The enlarged images are scanned into the computer and saved as a PICT file. Once all images have been scanned they are processed using NIH Image software.

6. Image Processing

A routine such as density slicing is carried out on the image to set the upper and lower grey scale limits that will res.ult in the image appearing as a complete and continuous line. Under normal circumstances a simple threshold routine will accomplish this. Images are converted to a binary form by setting the image pixels to black and all other pixels to white. A skeletonisation routine is applied which reduces the lines in the image to a thickness of one pixel. The processed images are saved in a MacPaint format (72 dpi).

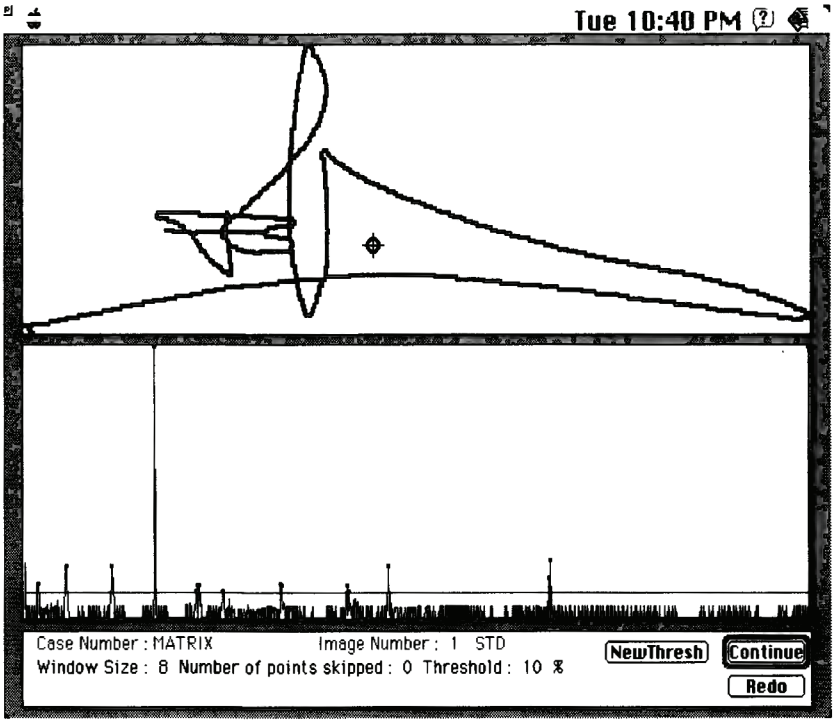


FIGURE 5. The Angular Differential results screen.

7. The *Matrix Analysis* Technique

The *Matrix Analysis* Technique is a stand alone module that utilizes the PEAT file management, calibration and Angular Differential software (Found, Rogers & Schmittat, 1994; Found, Rogers & Schmittat, 1997). Once the image has been opened into the Matrix package, the operator is prompted to enter information regarding the location of the starting point, the terminating point and pen direction for each line segment in the image. From this information the Angular Differential result screen is generated.

Figure 5 displays a result screen of the Angular Differential module (Found, Rogers, Schmittat & Metz, 1995; Found, Rogers & Schmittat, 1997). This technique identifies curvature maxima in the line. The screen is divided into three windows. The uppermost screen is a reproduction of the image being analyzed. The cross-hairs represent the average x and y values of the signature image pixels, where intersections and retraced portions of the line have been appropriately measured. The middle window is a plot of angular difference versus pixel number. The horizontal line represents the threshold value entered by the operator. The peaks above this line are colored. The corresponding pixel in the image window is also colored for identification. It is these colored pixels that are used for *matrix analysis*. The bottom window is the information and instruction window. This window shows the values for. all variables entered by the operator. The maxima present themselves either as a single blue pixel, a black pixel between two blue pixels or a blue pixel between two other blue pixels. It can also be observed that the start and endpoints of the image bear no curvature maxima. The start and end points are, however, candidates for measurement. These points can be added in the editing screen of the software.

Figure 6 displays the editing screen of the *matrix analysis* module. Blue pixels can either be added to the image (for example at the start and end-point) or removed from the image. The operator is required to locate the pixel of interest using the cursor and the PIG routine as has been described (Found, Rogers & Schmittat, 1994). At the end of the editing process the image is in a form where each measurement point is represented by a single blue pixel.

The total number of peaks present in the signature showing all of the required measurement points are

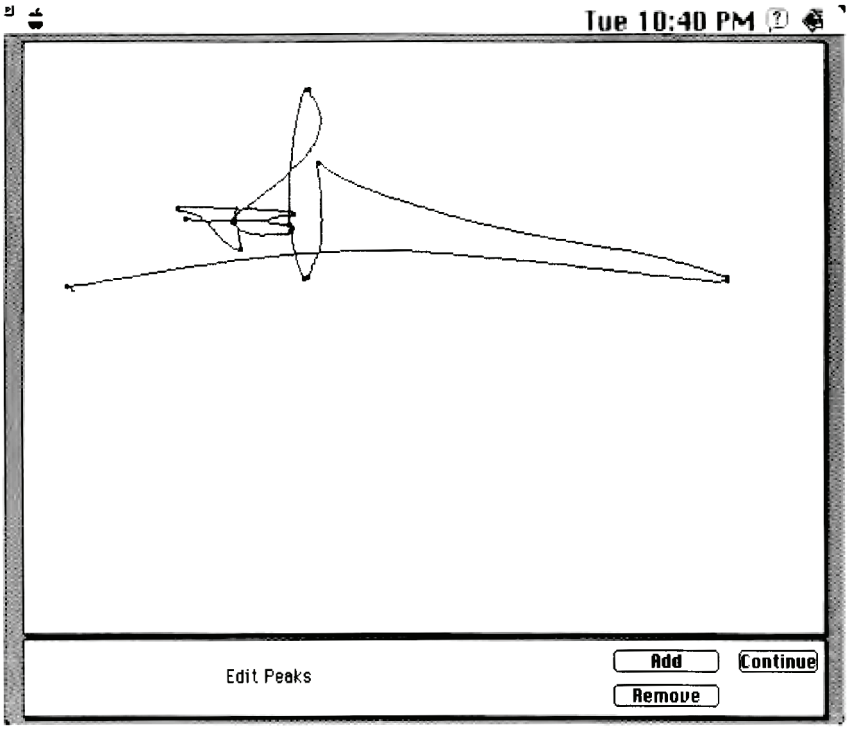


FIGURE 6. The peak editing screen. The operator can interact with the software to add or subtract measurement points on the image.

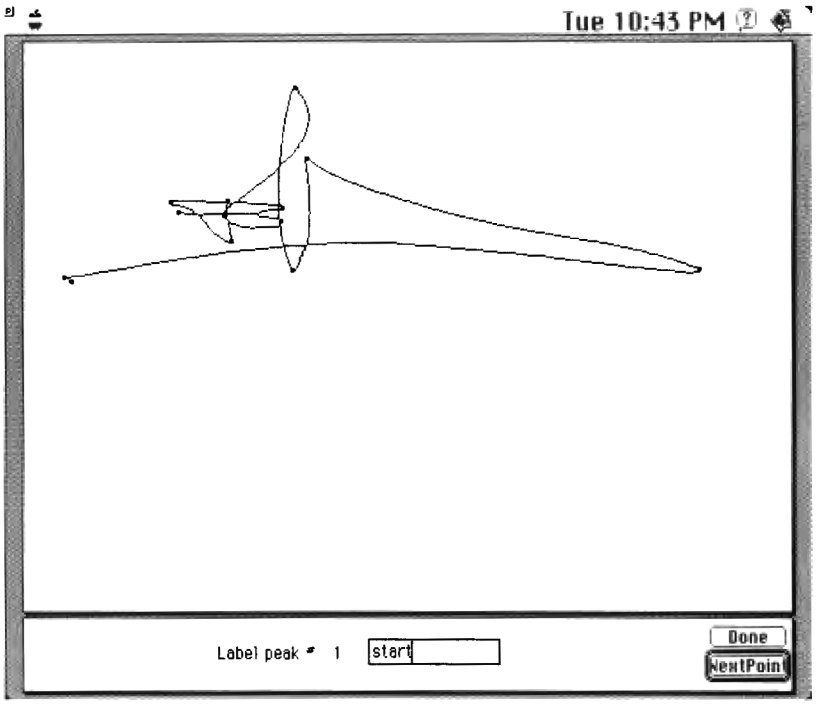


FIGURE 7. The peak labelling screen. The operator can interact with the software to label the measurement points used in the analysis. These labels are used to identify data in the results spreadsheet.

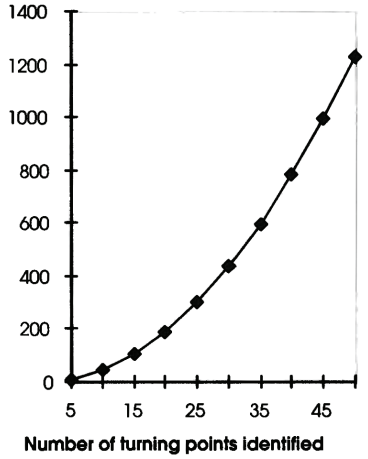


FIGURE 8. A plot of the number of.turning points identified versus the number of raw measurements generated per signature.

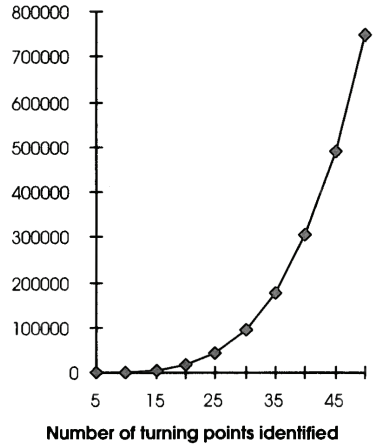


FIGURE 9. A plot of the number of turning points identified versus the number of ratio comparison measurements generated per signature.

now entered into the computer. This value sets the number of rows in the raw results spreadsheet.

The operator is now required to number each of the peaks in order of temporal production. Peaks not numbered are automatically deleted from the calculation. Once this process is completed the operator can enter descriptive labels for each of the peak numbers to assist in the identification of the points measured from the results spreadsheets (Figure 7).

The program now stores the coordinates for the turning point pixels associated with the image analysed, along with the calibration data. The next signature can then be analysed in the same manner (excluding the need to enter labels as the relationship between peak number and descriptive label is now set). At the end of the standard image group the number of questioned images is entered. The same process is conducted on the questioned signature group. When all of the questioned signatures have been analysed, the operator is given the facility to calculate the comparison results in two ways: raw results and ratio results.

8. Raw Results

Raw results are a calculation of the mean, minimum value, maximum value and standard deviation of the data for each of the distances between each of the measurement points for the standard image group. Each of the values for the corresponding data point for each of the questioned images is then calculated and compared to the minimum and maximum value of this point in the standard group. Results of this comparison are flagged as either inside or outside the range of variation on the spreadsheet using a tick or cross symbol. A total of the number of questioned values falling in the range or outside the range of the standard image group is presented at the base of the results spreadsheet. The percentages calculated from these values is the % spatial consistency score. Participant data points contributing to this final score can be derived by visual inspection of the results in the spreadsheet in combination with the labels. Figure 8 provides a plot of the number of turning points identified versus the number of raw measurements generated per signature.

9. Ratio Results

The ratio results subroutine recalculates from the raw data files all combinations of ratios between data points. This provides a total spatial consistency score in the same way as has been discussed in the previous section. Figure 9 provides a plot of the number of turning points identified versus the number of ratio comparison measurements generated per signature.

10. Discussion

The controversy surrounding the relationship between Document Examination and science has been reported (Risinger, Denbeaux & Saks, 1989; Huber & Headrick, 1990). This controversy has culminated in a recent pivotal court decision where it was concluded that forensic handwriting examination could not properly be characterized as scientific in nature (United States v. Starzecpyzel, 1995). Although a variety of factors contributed to this decision, methodological shortfalls in both research and casework investigations arising out of the almost total absence of objective comparison techniques can be seen as a major issue for the field. The most relevant work which may be applied to this problem can be found in the signature identification, signature verification and optical character recognition literature. To date, however, there has been no report of a technique that has been directly carried across and applied routinely in the field of forensic document examination. This is not so surprising given that forensic examiners deal with static images and comparison samples which do not necessarily capture the normal range of behaviour of a particular writer. Compounding this problem is the reality that the analysis of space is only one factor in the overall decision making process regarding the authorship of the image. Spatial information is important, however, when examiners make determinations about whether or not a particular image is consistent or inconsistent with a body of standard material. This is carried out not only on the basis of space, but also on line quality. Any opinion settled upon at this stage is not about authorship but rather about formulating an appropriate set of hypotheses such that issues of authorship based on theoretical considerations of image complexity can be investigated (Found & Rogers, 1995; Found & Rogers, 1998). Given this philosophical approach, our software research to date has focused purely on the issue of spatial scoring and not on the issue of predicting writer identity directly from this score. In view of the most recent submission (Found & Rogers, 1998) regarding underlying theoretical considerations associated with forensic handwriting examinations, this would seem the most appropriate starting point for the eventual inclusion of techniques such as those described here.

The philosophy behind the PEAT program was basically to introduce into the field of handwriting comparison research and casework a technique that generates objective spatial comparison data from common static handwriting traces. Early versions of the program provided the tools necessary to take a variety of length, area and angle measurements which offered solutions to the measurement problems associated with non-linear and intersecting handwriting traces. Although these techniques did provide useful data when comparing genuine to simulated questioned images, the overall approach fell short of the ultimate aims of spatial analysis in casework. The primary concerns proved to be in the areas of operator analysis time and ultimate objectivity in measurement point selection and data generation.

The number of features that could be measured from a two dimensional image to generate a spatial consistency score are very large should curved line length, distance between two points, areas and angles be taken into consideration, both independently and in combination. A whole variety of these measurements could be taken, in each of the modules, by the operator. Of critical importance, given the spatial measurement strategy, was what should be measured and, given the time intensive nature of the task, what the operator neglected to measure. Given this shortfall, similar criticisms could be made of the technique, with regards to its subjectivity, as can be levelled at existing visual comparison methodologies.

The angular differential software provides a method to at least in part compensate for this potential source of criticism. This module could be used in two ways: either to directly identify turning points from which the operator could manually take distance between two point measurements using the appropriate PEAT module, or to validate measurement points identified by the operator's eye. This module, however, did not provide a fast method to edit the points, or to actually measure the distances between the points. In addition, even though the act of measuring between the points was simple, within the distance between two points module alone, if the number of points that were chosen to be measured was large, the process would become extremely time-consuming and open up opportunity for operator errors. Given an image exhibiting 10 turning points

(which for a typical signature is low), the operator would in each instance need to identify, using the PIG grid, the two relevant data points for each of the ten measures. Up to the point where the *matrix analysis* technique was developed the emphasis remained, to a certain extent, on the ability of the operator to choose what actual measurements would be made. Although this process can be criticised, it is still an improvement on the traditional technique of visual inspection and estimation that have been, and are, used.

The most important feature of the *Matrix Analysis* technique is that the spatial score that is produced combines every possible combination of straight line measurements from the data points. There can, therefore, be no criticism that relevant measurements have been excluded. In addition, the technique is very fast. Once the measurement points have been selected and numbered, the actual calculation of the measurements is performed automatically.

11. Future Directions

The future for techniques of this type can be thought of in terms of speed and the relationship between spatial analysis, line quality analysis and issues regarding authorship. Analysis time, although not a fundamental concern in research of this type, may ultimately impact on whether techniques such as those described here will be introduced into routine casework. There is clearly a significant time difference between looking at an image and making a spatial consistency judgement, and analysing the image objectively and making a judgement. Time savings could be made at a number of levels in the analysis process. The following suggestions are but a few.

12. Scanning

The scanning and photocopying enlargement process can very easily be replaced using a CCD camera linked directly to the computer. Images appearing on documents can then be directly stored in a digital form along with the calibration grid. These images can be scaled simply by altering the zoom on the camera.

13. *Curvature Maxima Selection*

Studies could be undertaken to determine whether handwriting experts are able to accurately and repeatedly pick the points of maximum curvature by eye. This study would involve determining the relationship between the curvature maxima identified by the operator within the *matrix analysis* peak edit section, with those curvature maxima selected by the angular differential program. A correlation could be calculated to illustrate the strength of the relationship between these variables.

14. Manual Systems

Given the success of the above proposed experiment, the need to actually enter the entire image into the computer and process that image could theoretically be avoided. A digitising pad and associated pen could be connected to the *matrix analysis* program. The pad would be calibrated. The document or a copy bearing the image of interest could be placed on the pad. The operator would identify the curvature maxima by placing the pen on the turning point and providing a signal through the attached switch. The program would perform the normal analysis functions on the array of turning points provided from each image. We estimate that this would decrease the analysis time by up to 80%. We stress, however, that it moves back from the entirely objective approach as has been described. Ultimately, compromise of this type may be the only way to elicit change in the short term.

Ultimately we are moving towards systems which would employ technology such as neural networks to predict whether questioned signatures are genuine or simulated. Data generated from software such as the *matrix analysis* technique, from complexity models such as that reported by Found and Rogers (1995), and from either a validated subjective line quality scoring technique, or a yet to be reported objective method, would be put into such a network, along with the identity of the questioned signature. Variations in the amount and date range of questioned and standard material could also be introduced. Such a technique could then be subjected to validation trials and the error rate calculated. This type of objectivity in forensic science forms the future goal for

research of this type and offers considerable promise to the field of forensic handwriting examination. The timeliness of the availability of such systems is almost exclusively dependent on the enthusiasm of researchers in the forensic, signature verification, signature identification, optical character recognition and behavioural science fields, in combination with the participation of financial supporters to fund the required research.

15. References

Ammar, M. (1995). Portable software for signature verification and analysis (SIGV A 1.0), useable with IBM-PC compatible machines. *Proceedings of the 7th Conference of the International Graphonomics Society*, London, Canada, August 6-10.

Conway, J.V.P. (l 959). *Evidential Documents*. Illinios: Charles C. Thomas.

Ellen, D. (1989). *The Scientific Examination of Documents: Methods and Techniques.* West Sussex: Ellis Horwood Limited.

Found, B., Rogers, D., & Schmittat, R. (1994). A computer program designed to compare the spatial elements of handwriting. F*orensic Science International*, 68, 195-203.

Found, B., & Rogers, D. (1995). Contemporary issues in forensic handwriting examination. A discussion of key issues in the wake of the Starzecpyzel decision. *Journal of Forensic Document Examination*, 8, 1-31.

Found, B. & Rogers, D. (1998). A consideration of the theoretical basis of forensic handwriting examination: The application of "Complexity Theory" to understanding the basis of handwriting identification. *International Journal of Forensic Document Examiners*, 4, 109 - 118.

Han, K. & Sethi, I.K. (1995). *Proceedings of the Third International Conference on Document Analysis and Recognition.* Montreal,· Canada, August 14-16.

Harrison, W.R. (1958). *Suspect Documents, their Scientific Examination*. New York : Praeger.

Hecker, M. (1995). *Fish and Chips. A video presented at the Proceedings of the 7th Conference of the International Graphonomics Society.* London, Canada, August 6-10.

Hilton, 0. (1982). *Scientific Examination of Questioned Documents*. New York: Elsevier Science Publishing Co., Inc.

Huber, R.A., & Headrick, AM. (1990). Let's do it by numbers. *Forensic Science International*, 46, 209-218.

Murshed, N.A., Bortolozzi, F. & Sabourin, R. (1995). *Proceedings of the Third International Conference on Document Analysis and Recognition. Montreal*, Canada, August 14-16.

Osborn, A S. (1929). *Questioned Documents* (2nd ed.). Chicago : Nelson-Hall Co.

Philipp, M. (1992). Efficiency control study of the FISH system under real world conditions. *Proceedings of the 3rd European Conference for Police and Government Handwriting Experts*. Rome, Italy, October 5-7.

Philipp, M. (1996) On the use of signature verification systems in handwriting identification service. *Proceedings of the Fifth European Conference for Police and Government Handwriting Experts.* The Hague, The Netherlands, November 13-15.

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

Sagar, V.K. & Leedham, C.G. (1996). Forensic handwriting comparison using computer tools. *Proceedings of the Fifth European Conference for Police and Government Handwriting Experts.* The Hague, The Netherlands, November 13-15.

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

Statistical modelling of experts' perceptions of the ease of signature simulation

Bryan Found[1] Doug Rogers[2], Virginia Rowe[3] and David Dick[3]

Abstract: *The perceived complexity of handwriting traces by forensic experts is a critical element in the process by which opinions regarding the authorship of handwriting are formed. Variations in experts' perceptions of how complex an image is can significantly impact on the appropriate administration of social justice . There currently exists no test which is available in forensic science which provides a guide for the expert. This study used discriminate function analysis to construct a model which can be used for such a test. The model is based on 13 government forensic experts' perceptions of how easy or difficult it would be to successfully simulate each of 300 signatures. The variables used by the model to classify these signatures into three complexity groupings were 'number of turning points' and 'number of intersections and retraces'. The test was validated by comparing the model's calculation of complexity grouping versus fourteen forensic experts' groupings of an additional 197 signatures. Although substantial variation was found between the experts' perceptions overall, up to 72.9 % of their perceptions of complexity could be predicted by the model. Misclassification rates were found to be highest when discriminating between signatures where a qualified opinion in the direction of identification would be expressed versus those where a full opinion would be expressed. There was no misclassification associated with signatures where a full opinion would be expressed versus those for which no opinion would be expressed. This test can now be trialed in routine forensic casework and should provide forensic experts with a guide to signature complexity. Research should now be focused on validating the expert perceptions outlined in this paper*

Reference: Bryan Found, Doug Rogers, Virgina Rowe and David Dick (1992 Vol. 11 - reprinted and reformatted) Statistical Modelling of Experts' Perception of the Ease of Signature Simulation. J. Forensic Document Examination, Vol.29 pp. 35 - 51.

Keywords: Signatures, simulation, assessment of complexity, FHEs' perception of the ease of simulating signatures

1. Introduction

Nearly all routine forensic examinations of signature formations are carried out in order to investigate whether there is any likelihood of a nexus, by writer, between questioned material and a body of standard material. The presentation of the evidence, should there be any offered, is based on what is largely a subjective decision by the forensic handwriting expert and is documented in the form of an opinion. In many laboratories quality assurance systems are in place and the opinion reached by an examiner is reviewed by a peer. This process does not, of course, imply that the quality of the result is enhanced, but rather is designed to detect perceived shortfalls in the logic and the process of application of theory to a particular case.

It is the nature of the subjective approach to forensic handwriting examination that has interested

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, Macleod, Victoria, 3085.
3. Document Examination Section Australian Federal Police, LaTrobe Street, Melbourne, Victoria, 3000, Australia.

the authors for some time (Found, Rogers & Schmittat, 1994; Found, Rogers, Schmittat & Metz, 1994; Found, Rogers & Schmittat, 1997). Of particular interest is the relationship between existing theory and numerical assessments of the perceptions of handwriting experts regarding how easy or difficult images are to simulate (Found & Rogers, 1996). Current models of forensic handwriting theory suggest that the experts make a number of judgments prior to expressing a final opinion regarding authorship. It is thought that experts make a comparison of spatial features associated with the line trace and from this visual information reach a decision regarding whether they believe that the questioned image is consistent or inconsistent with the feature range of variation in the body of standard material. The opinion at this stage is not one regarding the authorship of the image. At this stage the method is purely focused on the proposition of the appropriate set of plausible explanations that could account for the observations. Once the appropriate explanations have been proposed, then the examination focuses on issues of authorship and relies on different theory (Found & Rogers, 1998). Should the decision be that the questioned image is consistent, then a number of explanations are proposed that could account for this. One explanation could be that a chance match has occurred whereby the questioned writings just happen to be consistent with the standard writings although they were in reality written by different persons. A second explanation could be that even though the questioned and standard images may be deemed consistent, this may be associated with a person simulating the handwriting characteristics of the standard writer without leaving indicators of this process. The third explanation, excluding the possibility of mechanical writing simulators (Schneider-Pieters, ten Camp & Hardy, 1996, is that the writer of the standard material actually wrote the questioned material. Methodologically, the focus is now on the basis of support for one of these explanations by excluding the remaining as being implausible. It is the complexity of the image that is crucial to a decision at this stage. The ease or difficulty of a person simulating the feature characteristics of another is referenced by this factor. In the simplest case, a single horizontal or vertical line drawn on a page could constitute the entire signature of an individual. This line may satisfy both spatial and feature criteria of the comparison protocol and be consistent with the known material. To express an opinion as to its authorship would clearly be invalid, however, as the image could not be considered complex and could therefore be too easily simulated successfully. Judgments of this type are routinely made by handwriting examiners, however, in the absence of complexity tests or indices.

A pilot study in this area (Found & Rogers, 1996) indicated that a classification model could be developed based on three experts' assessments of signature complexity. This model was found to classify 73.5% of signatures in common with the experts, based on a number of predictor variables such as number of turning points, feathering points, line intersections and retraces. In addition, a small validation set was used which suggested the agreement rate between the model's classification prediction and the expert could be as high as 92%. On the basis of these results, a larger study was designed, funded by the National Institute of Forensic Science (Australia).

The assessment of the complexity of handwritten images has been reported on previously in related fields of research. Kao, Shek and Lee (1983) reported a study of the effects on writing time and writing pressure when tracing or free-hand writing images of differing complexities. Wing (1978) and van Galen (1984) presented the results of reaction time studies on handwriting tasks of differing complexity. Meulenbroek and van Galen (1990) investigated the motoric complexity of cursive letter writing by children by analysing writing velocity, dysfluency and curvature measurements of grapheme segments. Changes in latency, movement time, trajectory length and pen pressure were analysed by van der Plaats and van Galen (1990) with respect to writing complexity. Other research in the forensic environment provide evidence that simulators are more likely to concentrate on eye-catching characteristics and therefore less likely to successfully imitate inconspicuous features (Leung, Cheng, Fung & Poon, 1993). Prolonged reaction times, increased movement times, increased dysfluencies and evidence suggesting a high degree of limb stiffness were found by Van Gemmert and van Galen (1996) to be associated with simulation behaviour. Similar evidence of the failure to faithfully reproduce fine features in handwriting can be found in case examples

in the standard forensic document examination texts (Osborn, 1929; Harrison, 1958; Conway, 1959; Hilton, 1982; Ellen, 1989). Clearly these inconspicuous features contribute to the difficulty of the simulation process and therefore to the overall complexity of the image.

Our research is most closely related to a detailed work by Brault and Plamondon (1993) into the relationship between signature complexity and the dynamic features associated with the process of signature forgery. Their work is particularly relevant to the improvement of the performance of signature verification systems where dynamic information can be monitored directly. These authors developed an imitation difficulty coefficient to estimate the relative difficulty that an imitator would have in producing an acceptable forgery. Many of the ten basic criteria, which they review in detail and on which their model was based, are also applicable to our complexity model. The difference with our model is that we are constrained in the forensic environment by the examination of handwritten images that are static. Direct dynamic data is not attainable and cannot be used. Limited dynamic information may be inferred, depending on the type of predictor variables used (Hardy, 1992; Found, Rogers, Schmittat & Metz, 1994; Found & Rogers, 1997; Van Galen, Hardy & Thomassen, 1997). In addition, the complexity research presented in this paper is based on the reality of casework in that the conditions under which the questioned signature was performed are unknown. Ultimately our complexity model is not aimed at detecting forgeries, but rather at providing a guide to handwriting experts to prevent the expression of an erroneous decision when the signature appears to be consistent with the genuine signature.

There are a number of parameters that have been or could be proposed that are either singularly or jointly responsible for the complexity of the final image and that can be detected from a static image. Examples of these are: the number of turning points in the line, the total line length over which the turning points occur, the number of line intersections including retraced line sections, the number of pen lifts, the number of line portions where superimposition of other line portions has occurred, the presence of feathering of the line as an indicator of pressure differentials and a lack of unique characters (ie. the signature is composed of one or more repeating units). The rationale for regarding many of these parameters as components of complexity have been reviewed by Brault and Plamondon (1993), summaries of which appear in Found and Rogers (1996).

The results of our pilot study provided evidence that the most useful predictors of experts' perceptions of image complexity is a measure of the number of turning points, the number of feathering points and the number of intersections and retraces. It was found that the total line length and the number of pen lifts were not of use. The total line length was most likely excluded from the statistical model due to the high correlation between this measure and the occurrence of other parameters; that is, the longer the signature, the more likely it is to exhibit a greater number of turning points, intersections and retraces, etc. There is also a practical advantage for the absence of a requirement for examiners to take a measurement of total line length as it requires specific software and can be time consuming. It was thought that visual counting methods for predictor variables would be more likely to produce a model that was useful.

An explanation of the reason for the participation of the measures used in the complexity assessment is given below:

2. The number of turning points (TP) in the line

It is this number that results in the curviness of the line. For any given line length an increase in this number would result from the pen increasing the frequency of direction change. This is indirectly a measure of the dynamics of signature formation summarized in Brault and Plamondon (1993) in terms of biomechanical modelling and referred to in terms of possible measurement points in Hardy (1992) and Found, Rogers, Schmittat and Metz (1994).

3. The number of line intersections including retraced line sections (INTRT)

This is a measure of the degree to which earlier sections of the line are overwritten by later sections. This element is important, as it can confuse the simulator as to the pen direction of any given intersecting portion. In addition, the pattern formed

may be difficult to simulate purely on the grounds that the features and proportions ultimately formed may be a composite of intersecting portions of the signature separated in time but not space.

4. The presence of feathering of the line: Number of feathering points (FEATH)

Feathering of the line is usually a result of pressure differentials between the writing surface and the writing implement. These types of features are usually associated with a fluently written formation. Clearly, it is more difficult to correctly simulate a signature, capturing not only the spatial likeness but also the fluency of the line itself. If, alternately, the standard signature displays no feathering and has poor line quality which is evident in pauses, tremor, etc., then this greatly diminishes the difficulty associated with simulating that image.

5. Experiment 1. Construction of the model

The aim of the experiment was to investigate whether experts' perceptions of the complexity of a static signature could be predicted by a statistical model based on a discriminant function analysis. The classification scheme constructed was then used to determine which predictor variables were most useful. The validity of the model was tested in Experiment 2.

6. Method

Thirteen forensic handwriting examiners employed at Police forensic laboratories were asked to independently group 300 signatures (collected from university students) according to the following criteria:

Group 3: In the expert's opinion, given that the features fall within the range of variation of the standard signature group, these signatures are simplistic and would not warrant any opinion with respect to whether or not they are genuine.

Group 2: In the expert's opinion, given that the features fall within the range of variation of the standard signature group, these signatures exhibit some elements which would be difficult to simulate and therefore a qualified opinion would likely be expressed that they are genuine.

Group 1: In the expert's opinion, given that the features fall within the range of variation of the standard signature group, these signatures exhibit many elements which would be difficult to simulate and therefore a full (unqualified) opinion would likely be expressed that they are genuine.

Forward stepwise discriminant function analyses were performed with SPSS software using the three feature variables TP, INTRT and FEATH as predictors for classifications into the three groups. These predictor variables were determined visually by individuals trained in the technique and were independently checked by a forensic specialist.

TP was determined according to the following criteria. The starting point and terminating point of any continuous line trace was counted as one point each. To count the major turning points along the line, a small pointer was used to follow the trajectory of the line according to the sequence of formation. Whenever the pointer had to be pushed in a new direction, that point was counted as one. The total score was the sum of starting and terminating points and the number of points counted along the line. Diacritic marks were excluded from the counting process. Figure 1 shows an example of a signatures and its TP score.

To calculate INTRT, the trajectory of the line trace in the direction of formation was followed. The number of times where the line either intersected with, or retraced over, previously formed sections were counted. Figure 2 is an example of a signature and its INTRT score.

FEATH were determined by counting the number of times the line tapered to a significant extent. An example of this feature would be where the width of the line trace reduced as the pen was lifted off the page whilst it was still moving across the paper. Since this parameter was entirely subjective, the result was confirmed independently by two additional examiners.

Using discriminant function analyses, a number of models were constructed. These included models for each expert, group models, and a model for experts who classified signatures similarly.

7. Results

7.1 Experiment 1

To consider the variations in experts' perceptions of complexity, we chose to model each of the thirteen expert's results independently. Two examples of how well the model (derived from an individual's ratings)

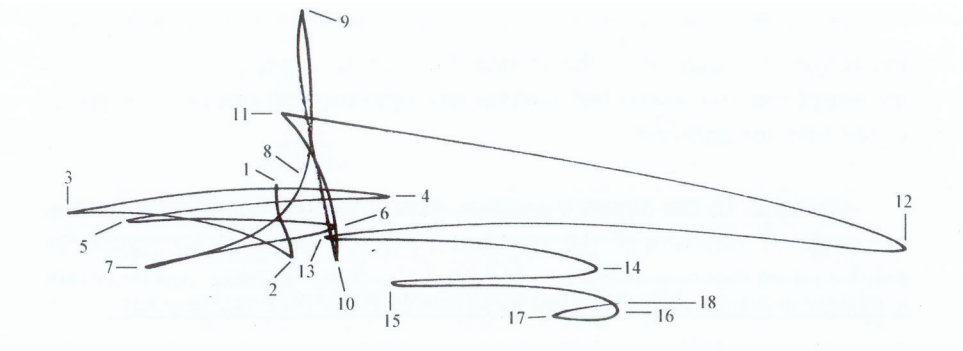


FIGURE 1. Example of a signature illustrating the application of the method used to manually count the number of turning points associated with each signature (TP=18).

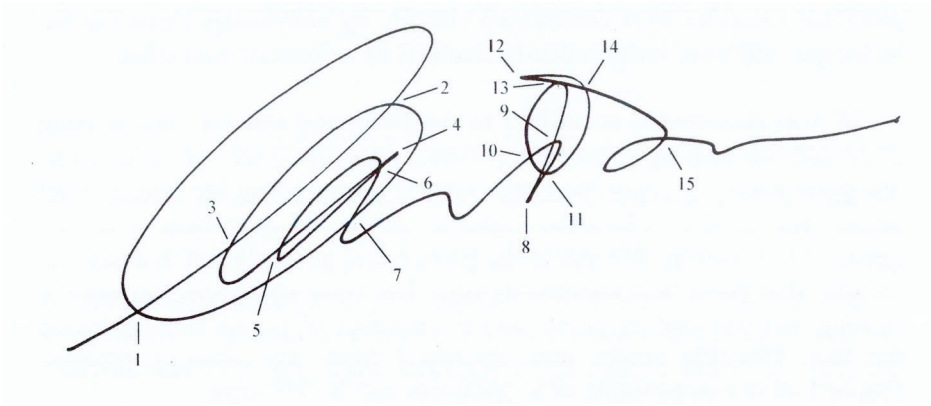


FIGURE 2. Example of a signature illustrating the application of the method used to manually count the number of intersections and retraces associated with each signature (INTRT=15).

predicted an individual's actual stated perception of difficulty are shown in Tables 1 and 2. For each table the second column shows the number of signatures that the examiner classified as either Group 1, 2 or 3. The three right hand columns show the number (and percentage) of those classified by the model as either Group 1, 2 or 3. For example, the subject whose results are shown in Table 1, considered 44 of the signatures to belong to Group 1, whereas the model for this subject predicted that of those 44 signatures, 21 belonged to Group 1, 14 belonged to Group 2 and 9 belonged to Group 3. For this subject the overall agreement between the model and the individual's actual classification (percentage of grouped cases correctly classified) was 58.7%. For expert 13 whose results are shown in Table 2, the model based on this individual's groupings would have predicted a total of 82.9% of groupings in common with the expert. As can be seen in the table for this subject, the model never predicted a signature as Group 3 when the expert rated the signature as Group 1 and never predicted a signature belonged to Group 1 when the expert had rated the signature as Group 3.

Across all of the experts tested there was a variation in the ability of the discriminant analysis to use the predictor variables to construct a model. Table 3 shows for each subject the percentage of cases correctly classified by a model derived from each examiner's assessment of complexity. In eight instances the models were calculated using only two predictor variables (TP and INTRT), as the discriminant function analysis rejected the third variable because the inclusion of the third variable (FEATH) did not increase the percentage of grouped cases correctly classified.

The percentage of grouped cases correctly classified for all experts combined is also shown in Table 3. The criteria of signature group inclusion into

Actual Group	No. of Cases	Predicted Group Membership		
		1	2	3
Group 1	44	21 (47%)	14 (31.8)	9 (20.5%)
Group 2	125	36 (28.8%)	60 (48.0%)	29 (23.2%
Group 3	131	6 (4.6%)	30 (22.9%)	95 (72.5%)

% of Grouped cases correcly classified = 58.7%

TABLE I. Results of the classification scheme from expert 1 ‘s assessment of complexity using the three predictor variables TP, INTRT and FEATH.

Actual Group	No. of Cases	Predicted Group Membership		
		1	2	3
Group 1	222	178 (80.2%)	44 (19.8%)	0 (0.0%)
Group 2	37	0 (0.0)	33 (89.2%)	4 (10.8%)
Group 3	40	0 (0.0)	3 (7.5%)	37 (92.5%)

% of Grouped cases correcly classified = 82.9%

TABLE 2. Results of the classification scheme derived from expert l 3’s assessment of complexity using the two predictor variables TP and INTRT

the calculation of the model is that six or more of the experts grouped the signature in common. Clearly there is a filtering of the data before the model is calculated and a finite number of the original signature set is excluded due to a wide range of responses regarding the grouping. In this instance, 62.9% of the signatures could be correctly classified by the model constructed. Table 4 summarizes the classification scheme derived from all experts’ results using the predictor variables TP and INTRT.

As can be seen from Table 4, for those signatures classified by the experts as being Group 1, the model calculated a proportion of these sig.natures as Group 2 and a smaller proportion as Group 3. For those signatures classified by the experts as Group 3, the model calculated a proportion of these signatures as Group 2 and a smaller proportion as Group 1. The results for misclassification of Group 1 signatures as Group 3 (4.9%) and Group 3 signatures as Group 1 (4.3%) indicate that the model was able to effectively

Expert Code	% of grouped cases correctly classified using TP, FEATH and INTRT	% of grouped cases correctly classified using TP and INTRT
1	58.7	57.0
2		78.3
3	58.9	58.2
4		68.7
5		76.5
6	81.3	81
7		67.3
8	60.3	59.7
9	60.7	59.7
10	62.0	63.3
11		64.7
12	81.2	82.2
13		82.9
All Experts		62.9
Experts 2, 5, 6, 12 & 13		83.2

TABLE 3. Summary of"% of grouped cases correctly classified' results of the classification scheme derived from examiners' assessments of complexity using two and three predictor variables.

categorize signatures as being more likely to be identifiable versus those where no opinion should be expressed.

The decision was made, based on the pilot study and the profile of the percentage of grouped cases correctly classified for the individual results, that the final model would be constructed only on those experts where more than 75.0% of signatures could be correctly classified by the model. Experts 2, 5, 6, 12 and 13 fell into this group (see Table 3). A new model was constructed on the basis of these experts' classifications which we have termed the concordant model. The criteria for assigning a signature to a particular classification group was that three or more of the five experts classified the signature in common. Table 5 represents the classification rates for the concordant model calculated on this basis. The concordant model correctly classified 83.2% of signatures, which was the highest percentage of grouped cases correctly classified for all the models used (see Table 3). This overall percentage corresponded to the correct classification of 80.0% for Group 1, 84.2% for Group 2 and 95.6% for Group 3. Although there is substantial misclassification relating to Group 2, there were no expert-grouped signatures misclassified as Group 3 when they were classified as Group 1 and vice versa.

Actual Group	No. of Cases	Predicted Group Membership		
		1	2	3
Group 1	123	75 (61.0%)	42 (34.1%)	6 (4.9%)
Group 2	47	14 (29.8%)	14 (29.8%)	19 (40.4%)
Group 3	94	4 (4.3%)	13 (13.8%)	77 (81.9%)

% of Grouped cases correcly classified = 62.9%

TABLE 4. Results of the classification scheme derived from all experts' assessments of complexity using the two predictor variables TP and INTRT. Final signature groupings were determined when six or more of the experts grouped a signature in common.

Actual Group	No. of Cases	Predicted Group Membership		
		1	2	3
Group 1	195	156 (80.0%)	39 (20.0%)	0 (0.0%)
Group 2	57	2 (3.2%)	48 (84.2%)	7 (12.3%)
Group 3	45	0 (0.0%)	2 (4.4%)	77 (81.9%)

% of Grouped cases correcly classified = 62.9%

TABLE 5. Results of the classification scheme derived from experts who scored above 75.0% in the individual test assessment of complexity using the two predictor variables TP and INTRT. Final signature groupings were determined when three or more of the five experts grouped a signature in common.

Table 6 summarizes the classification function coefficients for the concordant model constructed from the five experts, the results of which appear in Table 5. These classification function coefficients are what can be used to classify signatures whose groups are unknown. This is accomplished by placing the value of the TP and INTRT into the three equations constructed from this Table. These equations are:

Group I value= (0.3407762 x TP) + (0.2397084 x INTRT) -9.418039
Group 2 value= (0.1685134 x TP) + (0.08713504 x INTRT) -2.915064
Group 3 value = (0.09862483 x TP) -(0.02637828 x INTRT) -1.508095

From these calculations three numbers are generated, one for each of the groups. The classification

	Predicted Group Membership		
	1	2	3
TP	0.3407762	0.1685134	0.09862483
INTRT	0.2397084	0.08713504	-0.02637828
Constant	-9.418039	-2.915064	-1.508095

TABLE 6. Classification function coefficients for the concordant model constructed on experts 2, 5, 6, 12 and 13.

prediction based on the model for an unknown signature is simply the equation whose value is higher than the other two. In this way new signatures can be classified. It is also by this process that the model itself can be validated.

7.2. Experiment Validation of the model

As an indicator of the validity of the model constructed, fourteen experts, including those used to construct the initial model, were given 193 new signatures approximately six months after the original classification test. The same instructions, outlined in the Methods section of Experiment I, were given to the experts regarding these signatures. The value of the predictor variables for each signature was determined and their group classification calculated using the equations given above, based on the classification function coefficients given in Table 6. The classification groups calculated using the model and assigned by each expert were then compared. Table 7 is a summary of the results of this comparison and shows the range of total percentage agreement scores for the fourteen experts tested. These scores range from 34.9% to 70.9%.

We note from the raw data, which is reflected in the breakdown of the error scores in Table 7, that the 34. 9% agreement rate for expert F was unusual when compared with the remaining experts. For example, there is a 17.2% misclassification of signatures that the model would have predicted were signatures that were complex and that expert F registered as simplistic. This compares to no misclassification where the model predicted the signatures were simplistic and expert F believed that they were complex. This, in combination with the remaining error data, suggests that expert F was considerably more conservative and therefore had vastly different perceptions of the complexity of formations than the remaining subjects, or there was a basic misunderstanding of the basis of the test associated with this expert. In any event, the results of this expert are largely filtered out by the techniques used to generate the mean scores represented in Table 7.

The mean values in Table 8 were calculated by averaging experts' complexity groupings and rounding the final value to an integer. This final score was then compared to the concordant model's classification for each signature and the total % agreement and distribution of misclassification scores calculated. This process was carried out for all experts, for all subjects excluding expert F, and for the experts 2, 5, 6, 12 and 13. The exclusion of expert F makes only a small difference to the final distribution of error scores.

The last three columns in Table 8 provide the general misclassification rate: that is, when either the model predicted that a signature was Group 3 and the expert's perceptions were that it was Group 1 or vice versa. As can be observed, there was no error associated with this type of misclassification. The majority of the errors are associated with misclassification of Group 1 and Group 2 signatures. A comparison of the error

Expert	Total % Agreement	*Error 1:3*	*Error 3:1*	*Error 2:3*	*Error 3:2*	*Error 1:2*	*Error 2:1*	*Error 1/3*	*Error 2/3*	*Error 1/2*
A	62	1	0	8.3	5.7	8.3	14.6	1	14.1	22.9
B	61.5	1	0	11.5	3.6	11.5	10.9	1	15.1	22.4
C	62	2.6	0	14.4	2.1	14.1	5.2	2.6	16.1	19.3
D	54.7	9.4	0	22.4	1	7.3	5.2	9.4	23.4	12.5
E	61	1	0	16.7	0.5	16.1	4.7	1	17.2	20.8
F	34.9	17.2	0	39.6	0	6.3	2.1	17.2	39.6	8.3
G	62.5	0.5	0	6.8	3.1	14.6	12.5	0.5	9.9	27.1
H	50.5	0	3.1	2.1	12.5	1	30.7	3.1	14.6	31.8
I	60.9	0	0.5	8.9	3.1	14.1	12.5	0.5	12	26.6
J	57.4	0	1	3.1	7.8	7.8	22.9	1	10.9	30.7
K	58.3	0	0.5	2.6	6.8	0.5	31.3	0.5	9.4	31.8
L	70.9	0	0	5.2	3.1	3.1	17.7	0	8.3	20.8
M	50.5	0	1	2.1	15.1	0	31.3	1	17.2	31.3
N	59.9	0	0	2.1	9.9	1	27.1	0	12	28.1

'Error x:y' indicates the % error where the model calls a signature as 'x' and the specialists call it as 'y'. 'Error x/y' indicates the % error where the model calls a signature as 'x' or 'y' and the specialists call it as 'y' or 'x'. For example Error 1:3 is where the model predicted a signature belonged to group 1 and the examiner rated the signature as group 3. Error 3: 1 is where the model predicted a signature belonged to group 3 and the examiner rated the signature as group 1. Error 1/3 is the total of these mismatched groupings.

TABLE 7. Total percentage agreement and distribution of misclassification by the concordant model when compared to experts' results on the validation set of signatures. Results presented by expert.

values given in Table 7 shows that there is no difference in the error rates within a group; that is, when the model predicts Group 3 and the expert's opinion was that the signature was group 1 versus the reverse of this, for comparisons between Groups I and 3 and 2 and 3. There was, however, a significant difference at $p<0.05$ between Groups I and 2 (see Table 9). The data indicates that the model is more conservative than the experts at the 2: 1 level, with more errors associated with the model predicting signatures as being Group 2 signatures where the experts grouped them as Group 1.

Expert	Total % Agreement	*Error 1:3*	*Error 3:1*	*Error 2:3*	*Error 3:2*	*Error 1:2*	*Error 2:1*	*Error 1/3*	*Error 2/3*	*Error 1/2*
Mean (all experts)	66.2	0	0	4.7	3.6	9.9	15.6	0	8.3	25.5
Mean (all experts-F)	67.8	0	0	5.7	3.6	13	9.9	0	9.3	22.9
Experts 2, 5, 6, 12 & 13 validation results	72.9	0	0	3.1	0	26	21.4	0	3.1	24

'Error x:y' indicates the % error where the model calls a signature as 'x' and the specialists call it as 'y'. 'Error x/y' indicates the % error where the model calls a signature as 'x' or 'y' and the specialists call it as 'y' or 'x'. For example Error 1:3 is where the model predicted a signature belonged to group 1 and the examiner rated the signature as group 3. Error 3: 1 is where the model predicted a signature belonged to group 3 and the examiner rated the signature as group 1. Error 1/3 is the total of these mismatched groupings.

TABLE 8. Total percentage agreement and distribution of misclassification by the concordant model when compared to experts' results on the validation set of signatures. Results calculated by the mean signature classification over all ex.perts and the majority view of signature classification for experts 2, 5, 6, 12 and 13 (validation expert codes J, E, K, N and M respectively)

Error Type	3 and 1	3 and 2	2 and 1
1 and 3	0.3925	*	*
2 and 3	*	0.4138	*
1 and 2	*	*	0.044

TABLE 9. P values calculated for t-tests comparing direction of misclassification error rates for groups I and 3, 2 and 3 and I and 2.

Table 10 provides p-values for comparisons between non.directional group misclassification derived from Table 7. As can be observed, there is a significant difference, at $p<0.001$, associated with errors between each of the groups. In general, based on the perceptions of a limited expert group, the model is very good at discriminating between Group 3 and Group 1 signatures, has a small error rate associated with discriminating between Group 3 and Group 2 signatures, and has quite a large error rate when discriminating between Group 1 and Group 2 signatures.

8. Discussion

Discriminant function analysis is a commonly used statistical technique which provides a means of classifying objects into groups according to the value of variables associated with the objects that can be measured, taking into account an actual classification independently performed. In this experiment the objects for classification were signature formations and the variables were TP, INTRT and FEATH. The values for these variables were counted for 300 signature formations and separately checked.

Error Type	3 and 1	3 and 2
1 and 3	0.0001	0.0001
2 and 3	*	0.0007

TABLE 10. P values calculated fort-tests comparing direction of misclassification error rates between groups I and 3, 2 and 3 and 1 and 2.

The independent classification was performed by thirteen forensic handwriting experts according to the descriptions given in the methods section of Experiment 1. The strategy by which these experts classified the signatures was not investigated by the experimenters. The experts were not provided with any cues regarding the process by which the investigation of their perceptions would be carried out.

If we accept that there is validity associated with expert opinion regarding the authorship of questioned writings, then we must make an inference that experts are able to make valid judgements regarding when it is that an image is too simplistic to warrant an opinion. The relationship between image complexity and issues of writer identification have been articulated and form the basis of alternate forensic theory regarding writer identification (Found & Rogers, 1995). It is thought that visually identifiable features associated with the questioned writing provide the examiner with information of some type which would support the hypothesis that the image would be difficult to simulate successfully. Although a mathematical delineation of the identity of these features has not been carried out, it may be that simple and relatively accessible image characteristics could be used to predict the perceptions of the experts. Potential predictor variables used in this study were based on the findings of previous experimentation (Found & Rogers, 1996). This previous study was undertaken as a preliminary investigation of the theory and was based on a small number of signatures in both the model construction and validation stage of the experiment. In addition, the forensic experts used in the pilot study were trained and employed in one organisation only. The perceptions of these experts could not, therefore, be easily justified as representing the majority of government experts in the field nationally. The thirteen experts used in the current study were drawn from four police forensic laboratories and were the product of a greater number of training regimes. In addition, the experts varied with respect to their age, sex, and the number of years that they had been exclusively examining handwriting as Document Examiners.

The discussion of the results of the current study is divided into two stages. The first deals with issues associated with the construction of the classification model. The second is the validation stage of the classification model.

9. Construction of the classification model

There are a number of factors that can affect the process by which the classification models the entered data and the final accuracy of the model based on both the misclassification rate of the original data and the validation data. The choice of potential predictor variables can have a significant impact on the accuracy of the model, particularly when at.tempting to simplify a three-dimensional static handwritten image into a series of numbers. Clearly, these numbers cannot accurately describe a given image and can therefore only be seen as a sample of the information that we observe.

The mathematics underlying discriminant analysis are also based on a number of assumptions about the data. For example, it is assumed that each group is a sample from a population that is normal and multivariate, and that the variables are independent. Data such as that calculated for total line length, the number of turning points and the number of feathering points in handwriting traces needs to be approached with some caution, as previous unpublished studies by the authors indicates that there can be a significant

correlation between these factors. The discriminant function has been found to dispose of these variables in the calculation of the model as the high correlation results in functions becoming mathematically redundant due to the inability of correlated data to efficiently discriminate between groups. Two variables that are highly correlated, such as total line length and the number of turning points, are unlikely to end up as both being predictors in a model where the values of these variables are both entered.

Another source of variation is associated with the perceptions of the complexity of the signature by the experts themselves. In each case the experts classified the signatures without collaboration with other experts and in the absence of known techniques to do so in an objective fashion. The treatment of the data in the pilot study reflected this variation by having to apply criteria by which the final grouping of any one signature was made. There are a number of ways that this can be approached. The average result can be taken and rounded to an integer value representing the complexity grouping, the most frequent common grouping can be calculated or the majority view, if one can be found, can be utilised. This variation between experts is in reality quite complex in nature and can be related to factors such as the training they received, how conservative they are and the validity of opinion levels regarding authorship.

Possibly the most important issue with investigations of this type is the determination of the relationship that exists between expert perceptions and case realities. A discussion of this issue was presented by Hecker (1996) and focused on the question of whether experts may be too conservative regarding the ease or difficulty simulators experience in copying an image successfully. The perceptions of experts ultimately can only be tested through validation studies whereby, for example, the expert is forced to express an opinion regarding the authorship of a questioned signature in spite of its apparent complexity. The expert's perception of the complexity could be recorded, or the complexity grouping could be provided by a model such as is being developed here, and then compared to the error rates associated with the opinions expressed. Should a significantly higher error rate be found with those signatures that the expert or the model predicted as being simplistic, this would provide support for the validity of the expert's complexity prediction.

It is optional whether classification models are generated purely on expert group averages or concordant groups according to the criteria already mentioned. These models are therefore constructed on group data that is, to some extent, filtered. To enhance the discussion regarding the variations on experts' perceptions of complexity, we chose to model each of the thirteen expert's results independently. The data used to calculate each of these models represent the perception of the relative complexity of each of the 300 signatures by the experts. As can be observed, the models used either all three variables TP, INTRT and FEATH as predictors, or two of the variables to the exclusion of FEATH. For each expert there is a misclassification rate; that is, an error where the model, based on the predictor variables used, would not have predicted the actual expert's classification. Across all of the experts tested we found a variation in the ability of the discriminant analysis to use the predictor variables to construct a model. This illustrates the diversity of experts' perceptions regarding the complexity phenomena. It must be stressed at this point that '% of grouped cases correctly classified scores do not necessarily indicate that any given expert is grouping according to perceptions that are incorrect. It may be that it is just that the predictor variables being used are able to better predict the grouping of some experts' perceptions over others. For those experts that scored well in the '% of grouped cases correctly classified' score, it does however indicate that there is an illustratable mathematical relationship between the basis of their perception and variables associated with the images that are being subjectively processed by them.

For the classification scheme derived from all of the experts' results using the predictor variables TP and INTRT, 62.9% of the signatures were able to be correctly classified by the model constructed. This com.pares with 73.5% for the model calculated by Found and Rogers (1996). The discrepancy in this score is not surprising, due to the increased number of experts participating in the study in conjunction with the significantly larger test signature set (126 in the pilot study versus 300 in the current model). The most significant misclassification associated with this

section of the study appears to be associated with the Group 2 signatures. As can be seen from Table 4, the misclassification profile is somewhat similar for those signatures classified by the experts as being Group 1, where the model calculated a proportion of these signatures as Group 2 and a smaller proportion as Group 3, and Group 3 where the model calculated a proportion of these signatures as Group 2 and a smaller proportion as Group 1. The most significant finding from these observations is that there is much variation in the perceptions by experts of the complexity of signatures where a qualified level of opinion would be expressed. The model constructed on the results of all experts was ineffective in grouping signatures of this type and in fact was found to misclassify these signatures mostly as Group 3 signatures. These results did, however, indicate that a model could be constructed which was able to effectively categorize signatures as being more likely to be identifiable versus those where no opinion should be expressed. The misclassification rate with respect to this, excluding that rate associated with the Group 2 qualified level of opinion, was found to be 4.9% and 4.3% respectively.

The concordant model was constructed on the basis of five experts whose individual model correctly classified more than 75.0% of signatures. In constructing the concordant model, the criteria used to classify signatures into the expert classification groups was that three or more experts classified the signature in common. This model had the highest percentage of signatures correctly classified (83.2%). In addition, the profile of misclassification proved to be more acceptable. The finding that there were no expert grouped signatures misclassified as Group 3 when they were classified as Group 1 and vice versa, was a particularly useful result indicating the model clearly distinguished between signatures considered identifiable versus ones for which no opinion should be expressed.

10. Validation of the model

For the validation trials there was a range of total percentage agreement scores for the fourteen experts who participated. These scores ranged from 34.9% to 70.9%. The mean values calculated by averaging experts' complexity groupings and rounding the final value to an integer provided total percentage agreement scores better than the majority of the scores for the individual examiners. In addition, the misclassification rate was generally better for group results than for individual results. For example, there were no errors when either the model predicted that a signature was Group 3 and the experts' perceptions were that it was Group 1 or vice versa. The majority of the errors are associated with misclassification of Group 1 and Group 2 signatures. The results indicate that the model is more conservative than the experts at the 2:1 level, with more errors associated with the model predicting signatures as being Group 2 signatures where the experts grouped them as Group 1.

In general, based on the perceptions of a limited expert group, the model is very good at discriminating between Group 3 and Group 1 signatures, has a small error rate associated with discriminating between Group 3 and Group 2 signatures, and has quite a large error rate when discriminating between Group l and Group 2 signatures. Again, this error is likely to reflect a problem regarding the validity of expressing opinions according to levels whose meaning is not clearly defined or able to be articulated easily (Sjerps, Massier & Wagenaar, 1996).

The previous pilot study conducted by the authors indicated that the agreement rate with the model rose significantly when the validation phase was approached from an alternate direction. Instead of re-testing experts independently, it is possible to use the model to classify the validation set and then present each of the validation signatures to the experts, inform them of the model's classification, and ask them to either agree or disagree with the model. The agreement rate in the pilot study rose from 64.5% for both experts, to 92 and 85%. There is no evidence that would suggest that a similar result would not be found with this study although, because of a lack of experts, we have been unable to investigate this.

The model developed during this study was successful in predicting a total of between 67.8 and 72.9% of the experts' grouped perceptions as indicated in the validation experiment. It should be noted that the method of grouping used to construct the model was in a sense artificial, in that the normal questioned-to-standard examination protocol used in these cases was not adhered to. Issues associated with the relationship, if one exists, between the complexity

grouping and the range of variation in the known material were ignored in these trials. The experiment also excluded signatures exhibiting poor line quality. There were no examples of these signatures in our sample. There is, however, a theoretical relationship between complexity and line quality in that it would be expected that as the line quality decreased so would the assessment of complexity, as the ease with which the image could be copied would increase.

The perceptions of our experts as to the ease or difficulty with which an image could be copied have not been validated. Brault and Plamondon (1993) used an 'Expert Examiner Opinion' classification of complexity and compared this to the opinions of imitators (forgers) and to a mathematically generated dissimilarity index. They found poor agreement between the expert's classification and the other two measures. Although an explanation for this finding was proposed, the validity of expert opinion on this point still remains unreported. Our study was not designed to validate expert opinion regarding complexity. It does, however, provide support for the notion that this profession could introduce standard tests where collective perceptions, such as were tested here, could be standardised. Standardisation of tests into statistical forms makes the process of validation significantly more straightforward. This applies not only to decisions regarding complexity, but also to the area of methodology.

We suspect that the role of complexity in handwriting may be far more central to the field than the aspects that we have investigated here suggest (Found & Rogers, 1998). We have proposed a number of theoretical relationships between the elements that determine an image's complexity and the theory of the basis of how a nexus is able to be established between populations of written images. These relationships are:

1. as we increase the number of strokes in an image its complexity increases;

2. as the complexity of the image increases, the likelihood of another writer sharing the same elements in the handwriting decreases; and

3. as we increase the complexity of an image, we decrease the likelihood of that image being successfully reproduced by another individual.

We would argue that it is these fundamental relationships that allow opinions to be expressed regarding the authorship of handwriting. Each of these relationships is theoretically able to be validated. The complexity theory is an alternative paradigm to the notion of handwriting identification on the basis of class and individual characteristics.

The field of forensic handwriting examination has been criticized on scientific grounds from a number of sources (Risinger, Denbeaux & Saks, 1989; Huber & Headrick, 1990). This study is one of a number of research projects carried out by the authors in response to these criticisms, whose aim is to inject more objectivity and accountability into the methodology. Tests similar to this one can be designed to standardize opinions regarding spatial consistency of questioned signatures and line quality assessments. Modifications of the sorts of models statistically constructed can also be used to supplement existing training methods.

Any index of complexity finally settled upon can at best be a guide for examiners. There may well be instances where a particular signature would fall short of the complexity criteria for some previously unaccountable reason, but would be, in the opinion of the examiner, worthy of judgment. At least, however, the signature would be flagged as less than optimal and the precise reasons for its upgrading would need careful consideration and explanation in the courtroom environment.

11. Conclusion

The study presented here provides handwriting experts with a test that can be applied during casework to supplement individual perceptions as to the ease or difficulty with which an image could be simulated successfully. This may prove particularly useful for those examiners who work alone and whose individual perceptions cannot be balanced by alternative views. It is hoped that the model presented here will not only assist in individual casework, but will provide a mechanism by which the elements of expert disagreement in this area can be more easily investigated.

12. References

Brault, J., & Plamondon, R. (1993). A complexity measure of handwritten Curves: Modeling of dynamic signature forgery. IEEE Transactions on Systems, Man, and Cybernetics, 23, 400-412.

Conway, J.V.P. (1959). Evidential documents. Illinois: Charles C Thomas.

Ellen, D. (1989). The scientific examination of documents: Methods and techniques. West Sussex: Ellis Horwood Limited.

Found, B., & Rogers, D. (1995). Contemporary issues in forensic handwriting examination. A discussion of key issues in the wake of the Starzecpyzel decision. Journal of Forensic Document Examination, 8, 1-31.

Found, B. and Rogers, D. (1996). The forensic investigation of signature complexity. In M. Simner, G. Leedham & A. Thomassen (Eds.), Handwriting and Drawing Research: Basic and Applied Issues, Amsterdam: IOS Press, pp. 483-492.

Found, B. & Rogers, D. (1998). A consideration of the theoretical basis of forensic handwriting examination: The application of "Complexity Theory" to understanding the basis of handwriting identification. International Journal of Forensic Document Examiners, 4, 109-118.

Found, B., Rogers, D., & Schmittat, R. (1994). A computer program designed to compare the spatial elements of handwriting. Forensic Science International, 68, 195-203.

Found, B., Rogers, D., Schmittat, R., & Metz, H. (1994, November). A computer technique for objectively selecting measurement points from handwriting. Paper presented at the 12th Australian and New Zealand International Symposium of the Forensic Sciences, Auckland, New Zealand.

Hardy, H.J .J. (1992). Dynamics of the writing movement: Physical modelling and practical applications. Journal of Forensic Document Examination, 5, 1-34.

Harrison, W. R. (1958). Suspect documents, their scientific examination. New York: Praeger.

Hecker, M.R. (1996). Subjective elements in the evaluation process of disputed signatures. Proceedings of the 5th European Conference for Police and Government Handwriting Experts. The Hague, The Netherlands, 13-15 November.

Hilton, 0. (1982). Scientific examination of questioned documents. New York : Elsevier Science Publishing Co., Inc.

Huber, R.A., & Headrick, A.M. (1990). Let's do it by numbers. Forensic Science International, 46, 209-218.

Kao, H.S.R., Shek, T.L., & Lee, E.S.P. (1983). Control modes and task complexity in tracing and handwriting performance. Acta Psychologica, 54, 69-77.

Leung, S.C., Cheng, Y.S., Fung, H.T., & Poon, N.L. (1993). Forgery I-Simulation. Journal of Forensic Sciences, 38, 402-412.

Meulenbroek, R.G.J., & van Galen, G.P. (1990). Perceptual-motor complexity of printed and cursive letters. Journal of Experimental Education, 58, 95-110.

Muehlberger, R.J. (1990). Identifying simulations. Journal of Forensic Sciences, 35, 368-374.

Osborn, A. S. (1929). Questioned documents (2nd ed.). Chicago: NelsonHall Co.

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

Schneider-Pieters, H., ten Camps, C. & Hardy, H. (1996). The computer - friend or foe? Proceedings of the 5th European Conference for Police and Government Handwriting Experts, The Hague, The Netherlands, 13-15 November.

Sjerps, M.J., Massier, R.E.F., & Wagenaar, W.A. (1996). Expressing expert opinion using a verbal probability scale. Proceedings of the 5th European Conference for Police and Government Handwriting Experts. The Hague, The Netherlands, 13-15 November.

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

van der Platts, R.E., & van Galen, G.P. (1990). Effects of spatial and motor demands in handwriting. Journal of Motor Behaviour, 22, 361- 385.

Van Galen, G.P. (1984). Structural complexity of motor patterns: A study on reaction times and movement times of handwritten letters. Psychological Research, 46, 49-57.

Van Galen, G.P., Hardy, H.J.J., & Thomassen, A.J.W.M. (1997). State, trait and environmental influences on the dynamics of handwritinggeneration as possible clues for forensic analysis. In: W. de Jong(Ed.), Proceedings III International Congress of the Gesellschaft furForensische Schriftuntersuchung (GFS). Luzern, September 10-13,1997.

Van Gemmert, A.W.A. & van Galen, G.P. (1996). Dynamic features of mimicking another persons writing and signature. In M.L. Simner, C.G. Leedham & A.J.W.M. Thomassen (Eds.), Handwriting and drawing research: Basic and applied issues (pp. 459-471). Amsterdam: IOS Press.

Wing, A.M. (1978). Response timing in handwriting. In G.E. Stelmach(Ed.), Information processing

in motor control and learning (pp. 153-172). New York: Academic Press.

Acknowledgment: This research was funded by the National Institute of Forensic Science, Australia.

Journal of forensic document examination (Online)
ISSN 0895-0849

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.

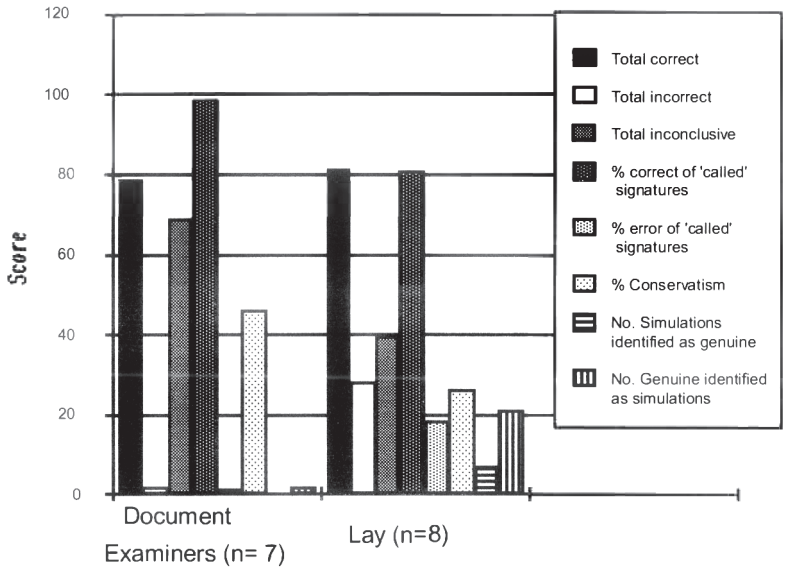


FIGURE 1. Mean results of document examiners and lay persons.

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).

Journal of forensic document examination (Online)
ISSN 0895-0849

THE OBJECTIVE STATIC ANALYSIS OF SPATIAL ERRORS IN SIMULATIONS

Bryan Found[1,2], Doug Rogers[1] and Hermann Metz[2]

Abstract. *The Pattern Evidence Analysis Toolbox software (Found, Rogers & Schmittat, 1994) has been specifically designed to take accurate spatial measurements from static handwriting traces including signatures. Forensic handwriting specialists in casework frequently encounter signatures of questionable authenticity. Some criticism has been levelled at this forensic field resulting from the lack of objective data used to draw conclusions regarding the authenticity of questioned signatures. In this study a range of spatial measurements of 200 known signatures, collected from 10 individuals, was compared to 140 forgeries of their signatures made by 14 forgers. It was found that the forgeries as a group did display significant numbers of spatial errors when compared to genuine signatures. The results indicate that measurement of spatial errors could be a source of information which can be used to discriminate between possible simulations and genuine signatures, and provide data on the types of errors likely to occur. Information obtained in this study has been used for the development of software (Found, Rogers & Schmittat, 1998), which may ultimately be practicable in the forensic environment.*

Reference: Bryan Found, Doug Rogers, Hermann Metz (1999, Vol. 12 - reformatted and reprinted). The Objective Analysis of Spatial Errors in Simulation. J. Forensic Document Examination, Vol. 29, pp. 61 - 71.

Keywords: Simulation, spatial errors, measurement strategies.

1. Introduction

Forensic handwriting specialists frequently encounter cases involving questioned signatures. Harrison (1958), in his chapter on signature forgery, lists seven classes that suspect signatures may fall into. There are signatures which, upon examination, appear completely unlike the signatures that they are purporting to be. There are forged signatures of individuals that do not exist. There are traced signatures drawn onto documents using a genuine signature as a guide, while freehand simulations are drawn onto a document freehand. Then there are questioned signatures that are genuine and are disputed either as a result of the signature being obtained by trickery, the author honestly not believing that he wrote it, or those signatures where the genuine writer has modified the formation, usually for the purpose of denial at a later date. A not uncommon form of forgery that must be added to this list is where a genuine signature has been photocopied onto a document. The task for the handwriting examiner is to distinguish between the possible classes that a questioned signature may fall into.

The main technique used to distinguish between classes of questioned signatures in the forensic environment is based on visual indicators (see Found & Rogers, 1999, this issue). Detailed descriptions of the features that are assessed subjectively in the determination of the authorship of a disputed signature appear in most forensic texts on the subject (Conway, 1959; Harrison, 1958; Hilton, 1982; Osborn, 1929; Ellen, 1989). However, it has been argued by some (Huber & Headrick, 1990, 1999) that forensic handwriting examination cannot be considered as a scientific discipline without the incorporation of objective measurement techniques. Huber and Headrick (1990) state, "Our studies of handwriting for identification purposes have always taken into consideration some measurable features, such as size, relative heights, spacing, though the recording of the

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

measurements has not been standard practice Until we do so we must accept the fact this area of our work does not meet the criteria of science." Totty and Hardcastle (1986) on assessing the 'SIGNCHECK' signature authentication system state that future systems "may produce information about signatures which could augment the information currently available to the document examiner." Clearly, there is some support for the incorporation of measurements into the existing comparison methodology.

There are many techniques available to the document examiner that measure handwriting. Potentially relevant software continues to come on line through advances in signature verification systems and optical character recognition research. For the forensic practitioner, however, the data produced does not necessarily provide clear answers regarding the *class* that a questioned signature may fall into. This is a result of both theoretical and practical considerations. In the casework environment, unlike the environment constructed for signature verification techniques, the examiner has to contend with usually limited amounts and qualities of both questioned and specimen material. In addition, the time period over which the material was produced may vary considerably. Theoretically, if it is found that a questioned signature is spatially dissimilar to a specimen group, then this does not imply that an individual other than the genuine writer wrote it. Consequently, the use of objective measures in signature comparisons will likely be limited to the stage in the method where a decision is made regarding whether the questioned signature is similar or dissimilar to the specimen material.

A promising technique to obtain objective measures of line quality from static images is being developed (Frank & Grube, 1998). The study reported here involves a technique that provides spatial consistency information only. It is an early work carried out prior to the development of computer software such as the 'Angular Differential' (Found, Rogers & Schmittat, 1997) and 'Matrix Analysis' (Found, Rogers & Schmittat, 1998). This study aimed to investigate only one type of signature, the freehand simulation, using the PEAT software (Found, Rogers & Schmittat, 1994). A simulated signature is one that has not been performed using the normal generalized motor program for the genuine signature. This may result from the use of a motor program by someone other than the genuine writer, or by the genuine writer using a different motor program. We will refer to the simulations in this study as forgeries, only because we know that they were written for the purpose of deception by individuals other than the genuine writer.

Simulations can be made under a variety of circumstances and on a variety of documents, which may make the act more or less difficult for the simulator. Many of the normal sources of variation in routine case examinations have been controlled. In this investigation we have chosen the *one-off* simulation as might occur at a transaction point. These simulations are produced on a specific document where the forger only has one attempt to reproduce it for the purpose of a deception. The forgers were, however, given unlimited practice prior to this attempt. In addition, the comparison material was comprised exclusively of requested signature specimens taken in one sitting. One would expect, therefore, on the basis of investigations of normal variations conducted by authors such as Evett and Totty (1985), that the normal range of variation in the signature would be unlikely to be captured fully.

The aims of the experiment were firstly to determine whether *transaction point* forgeries exhibited measurable spatial errors as compared with genuine signatures. Secondly, for the spatial errors detected, we aimed to determine which parameter type they were most likely to be associated with. Thirdly, our aim was to determine whether spatial errors could be a source of information which could be used to discriminate between possible simulations and genuine signatures.

2. Method

2.1 Participants

Ten volunteers from the Victoria Forensic Science Centre provided signatures and gave the experimenters permission for their signatures to be simulated and used in this study. For the purpose of the study, the providers of the genuine signatures will be referred to as victims. Fourteen staff members of the School of Human Biosciences, at La Trobe University participated as forgers.

3. Material and Apparatus

3.1 Signatures

Twenty-five signatures, each executed on blank sheets of A4 paper with a ball point pen, were received from each of ten volunteers from the Victoria Forensic Science Centre. A random sample of three signatures from each volunteer was provided to the forgers to use as models.

3.2 Measurement technique

The spatial parameters of the genuine victim signatures and the simulated signatures were measured using PEAT software in conjunction with an image processing package (NIH Image version 1.57) on a Macintosh II computer.

Since handwritten images are relatively small, it was necessary to enlarge them before the scanning process. This was achieved using an enlarging photocopier. Images requiring analysis were enlarged to approximately fit across an A4 sheet of paper. A calibration grid accompanied each image through the enlargement process. The enlarged images and calibration grid were scanned into a computer and saved as PICT files. Once all the images had been scanned, they were processed using NIH Image software. This processing routine was carried out on the image to set the upper and lower grey scale limits that resulted in the image appearing as a complete and continuous line. Images were converted to a binary form by setting the image pixels to black and all other pixels to white. A skeletonization routine was applied that reduced the lines in the image to a thickness of one pixel. The processed images were saved in a MacPaint format (72 dpi).

4. Procedure

The forging aspect of this investigation was run as a competition over a period of approximately six months. A small prize was offered to the most successful forger, according to the spatial analysis, and the running scores were updated publicly as each new forgery was completed by the group. In all, 140 forgeries were collected from the subjects. Simulations were made on blank sheets of A4 paper. The following instructions were given to the forgers:

You have been provided with 3 signatures taken from each of ten victims whose signatures you wish to forge. The plan is that you intend to pass at ten different banks withdrawal slips bearing the forged signature of each of the victims. However, this particular banking organization has introduced new security measures. They only provide you with one blank document on which to produce the signature and the signature must be produced on banking premises.

Your task is to learn to perform each of the signatures. You can take as much time as you like to practise each of the signature formations. You must sign your signature only once on the official banking document provided. You therefore only have one chance to produce the final forgery of each of the ten victims' signatures. Since the signature must be produced in the vicinity of a banking official, you cannot trace the signature or use mechanical aids (eg. a photocopier).

You must adhere to the following criteria:

1. The signature must be a freehand simulation of the victim's signature being copied.
2. The signature must be written using a ball-point pen.
3. When forging on the official document, only one attempt can be made for each signature. You may have a copy of each of the victims' signatures beside you for reference.

Subjects practised each signature between 50 and 250 times before providing the one-off simulation on the "official banking document".

5. Data Analysis

Measurements

Two forensic handwriting specialists and one academic jointly decided the parameters to be measured and compared. Parameters for 20 genuine signatures were measured to obtain the range of variation of the specimen material. This was done on each of the victims' signatures prior to the collection of the forgeries. Measurements were made of the 14 forgeries per victim and between 2 and 5 of the remaining genuine signatures. These measurements were used for comparison with the range of variation in the specimen signatures. The parameter types and the abbreviations used to refer to them are given below.

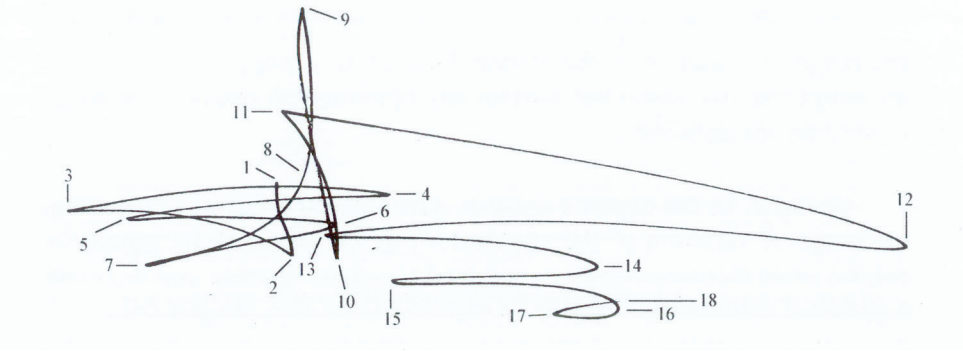


FIGURE 1. An example of a signature illustrating numbered feature points be tween which specific measurements were performed.

5.1 *Total line length (TLL).* This is a measure of the curvilinear line length for the entire image.

5.2 *Total area (TAREA).* This is a measure of the total area enclosed by the line trace forming the signature.

5.3 *Area of enclosed loops (LOOP).* This is a measure of specific areas enclosed by the line trace. An example of a loop can be seen in the sig nature of Figure 1 between points marked as 16 and 17.

5.4 *Length of specified lines (SPEC).* This is a measure of the line length between two specific feature points that can be visually identified. An example of this can be seen in the signature of Figure 1 between points marked as 11 and 15.

5.5 *Width (WIDTH*). This is a measure across the horizontal plane of the signature between two specific feature points that can be visually identified. An example of this can be seen in the signature of Figure 1 between points marked as 3 and 12.

5.6 *Diagonal (DIAGONAL).* This is a measure across the diagonal plane of the signature between two specific feature points that can be visually identified. An example of this can be seen in the signature of Figure 1 between points marked as 9 and 16.

5.7 *Height (HEIGHT).* This is a measure down the vertical plane of the signature between two specific feature points that can be visually identified. An example of this can be seen in the signature of Figure 1 between points marked as 9 and 10.

5.8 *Angle.* Two angle types were measured. Given that an angle is formed by three points in space, the ANGLE UP measurement was de fined as an angle where the middle point was taken at a feature that was at the apex of the signature. An example of this can be seen in the signature in Figure 1 between points marked as 7, 9 and 12. The ANGLE DOWN measurement was defined as an angle where the middle point was taken at a feature that was at the base of the signature. An example of this can be seen in the signature in Figure 1 between points marked as 3, 15 and 12. In both cases the first and last points from which the angle was constructed were in the medial plane of the signature, to the left and right of the signature formations.

For each parameter type listed above, the number of measurements taken for each signature varied according to the number of feature points that could be confidently identified. In general, no more than three measures of each of LOOP, SPEC, WIDTH, DIAGONAL or HEIGHT were taken for each of the victims' signatures.

6. The comparison method and calculating a spatial error score

The comparison method involved taking measurements of the same parameters for all the signatures of a victim and comparing them between the questioned and specimen groups of signatures. The range of variation for a particular parameter for the specimen group was deter mined. The measurement for this parameter for each of the victims' questioned signatures was then compared to this range. If the

measured parameter of the questioned signature fell outside the range of the specimens, this parameter was called an error (with a value of 1) for this signature. In this way a spatial error score could be generated for each of the questioned signatures, relative to the range of variation in measurements for the specimen group. In this context there may be an error score not only for forgeries, but also for genuine signatures included in the questioned group.

Since only a small number of comparison measures were taken for each signature, it was necessary to devise a scoring scheme which amplified any spatial error associated with the forgeries. Trials of these scoring procedures yielded the following scoring types, the scores for which were added to yield a compounded error score:

6.1 *Raw score (RAW).* The raw score was calculated by counting the number of times a measurement taken from a questioned signature fell outside the range of variation for that measurement in the specimen group. This error score was expressed as a percentage of the total measurements taken.

6.2 *Ratio score (RATIO).* The ratio for each of the measures associated with specified line lengths or distance between two points measurement (eg. HEIGHT, WIDTH), was calculated for each of the questioned sig natures. The ratio score was calculated by counting the number of times a ratio taken from a questioned signature fell outside the range of variation for that ratio in the specimen group. This error score was expressed as a percentage of the total ratio measurements.

6.3 *Normalized Scores.* Since questioned signatures can be larger or smaller than signatures in the victim's specimen group, and yet still retain the relative proportions of features in space, we incorporated into the error score a calculation that would compensate for this reality. Normalization selectively scales the signature features according to an adjustment made by one or more of the parameters to the mean for those parameters in the specimen group. For example, for a particular specimen signature, the width deviation from the overall width mean in the specimen group was calculated. This factor was then multiplied through the remaining parameters (compensations were made for area measurements and angles were excluded) in the specimen signature group to yield a new set of specimen comparison measurement ranges. Each questioned signature, once parameters had been normalized to the new specimen width mean, was then compared, and a normalized error score calculated. For each questioned signature, the error score was expressed as a percentage of the total measurements taken. Normalization scores were calculated for normalizations associated with total line length (NTLL), width (NWIDTH), height (NHEIGHT), total line length and width (NTLL&W), total line length and height (NTLL&H) and total line length, width and height (NTLL,W&H).

7. Statistical analysis

The error scores for each of the questioned signatures were calculated by expressing as a percentage the proportion of measures where the questioned value fell outside the range of variation of the specimen group for each test as indicated above. The error scores for each test were added to produce a final error score (compounded error score). This error score for the forged signatures in the questioned group was then compared to the error score for the genuine signatures in the questioned group, using unpaired two tailed *t*-tests to determine whether the spatial errors of these signature types differed.

8. Results

The questioned signatures analyzed for each victim included 14 forgeries and 2 to 5 genuine signatures. The error scores for each of the questioned signatures for four victims are represented in Figures 2 to 5. The error scores for each test are shown, along with the compounded error score. The full range of raw, ratio and normalized scores were made for nine victims. For victim 10 (Figure 5), a reduced number of measurements were taken, as measurement points were difficult to isolate because of the open and rounded formation of the signature. The same forgery number (x-axis on the graphs shown in Figures 2 to 5) were used for a particular forger for each victim. Inspection of the scores shown in the figures indicates a good deal of variation between forgers, and variation within forgers for different signatures.

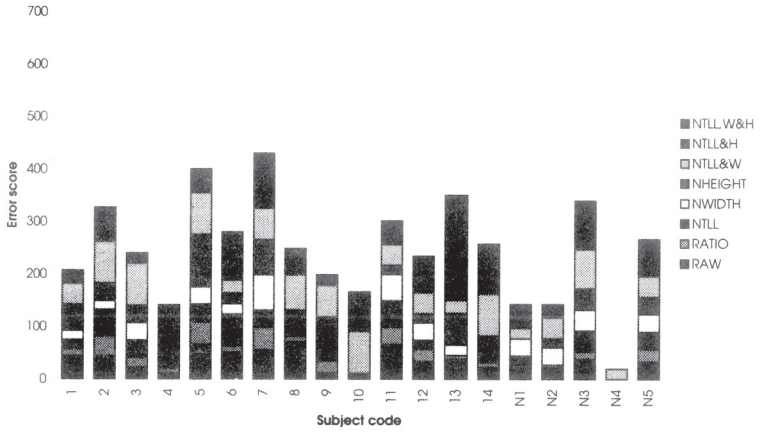


FIGURE 2. Compounded spatial error scores for the questioned signatures associated with victim 2. Forgeries are numbered 1 to 14 and genuine signatures numbered NI to N5. Maximum error score is 800. Error scores for the genuine signatures are not significantly different to the forgeries at p <.05.

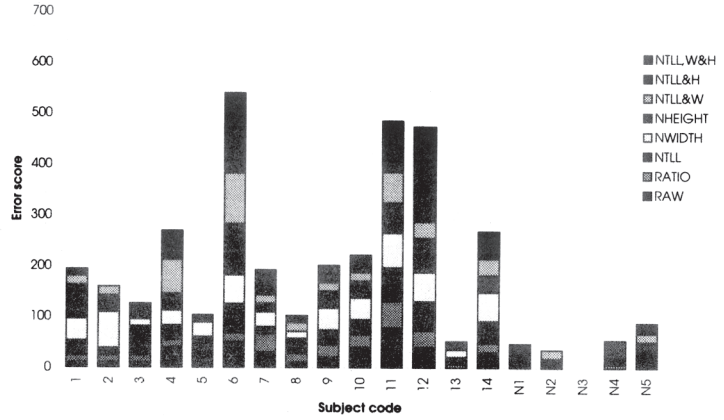


FIGURE 3. Compounded spatial error scores for the questioned signatures associated with victim 4. Forgeries are numbered I to 14 and genuine signatures numbered Nl to N5. Maximum error score is 800. Error scores for the genuine signatures are significantly different to the forgeries at p <.05.

The compounded error scores for the questioned genuine signatures were significantly less (p <.05) than for the forgeries for each of seven victims (Figures 3 to 5 are examples). For three victims (Figure 2 is an ex ample) there was no significant difference. When the forgery error scores for all victims' signatures were combined and compared to the error scores for questioned genuine signatures combined for all victims, there was a significant difference (at p<.05).

A comparison was made between the mean % spatial errors over all the victims' signatures, and the data types used to generate the compounded test score. In each case these data types could discriminate between the forged and genuine signatures in the questioned group (at p<.05). Figure 6 provides the mean percentage spatial error score for the questioned signatures for both forged and genuine signatures, versus the data test type used.

Figure 7 represents the proportion of occurrences, expressed as a percentage, where a particular parameter type was found to be in error in the forged signatures. WIDTH showed the greatest error, falling outside the range of variation for the specimen group in nearly 60% of cases, whereas TAREA had the lowest error(< 30%).

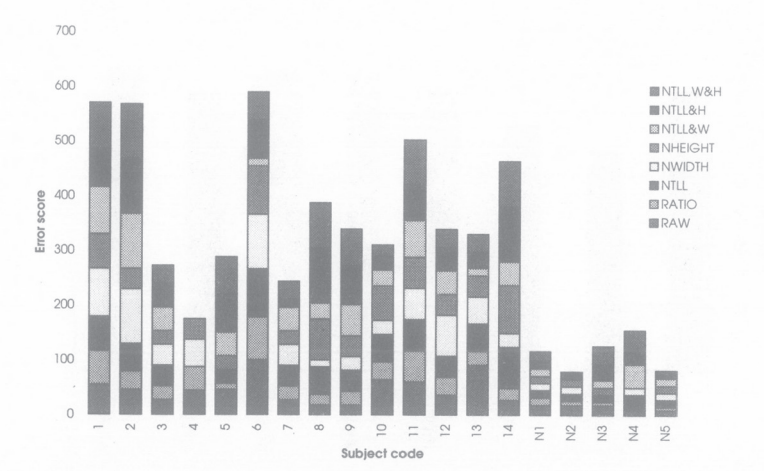


FIGURE 4. Compounded spatial error scores for the questioned signatures associated with victim 6. Forgeries are numbered 1 to 14 and genuine signatures numbered Nl to N5. Maximum error score is 800. Error scores for the genuine signatures are significantly different to the forgeries at p <.05.

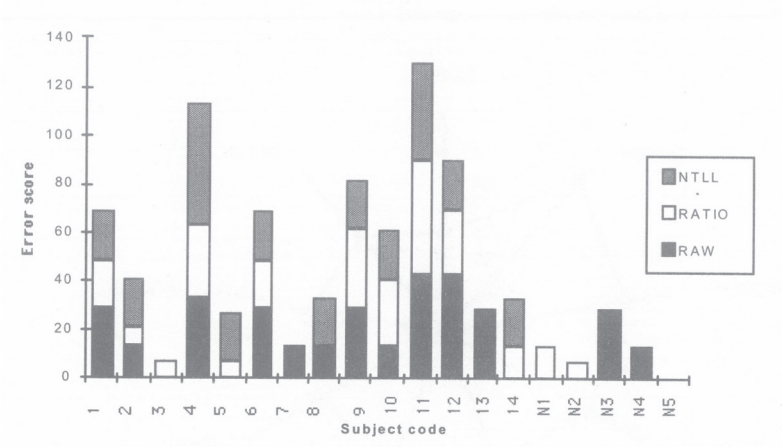


FIGURE 5. Compounded spatial error scores for the questioned signatures associated with victim 10. Forgeries are numbered 1 to 14 and genuine sig natures numbered Nl to N5. Maximum error score is 300. Error scores for the genuine signatures a re significantly different to the forgeries at p <.05

9. Discussion

Measurement strategies have been extensively used in the investigation of handwriting in the fields of motor control (eg Castiello & Stelmach, 1993; Phillips, Stelmach & Teasdale, 1991; Teulings, Thomassen & Van Galen, 1986; Wright, 1993), optical character recognition and signature verification (Han & Sethi, 1995; Leclerc & Plamondon, 1994), database searching systems both for forensic and signature authentication applications (Hecker, 1995) and to a much lesser extent in forensic handwriting examination (eg. Herkt, 1996; Philipp, 1996; Plamondon & Lorette, 1989). Many of these techniques rely on dynamic information which forensic specialists do not have direct access to. Research based on these dynamics, however, has proven directly relevant to forensic handwriting examination. Brault and Plamondon (1993) for example, investigated the relationship between signature complexity and the dynamic features associated with signature forgery. Van Gemmert and Van Galen (1996) used the dynamic investigative approach to illustrate the difference between forging and normal writing, using the relative power spectrum of the noise produced by writing

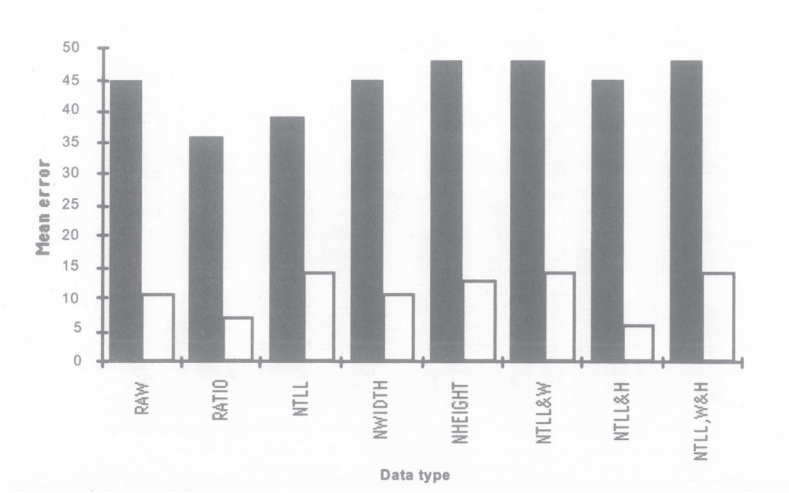


FIGURE 6. The mean% spatial error for the questioned signatures versus the data test type used. Forged signatures are represented by the black columns, and genuine questioned signatures are represented by the white columns. In each case the mean error for the forgeries is significantly different to the mean errors for the genuine signatures at $p < .05$.

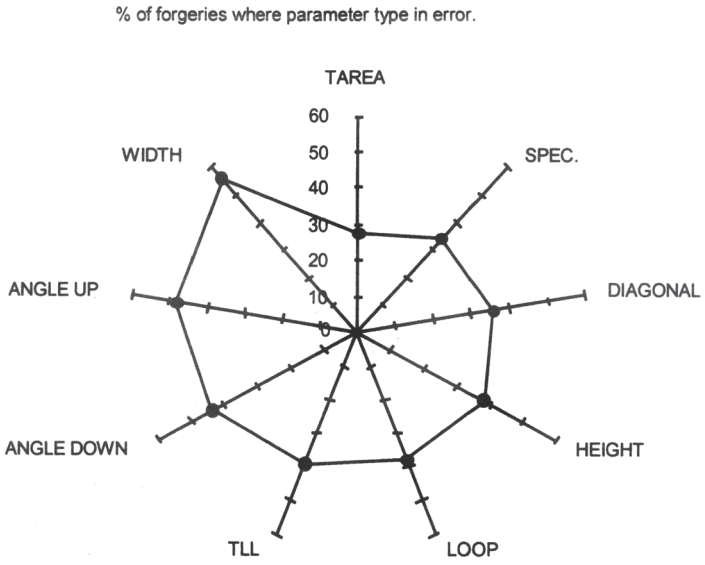


FIGURE 7. Number of occurrences in the forged signatures where the raw measurement of each parameter type fell outside the range of variation for the specimen group (expressed as a percentage).

under differing conditions. Van Gemmert, Van Galen, Hardy and Thomassen (1996) used a similar technique to investigate the dynamic characteristics associated with individuals disguising their writing. The advantage of data generated from studies of this type is that it is objective and can be tested.

Techniques based on research that can be incorporated into the forensic method for comparing handwriting traces, may ultimately be developed. The current study, although limited with respect to the number of measurement points compared and the method for selecting these measurement points, does provide significant support to the hypotheses that spatial disturbances can result from simulation behaviour, and that such disturbances can be objectively measured. Although line quality was not

considered in this study, the vast majority of signatures suffered reduced line quality in addition to the spatial errors detected. In comparison, al though spatial errors were also detected in the genuine questioned images, no such line quality deterioration was evident. Due to a number of factors, compounding the error scores was found to more successfully illustrate the spatial error, as compared with the error calculated from raw measurements alone. For example, the questioned signature may have been proportionally consistent with the range of variation in the specimen group, but may have been performed to a scale not characterized by that group. Therefore, if we were to take the raw measurement alone, then a signature even slightly larger or smaller than the range of sizes in the specimen group would produce a large error score. In forensic science it is not unusual to observe writing behaviour that varies over time with respect to the size of the signature. This variation may even be due to the size of the space allocated to the signer on the document. Raw measurements alone, therefore, may produce an unrealistic picture of the spatial consistency. Ratio scores compensate for any error in the raw score due to this factor. The normalization scores highlight proportional differences in a different way. Normalization effectively standardizes all signatures being compared to a mean measure of a particular parameter or combination of parameters. This technique would likely be more effective should a much larger sample of measurements be taken. Nevertheless, the technique used in the current study appeared effective, despite the normalization process reducing the number of available comparison measurements.

If we compare the error scores between the grouped questioned genuine signatures and the grouped forgeries for each victim, we find that three of the ten victims' signatures, did not exhibit significant error scores (at $p < .05$). One victim had only two signatures in the questioned genuine group which were likely to effect any calculation of significance. The signature of victim 2 was pictorially quite variable, and manifested in two of the questioned genuine signatures, generating a high error score (Figure 2). Clearly, variation of this nature is likely a limiting factor in interpreting the significance of spatial error scores. The signature of the other victim was relatively simplistic and variable. Forgers, therefore, had less difficulty capturing the spatial character of the signature, which is reflected in the non-significant p value.

The balance of the victims' signatures did show a significant difference between the spatial error scores of the forgeries as a group, and the questioned genuine signatures as a group. Figure 4 is an example where most forgers had difficulty capturing the spatial features, as evidenced by the high compound error scores. The genuine questioned signature error scores were relatively low, indicating that the genuine writer was fairly consistent.

Of interest to us was the fact that, except for a few instances, there was an error score for the genuine, questioned signatures. This indicates that the 15 signatures in the specimen group did not provide sufficient range of variation to include all spatial parameters of the genuine signatures taken from an individual. While this was expected in most cases due to the nature of the signatures we used and previous observations (Evett & Totty, 1985; Totty & Hardcastle, 1986), it needs to be taken into account in future refinements of such objective techniques.

The individual's ability-to capture the spatial features of the signature being forged does vary to some extent, as can be observed by the differential height of the graphs showing the compounded error scores. Subject 13 is an example of a good forger of many signatures (see for example Figures 3 & 5) yet relatively poor with others (eg., Figures 2 & 4).

Although it was advantageous to use the compounded error scores for the individual signatures, the comparison between the mean percentage spatial errors over all the victims' signatures, versus the data types used to generate the compounded test score (see Figure 6), indicated the data types were useful on their own. In each case these data types could discriminate significantly between the forged and genuine signatures in the questioned group. The technique used is, therefore, able to discriminate between these forged and genuine signatures under the strict controls of this experiment.

The parameters measured from the writing trace (raw measurements) were considered individually to see how well particular parameters correlated with the forgery process in our population of subjects. This

was done by comparing the number of occurrences in the forgeries where particular parameter appears to be relative measures of width. This does tend to make sense in that forgers, when drawing out the line, do so in a serial way. This may compromise their ability to reproduce spatial relationships separated in both time and space. The parameter least often found to be in error was measure of total area. It would appear that this results from the phenomena that this measure can vary quite markedly in response to slight differences in the movement of the pen. For example, if two portions of the line separated in time but not space did not intersect in one signature specimen but did in another, then the range of variation in the measure of that parameter could be very large. A large range of variation in a parameter provides the least difficulty for the forger to reproduce so that it falls within that range of the genuine signature group.

Experimental evidence (Leung, Cheng, Fung & Poon, 1993; Leung, Fung, Cheng & Poon, 1993; Van Gemmert & Van Galen, 1996) indicates that forgers concentrate on the spatial features of the handwriting they are producing in preference to capturing the dynamic features of the movement. Nevertheless, the results of the current study show that spatial relationships are difficult for individuals to capture accurately when forging signatures as a *one-off* simulation.

The analysis technique trialed here indicates that a number of aspects of the measurement of static signatures require development and improvement. Problems encountered include the significant amount of time taken, from scanning the images to generating a result, and the selection of appropriate measurement points. Examples of suitable solutions to these problems have been sought by the authors and have been reported (Found, Rogers & Schmittat, 1997; 1998). Future techniques should be aimed at incorporating spatial and line quality data together to objectively generate an error or consistency score. Handwriting specialists can then use this information at the stage where they determine whether the questioned image under examination is similar or dissimilar to the range of variation exhibited in the specimen material. Once this opinion has been reached, the expertise of the examiner can be used to focus on the appropriate propositions that explain the similarities and dissimilarities.

10. Conclusion

The technique employing PEAT software was successfully applied in this investigation to provide objective spatial error scores resulting from measurements of forged and genuine signatures. It was found that a significant number of spatial errors were made when individuals attempted to forge the signature of others. Techniques of this type have the potential in the future to offer forensic handwriting specialists methods to determine objectively those spatial features in signatures that are likely to reflect simulation behaviour. Future techniques should focus on characteristics associated with both space and line quality, to provide a useful scoring procedure.

11. References

Brault, J., & Plamondon, R. (1993). A complexity measure of handwritten curves: *Modelling of dynamic signature forgery. IEEE Transactions on Systems, Man, and Cybernetics*, 23, 400-412.

Castiello, U., & Stelmach, G.E. (1993). Generalised representation of handwriting: Evidence of effector independence. *Acta Psychologica*, 82, 53-68.

Conway, J.V.P. (1959). *Evidential Documents*. Illinios: Charles C Thomas.

Ellen, D. (1989). *The scientific examination of documents: Methods and techniques*. West Sussex: Ellis Horwood Limited.

Evett, I.W., & Totty, R.N. (1985). A study of the variation in the dimensions of genuine signatures. *Journal of the Forensic Science Society*, 25, 207-215.

Found, B., & Rogers, D. (1998). A consideration of the theoretical basis of forensic handwriting examination: The application of "Complexity Theory" to understanding the basis of handwriting identification. *International Journal of Forensic Document Examiners*, 4, 109-118.

Found, B., Rogers, D., & Schmittat, R. (1994). A computer program de signed to compare the spatial elements of handwriting. *Forensic Science International*, 68, 195-203.

Found, B., Rogers, D., & Schmittat, R. (1997). Recovering dynamic in formation from static handwriting traces using 'angular differential' software. *Journal of Questioned Document Examination*, 6, l, 17-38.

Found, B., Rogers, D., & Schmittat, R. (1998). 'Matrix Analysis': A technique to investigate the spatial properties of handwritten images. *Journal of Forensic Document Examination*, 11, 54-74.

Frank, K., & Grube, G. (1998). The automatic

extraction of pseudodynamic information from static images of handwriting based on marked gray value segmentation. *Journal of Forensic Document Examination*, 11, 17-38.

Han, K., & Sethi, I. (1995). Signature identification via local association of features. *Proceedings of the Third International Conference on Document Analysis and Recognition*, Montreal, Canada, pp 187-190.

Harrison, W.R. (1958). *Suspect documents, their scientific examination*. New York: Praeger.

Hecker, M. (1995). Fish and Chips. *A video presented at the Proceedings of the 7th Conference of the International Graphonomics Society*, London, Canada, August 6-10.

Herkt, A. (1986). S*ignature disguise or signature forgery? Journal of the Forensic Science Society*, 26, 257-266.

Hilton, 0. (1982). *Scientific examination of questioned documents*. New York: Elsevier Science.

Huber, R.A., & Headrick, A.M. (1990). Let's do it by numbers. *Forensic Science International*, 46, 209-218.

Huber, R.A., & Headrick, A.M. (1999). *Handwriting identification: Facts and fundamentals*. Boca Raton: CRC Press LLC.

Leclerc, F., & Plamondon, R. (1994). Automatic signature verification: The state of the art 1989-1993. lnternational Journal of Pattern Recognition and Machine Intelligence, 8, 643-660.

Leung, S.C., Cheng, Y.S., Fung, H.T., & Poon, N.L. (1993). Forgery I - Simulation. J*ournal of Forensic Sciences*, 38, 402-412.

Leung, S.C., Fung, H.T., Cheng, Y.S., & Poon, N.L. (1993). Forgery II - Tracing. J*ournal of Forensic Sciences*, 38, 413-424.

Osborn, A.S. (1929). *Questioned documents* (2nd ed.). Chicago: Nelson Hall.

Philipp, M. (1996). On the use of signature verification systems in handwriting identification service. Paper presented at the Fifth European Conference for Police and Government Handwriting Experts. The Hague, The Netherlands.

Phillips, J.G., Stelmach, G.E., & Teasdale, N. (1991). What can indices of handwriting quality tell us about Parkinsonian handwriting? *Human Movement Science*, 10, 301-314.

Plamondon, R., & Lorette, G. (1989). Automatic signature verification and writer identification - The state of the art. *Pattern Recognition*, 22, 107-131.

Teulings, H.L., Thomassen, A.J.W.M., & Van Galen, G.P. (1986). In variants in handwriting: The information contained in a motor pro gram. In H.S.R. Kao, G.P. Van Galen, & R. Hoosain (Eds.), *Graphonomics: Contemporary research in handwriting* (pp. 305-315). Amsterdam: North Holland.

Totty, R.N., & Hardcastle, R.A., (1986). A preliminary assessment of the SIGNCHECK system for signature authentication. *Journal of the Forensic Science Society*, 26, 181-195.

Van Gemmert, A.W.A., & Van Galen, G.P. (1996). Dynamic features of mimicking another person's writing and signature. In M.L. Simner, C.G. Leedham, & A.J.W.M. Thomassen (Eds.), *Handwriting and drawing research: Basic and applied issues* (pp. 459-471). Amsterdam: IOS Press.

Van Gemmert, AW.A., Van Galen, G.P., Hardy, H.J.J., & Thomassen,J.W.L. (1996). *Dynamical features of disguised handwriting*. Paper presented at the Fifth European Conference for Police and Government Handwriting Experts, The Hague.

Wright, C.E. (1993). Evaluating the special role of time in the control of handwriting. *Acta Psychologica*, 82, 5-52.

12. Acknowledgement

The authors acknowledge Mr David Black, Mr John Ganas and Mr Jim Brandi for their contribution in the data collection phase of this research.

This research was funded by the National Institute of Forensic Science, Australia.

Journal of forensic document examination (Online)
ISSN 0895-0849

The Skill Of A Group Of Forensic Document Examiners In Expressing Handwriting And Signature Authorship And Production Process Opinions

Bryan Found [1,2] Doug Rogers [1] and Allan Herkt [3]

Abstract. *Between March 1998 and June 2001, the six members of the New Zealand Police Document Examination Section completed six blind forensic handwriting and signature trials where the identity of the questioned writings were known by the experimenters but not by the document examiners. A total of 7494 opinions were expressed by the examiner group regarding the authorship of questioned handwriting and signature samples. Of these opinions, 2700 were correct, 11 were erroneous and 4783 were inconclusive. This translates into an overall raw error score of 0.1% of opinions, and a 'called error' score (one calculated by excluding the inconclusive data) of 0.4% of opinions. In addition, a total of 2982 opinions were expressed by the group on whether questioned signatures were the product of a simulation process. The group did not express any opinions that a simulated signature was the product of a genuine writing process, nor did the examiners express any opinions that a genuine signature was the product of a simulation process. Thus, for process opinions there was a zero error rate for the group. This paper overviews the individual and group opinion data associated with each of the trials. 'The results provide significant support to the validity and strong probative value of the skill that has been claimed by this group of examiners.*

Reference: Bryan Found, Doug Rogers, Allan Herkt (2001, Vol. 14 – reprinted and reformatted). The Skill of a Group of Forensic Document Examiners in Expressing Handwriting and Signature Authorship and Production Process Opinions. J. Forensic Document Examination, Vol. 29, pp. 73 - 82.

Keywords: Forensic Expertise Profiling Laboratory, handwriting, error rates, document examiners' skill

1. Introduction

The desire for the New Zealand Police Document Examination Section (NZPDES) to expose themselves to extensive and sustained blind testing of their claimed skill in forensic handwriting identification can be sourced originally to the concerns raised within the now historically significant Risinger, Denbeaux and Saks (1989) publication on the topic. Specifically of interest to the current study were the following criticisms raised by these authors:

- No court anywhere has ever explicitly considered and passed on its (handwriting identification) claim to validity.
- There exist almost no studies of its claims in any academic literature.
- Such studies as have been conducted, published and unpublished, raise serious questions as to its validity.
- The law has resisted requiring presentation of the asserted expertise in ways that would expose its validity problems.

It appeared obvious from the lack of published validation trials internationally that the criticisms raised were valid and, more importantly, could largely be addressed through the administration of

1 Forensic Expertise Profiling Laboratory, School of Human Bio.sciences, La Trobe University, Victoria, 3086, Australia.
2 Document Examination Team, Victoria Forensic Science Centre, Forensic Drive, Macleod, Victoria, 3085, Australia.
3 Document Examination Section, Wellington Central Police Station, Victoria Street, PO Box 693, Wellington, New Zealand

blind tests. The Special Advisory Group (Document Examination), which represents police and government document examiners in Australia and New Zealand, has been collaborating with the now Forensic Expertise Profiling Laboratory (La Trobe University, Australia) to carry out such trials since the mid 1990's. The NZPDES is one of the participants in this process and, in addition to the direct testing component of the trials, has participated in researching other aspects of forensic handwriting examination such as method development and documentation (Found & Rogers, 1999).

Since our collaborative research interest in validation testing of forensic handwriting examiners has commenced, a number of relevant events have occurred and important studies have been published on the topic. Judge McKenna in the Starzecpyzel decision (United States v. Starzecpyzel, 1995) stated that, 'The Daubert hearing established that forensic document examination which clothes itself with the trappings of science, does not rest on carefully articulated postulates, does not employ rigorous methodology, and has not convincingly documented the accuracy of its determinations. Forensic handwriting identification was, in spite of this statement, recognized as a practical skill and, therefore, held to be admissible in evidence. Risinger and Saks (1996) argued that an implication of this decision was the potential for plummeting validation standards for admissibility, which may result in the burden falling on the opponent, "to prove affirmatively that the skilled witnesses cannot do what they claim they can do". Clearly this is not a position that forensic examiners and legal specialists would be comfortable with. These authors then open the window to allowing some resolution of the concerns by stating that, "science can examine the dependability of such a process (handwriting identification) even when the process is not a science." Science has commenced to do so.

Forensic validation studies have been reported by Kam, Fielding and Conn (1997), Kam, Wetstein and Conn (1994) and Found, Sita and Rogers (1999). These studies have provided some support for the expertise claimed by practitioners, or at least those that have been tested within the trials, in terms of it being real and demonstrable. In each case an error score has been reported. It is this error score that is relevant to document examiner client groups, particularly the judicial system. It is the magnitude of the error score that best dictates the probative value of the evidence being presented.

In order to assess the magnitude of any error, the Forensic Expertise Profiling Laboratory has adopted a philosophy of testing, largely based on the criticisms of this field historically. Specifically, our philosophy attempts to address the following guiding statements:

1. "The level of correctness of the assertions made by examiners from day to day casework is not likely to prove to be a credible source for the (validation) data needed" (Huber & Headrick, 1999).
2. "A process such as handwriting identification presents a number of potential subtasks dealing with variables such as writing instruments, forgery of various sorts, age, health and so forth. No single test can map the abilities of any one practitioner, or any group of practitioners" (Risinger & Saks, 1996).
3. "A great many tests... would be necessary to know what, if anything, (examiners) can do accurately, and under what conditions"(Risinger & Saks, 1996).
4. "A complete testing regime would have tests which covered the entire spectrum of conditions and difficulties" (Risinger & Saks, 1996).

The results presented in this paper represent the NZPDES results on trials completed between March 1998 and June 2001. It should be noted that this laboratory has historically recorded one of the lowest error scores amongst the groups participating in our trials. In spite of this, these examiners were keen to bring into evidence issues surrounding the probative value of the skill that they had traditionally claimed. This overview of their testing results does not contain the minutia of details associated with the construction of each of the trials, a task we felt was best left to reports concerning each of the trials independently with the inclusion of all participants' data.

2. Overview of the trials included in this report

2.1. Trial 1

This trial was an upper-case handwriting trial. Examiners were provided with original samples of questioned and specimen writings. The specimen material was produced by three individuals. Examiners were required, amongst other tasks, to compare specimen writings with a total of 134 questioned samples. The questioned samples were requested normal (written by specimen writer and other writers), disguised (written by the specimen writer and other writers) or simulated writings (written by specimen writer and other writers). For each of the questioned samples examiners were required to express a '*direction of identification*' opinion, a '*direction of exclusion*' *opinion*, or an *inconclusive opinion*. Opinions were marked as either correct, incorrect or inconclusive. The opinions of each of the examiners were not subjected to a peer-review process.

2.2 Trial 2

This trial incorporated both questioned signatures and handwriting. Examiners were provided with 30 questioned documents (withdrawal slips), each with a signature, and 5 distinct samples of handwriting for opinion. Samples from 2 individuals were provided for comparison purposes. Each distinct sample of the questioned handwriting was written by one or other of the specimen writers. Each questioned signature was either a genuine signature by a specimen writer or a simulation. For each of the questioned handwriting samples examiners were required to express a direction of identification opinion, a direction of exclusion opinion, or an inconclusive opinion. For each of the questioned signatures examiners were required to express a direction of identification opinion, a simulation opinion, or an inconclusive opinion. Opinions were marked as either correct, incorrect or inconclusive. The opinions of each of the examiners were subjected to a peer-review process.

2.3 Trial 3

This trial was a handwriting trial. Examiners were provided with original samples of questioned and specimen writings. One individual produced the specimen material. Examiners were required, amongst other tasks, to compare the specimen writings with a total of 250 questioned samples. The questioned samples were requested normal (written by specimen writer and other writers), disguised (written by specimen writer and other writers), or simulated writings (written by specimen writer and other writers). For each of the questioned samples examiners were required to express a direction of identification opinion, a direction of exclusion opinion, or an inconclusive opinion. Opinions were marked as either correct, incorrect or inconclusive. The opinions of each of the examiners were subjected to a peer-review process.

2.4 Trial 4

This trial was a signature trial. Examiners were provided with examples of a specimen signature and were required to compare the specimen signatures with a total of 80 non-original (photocopied) questioned signatures. The questioned signatures comprised requested normal signatures and simulated signatures (written by the specimen writer and other writers). For each of the questioned signatures examiners were required to express a 'direction of identification' opinion, a 'simulation' opinion, or an inconclusive opinion. Opinions were marked as either correct, incorrect or inconclusive. The opinions of each of the examiners were not subjected to a peer-review process.

2.5 Trial 5

This trial was a signature trial. Examiners were provided with examples of a specimen signature and were required to compare them with a total of 260 original questioned signatures. The questioned signatures comprised requested normal signatures and simulated signatures (written by the specimen writer and other writers). For each of the questioned signatures examiners were required to express a direction of identification opinion, a simulation opinion, or an inconclusive opinion. Opinions were marked as either correct, incorrect or inconclusive. The opinions of each of the examiners were subjected to a peer-review process.

2.6 Trial 6

This trial was a signature trial. Examiners were provided with examples of a specimen signature and were required to compare them with a total of 250 questioned signatures. All signatures were high resolution scanned images, printed using a laser printer. The questioned signatures comprised requested normal signatures and simulated signatures (written by the specimen writer and other writers). For each of the questioned signatures examiners were required to express a direction of identification opinion, a direction of exclusion opinion, a simulation opinion, or an inconclusive opinion. Opinions were marked as either correct, in correct or inconclusive. The opinions of each of the examiners were subjected to a peer-review process.

3. Definition of scores used in this report

The development of methodology (Found & Rogers, 1999) was occurring during the administration of the trials described in this report. Incorporated in this process were changes in the definition of terms used to express opinions to more closely align to the reporting philosophies articulated in Evett (1998). To facilitate the compilation of results in this study, opinions were either treated as correct (in spite of the level of support for the proposition), incorrect (in spite of the level of support for the proposition), or inconclusive.

Examiners' authorship responses (opinion units) were marked as correct, incorrect or inconclusive. These marks were then analyzed to produce scores for each of the different questioned handwriting types (normal writing by the specimen writer, disguised writing by the specimen writer, simulated writing by the specimen writer, simulated writing not by the specimen writer, normal writing not by the specimen writer, and disguised writing not by the specimen writer). The scores are presented as numbers of opinions or as percentages, the latter representing opinion rates. The following definitions of the score categories are used in subsequent results tables in this report.

3.a # Correct
The number of authorship opinions that were correct.

3.b # Error
The number of authorship opinions that were incorrect.

3.c # Inconclusive
The number of authorship opinions that were inconclusive.

3.d % Correct
The number of correct authorship opinions divided by the total number of authorship opinions (expressed as a percentage).

3.e % Error
The number of incorrect authorship opinions divided by the total number of authorship opinions (expressed as a percentage).

3.f % Inconclusive
The number of inconclusive authorship opinions divided by the total number of authorship opinions (expressed as a percentage).

3.g % Correct called
The number of correct authorship opinions divided by the sum of the correct and erroneous authorship opinions (expressed as a percentage).

3.h % Error called
The number of incorrect authorship opinions divided by the sum of the correct and erroneous authorship opinions (expressed as a percentage).

The called scores do not include inconclusive opinions and, therefore, equate to a number that reflects the opinion rate when an examiner is expressing an opinion that is other than inconclusive.

4. Results

The results of all six of the authorized forensic document examiners with the NZPDES are included in this report. A total of 7494 authorship opinions

Writing Type	Opinion Scores							
Signatures	# correct	# error	# inc.	% correct	% error	% inc.	% correct called	% error called
	899	0	2791	23.4	0	76.6	100	0
Handwriting	1801	11	1992	47.3	0.3	52.4	99.4	.06
Handwriting and Signatures	2700	11	4783	36.0	0.1	63.8	99.6	.04

TABLE 1. Summary of authorship opinion unit scores for all opinions expressed in the trials for signature, handwriting and combined signature and handwriting samples.

Writing Type	Opinion Scores							
	# correct	# error	# inc.	% correct	% error	% inc.	% correct called	% error called
Normal by specimen writer	707	1	474	59.8	.01	40.1	99.9	.01
Normal not by specimen writer	525	1	614	46.1	.01	53.9	99.8	0.2
Disguise by specimen writer	301	1	94	76.0	0.3	23.7	99.7	.03
Disguise not by specimen writer	82	0	266	23.6	0.0	76.4	100.0	0.0
Simulated not by specimen writer	154	3	383	28.5	0.6	70.9	98.1	1.9
Simulated by specimen writer	32	5	161	16.2	2.5	81.3	86.5	13.5
Handwriting totals	**1801**	**11**	**1992**	**47.3**	**.03**	**52.4**	**99.4**	**0.6**

TABLE 2. The authorship opinion scores for the examiner group across all handwriting types represented across each of the three handwriting text trials.

have been expressed by the six examiners in the group. Five of the six examiners completed all six of the trials. One examiner did not examine one of the trials.

5. Handwriting text results

There were 3804 authorship opinions expressed by the group on handwriting text comparisons. Table 2 provides the authorship opinion scores for the examiner group across all handwriting types represented across each of the three handwriting text trials. As can be observed, the 'potential or estimated error rate' for handwriting types varies according to the questioned writing type. The % error is <1% for all handwriting text types except those samples that are simulated by the specimen writer, where the error is found to be 2.5% (a called error rate of 13.5%). Although the two simulation writing types have the highest error rates of the handwriting types, this must be balanced with the corresponding % inconclusive scores. These two categories of writing exhibit high % in conclusive scores, which indicates that examiners are more conservative when expressing opinions regarding samples of this type. In addition, the 5 errors made calling a simulated sample of writing (by the specimen writer) as not written by the specimen writer, were all made on a non peer reviewed trial, and 4 of the 8 errors were made by one individual.

Table 3 provides the scores for authorship opinions expressed for each examiner across all handwriting types represented in each of the three handwriting text trials. Note that the handwriting types are represented by the codes SP (by specimen writer), NSP (not by specimen writer), DSP (disguised by specimen writer),

examiner number	Writing Type	Opinion Scores							
		# correct	# error	# inc.	% correct	% error	% inc.	% correct called	% error called
1	DNSP	10	0	48	17.2	0.0	82.8	100	0
2	DNSP	12	0	46	20.7	0.0	79.3	100	0
3	DNSP	7	0	51	12.1	0.0	87.9	100	0
4	DNSP	10	0	48	17.2	0.0	82.8	100	0
5	DNSP	23	0	35	39.7	0.0	60.3	100	0
6	DNSP	20	0	38	34.5	0.0	65.5	100	0
1	DSP	50	0	16	75.8	0.0	24.2	100	0
2	DSP	50	0	16	75.8	0.0	24.2	100	0
3	DSP	50	0	16	75.8	0.0	24.2	100	0
4	DSP	50	1	15	75.8	1.5	22.7	98.0	2.0
5	DSP	51	0	15	77.3	0.0	22.7	100	0
6	DSP	50	0	16	75.8	0.0	24.2	100	0
1	NSP	95	1	119	44.2	0.5	55.3	99.0	1.0
2	NSP	34	0	31	52.3	0	47.7	100	0
3	NSP	93	0	122	43.3	0	56.7	100	0
4	NSP	98	0	117	45.6	0	54.4	100	0
5	NSP	106	0	109	49.3	0	50.7	100	0
6	NSP	99	0	116	46	0	54.0	100	0
1	SNSP	24	0	66	26.7	0	73.3	100	0
2	SNSP	22	0	68	24.4	0	75.6	100	0
3	SNSP	22	0	68	24.4	0	75.6	100	0
4	SNSP	25	0	65	27.8	0	72.2	100	0
5	SNSP	31	2	57	34.4	2.2	63.3	93.9	6.1
6	SNSP	30	1	59	33.3	1.1	65.6	96.8	3.2
1	SP	128	0	94	57.7	0	42.3	100	0
2	SP	62	0	10	86.1	0	13.9	100	0
3	SP	129	0	93	58.1	0	41.9	100	0
4	SP	130	0	92	58.6	0	41.4	100	0
5	SP	131	0	91	59.0	0	41.0	100	0
6	SP	127	1	94	57.2	0.5	42.3	99.2	0.8
1	SSP	6	0	27	18.2	0	81.8	100	0
2	SSP	6	0	27	18.2	0	81.8	100	0
3	SSP	3	0	30	9.1	0	90.9	100	0
4	SSP	1	4	28	3.0	12.1	84.8	20	80
5	SSP	10	0	23	30.3	0	69.7	100	0
6	SSP	6	1	26	18.2	3.0	78.8	85.7	14.3
	Totals	1801	11	1992	47.3	0.3	52.4	99.4	0.6

TABLE 3. The authorship opinion scores expressed for each examiner across all handwriting types represented in each of the 3 handwriting text trials.

Signature Type	Opinion Scores							
	# correct	# error	# inc.	% Correct	% error	% Inc.	% correct Called	% error called
Normal by specimen writer	712	0	27	96.3	0	3.7	100	0
Disguise by specimen writer	122	0	286	29.9	0	70.1	100	0
Simulated not by specimen writer	65	0	2322	2.7	0	97.3	100	0
Simulated by specimen writer	0	0	156	0	0	100	n/a	n/a
Totals	899	0	2791	24.4	0.0	75.6	100	0.0

TABLE 4. The opinion scores for all signature types represented in the trials.

DNSP (disguised not by specimen writer), SNSP (simulated not by specimen writer and SSP (simulated by specimen writer).

6. Signature results

There were 3690 authorship opinions expressed by the group on signature comparisons. Table 4 provides the grouped authorship opinion unit scores for all signature types represented in the trials. The scores are for the group of examiners as a whole, where all of the same questioned signature types from the different trials have been combined. As can be observed, no error has yet to be recorded by the group regarding the authorship of questioned signatures. It should be noted, however, that the group has not recorded any opinions where the specimen writer was excluded from having written a particular signature.

Table 5 provides the opinion scores for each examiner across all signature types represented in each of the signature trials. Note that the signature types are represented by the codes SP (by specimen writer), DSP (disguised by specimen writer), SNSP (simulated not by specimen writer and SSP (simulated by specimen writer). Although there were no errors in the direction of identification or exclusion, examiners were 100% inconclusive as to whether or not the specimen writer wrote any of the signatures that were the product of a simulation process.

7. Signature process

The determination of a writing process is not about whether or not the writer of the specimens did or did not write a particular entry, but is an opinion regarding the writing behaviour itself. From trials 2, 4, 5 and 6 it is possible to extract opinions by the group on whether or not examiners believed that questioned signatures were genuine (where it can be assumed that the examiners were of the opinion that the signatures were not the product of a simulation process), or produced using a simulation (or imitation) process. In many instances the authorship of simulated signatures is not determinable due to the difficulty in excluding the proposition that the specimen writer did not simulate his or her own signature for the purposes of denial at a later date. An opinion that a signature was produced using a simulation process can, however, be of assistance to the judiciary.

A total of 2982 opinions were expressed by the group on genuine and simulated signature samples. The scores for these process opinions are shown in Table 6. As can be observed, the group did not express any opinions that a simulated signature was the product of a genuine writing process, nor did the examiners express any opinions that a genuine signature was the product of a simulation process.

Table 7 provides the process opinion scores for each examiner for genuine and simulated signatures.

8. Discussion

There are many aims in conducting skill research of this type. Examples include whether the skills claimed by a particular group are real, what the error rate in decision making by individuals and groups is, and what the relationship is between results' profiles from different laboratories, including all the variables associated with qualifications, training programs, experience etc. At this stage in the documentation of

examiner number	Writing Type	Opinion Scores							
		# correct	# error	# inc.	% correct	% error	% inc.	% correct called	% error called
1	DSP	21	0	47	30.9	0.0	69.1	100	0.0
2	DSP	20	0	48	29.4	0.0	70.6	100	0.0
3	DSP	20	0	48	29.4	0.0	70.6	100	0.0
4	DSP	20	0	48	29.4	0.0	70.6	100	0.0
5	DSP	20	0	48	29.4	0.0	70.6	100	0.0
6	DSP	21	0	47	30.9	0.0	69.1	100	0.0
1	SNSP	0	0	400	0.0	0.0	100	n/a	n/a
2	SNSP	0	0	387	0.0	0.0	100	n/a	n/a
3	SNSP	0	0	400	0.0	0.0	100	n/a	n/a
4	SNSP	0	0	400	0.0	0.0	100	n/a	n/a
5	SNSP	0	0	400	0.0	0.0	100	n/a	n/a
6	SNSP	0	0	400	0.0	0.0	100	n/a	n/a
1	SP	123	0	3	97.6	0.0	2.4	100	0.0
2	SP	107	0	2	98.2	0.0	1.8	100	0.0
3	SP	118	0	8	93.7	0.0	6.3	100	0.0
4	SP	121	0	5	96	0.0	4.0	100	0.0
5	SP	124	0	2	98.4	0.0	1.6	100	0.0
6	SP	119	0	7	94.4	0.0	5.6	100	0.0
1	SSP	0	0	26	0.0	0.0	100	n/a	n/a
2	SSP	0	0	26	0.0	0.0	100	n/a	n/a
3	SSP	0	0	26	0.0	0.0	100	n/a	n/a
4	SSP	0	0	26	0.0	0.0	100	n/a	n/a
5	SSP	0	0	26	0.0	0.0	100	n/a	n/a
6	SSP	0	0	26	0.0	0.0	100	n/a	n/a

TABLE 5. The opinion scores expressed for each examiner across all signature types represented in the trials.

forensic handwriting examiners' skills, the most critical factors being investigated were the characterisation of examiners' skill at providing identification/exclusion evidence on different categories of writing, and the potential error rate associated with expressing opinions on those writing types. The determination of the potential error rate of the technique is important, such that the client group can choose whether the result-generating system is appropriate to that claimed and has probative value suitable for judicial use.

The results generated by the NZPDES as a group are characterised by low error rates (< 1% overall), which provide significant support to the validity of the skill that has been claimed by this group. Larger error rates are associated with opinions regarding samples of handwriting text that have been 'simulated'. The errors associated with the two 'simulation' writing types are, however, not shared by all members of the group. Eight of the eleven authorship opinion errors were made on non-peer reviewed trials and it is not unreasonable to expect that errors of this type would be significantly reduced through the normal quality peer-review practices used by this group. In addition, the continued participation in expertise profiling trials, which offer a revision and corrective action component, should maximize the opportunity for

Sample Type	Opinion Scores							
	# correct	# error	# inc.	% correct	% error	% inc.	% correct called	% error called
Genuine Samples	536	0	29	94.9	0	5.1	100	
Simulated not by specimen writer	2356	0	61	97.5	0	2.5	100	0

TABLE 6. The process opinion scores expressed by the group in the trials.

examiner number	Writing Type	Opinion Scores							
		# correct	# error	# inc.	% correct	% error	% inc.	% correct called	% error called
1	SP	94	0	3	96.9	0	3.1	100	0
2	SP	78	0	2	97.5	0	2.5	100	0
3	SP	89	0	8	91.8	0	8.2	100	0
4	SP	92	0	5	94.8	0	5.2	100	0
5	SP	94	0	3	96.9	0	3.1	100	0
6	SP	89	0	8	91.8	0	8.2	100	0
1	SNSP	384	0	21	94.8	0	5.2	100	0
2	SNSP	384	0	8	98.0	0	2.0	100	0
3	SNSP	378	0	27	93.3	0	6.7	100	0
4	SNSP	405	0	0	100	0	0.0	100	0
5	SNSP	403	0	2	99.5	0	0.5	100	0
6	SNSP	402	0	3	99.3	0	0.7	100	0

TABLE 7. The process opinion scores for each examiner for signatures by the specimen writer (SP) and simulations of the specimen writer's signature not by the specimen writer (SNSP).

perceptual and cognitive revision where the system has not produced the correct response.

Although it appears that the judiciary invests strongly in examiner experience to gauge the reliability of opinion, studies conducted at the Forensic Expertise Profiling Laboratory, incorporating the data presented here, have yet to find a simple correlation between experience (that is the number of years that an examiner has been practising forensic handwriting examination), and their correct, error and conservatism scores. Given this reality, it is proposed that the only mechanism by which the judiciary can assess the value of examiner opinion is through examiner results on independent blind trials of the types presented.

9. Utilizing potential or estimated error rates

Because of the number of varied trials undertaken by this group, we consider that the error shown is a good estimate of the group's potential error rate that can be considered when applying the technique in the casework setting. This error rate can, therefore, be reported as the group's potential error rate. It is important to consider that, although a potential or estimated error rate of < 1% is appropriate to discuss, this rate is associated with examiners making decisions on blind validation trials and then grouping the results. The grouping of results does dilute the data, as the overall data set contains a number of distinct categories of writing, and examiners' relative skill in expressing opinions about these categories does vary between the group and between wexaminers. A single trial, or even a series of trials, is unlikely to capture all of the variables associated with the routine presentation of forensic casework. Forensic handwriting examination involves an enormous number of tasks prior to a final opinion being expressed. In addition,

questioned and specimen writing can vary with respect to quantity, quality, complexity, skill etc. The error quoted is, therefore, without question an estimate based on the application of the same cognitive skill set to different types of blind trials that is used to examine handwriting and signatures in the casework environment. Since we observe an enormous amount of casework variables, the only approach available to examiners at this time is constant exposure to blind trials that emulate casework as closely as possible.

It is still the case that most examiners internationally have not been exposed to the rigours of testing of the magnitude described in this paper. For courts to take holistic comfort in error scores generated by blind trials, if in fact they take comfort at all, would be a precarious position. To take this position would be to embrace an underlying assumption that the error scores generated by the individuals taking part in the reported trials are representative of error rates over larger groups of document examiners. There is, at this point in time, no clear evidence to support this proposition. It is, therefore, in no way possible to suggest that individuals not covered by this report (that is, outside the New Zealand Police Document Examination group), should be attributed with a similar skill profile and associated error rate.

10. References

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3), 198-202.

Found, B., & Rogers, D. (Eds.). (1999). Documentation of forensic handwriting comparison and identification method: A modular approach. *Journal of Forensic Document Examination*, 12, 1-68.

Found, B., Sita, J., & Rogers, D. (1999). The development of a program for characterising forensic handwriting examiners' expertise: Signature examination pilot study. J*ournal of Forensic Document Examination*, 12, 69-80.

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

Huber, R.A., & Headrick, A.M. (1999). *Handwriting Identification:Facts and Fundamentals*, Boca Raton, CRC Press.

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

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

Risinger, D. M., Denbeaux, M.P., & Saks, M.J., (1989). Exorcism of ignorance as a proxy for rational knowledge: The lessons of handwriting 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.

Acknowledgement

The trials have received financial sponsorship from the National Institute of Forensic Science, Australia and the Senior Managers of Australian and New Zealand Forensic Science Laboratories. We acknowledge the following examiners: David Boot, Allan Herkt, Trish James, Gordon Sharfe, John Walker and Delwynne Walsh, for their participation in this project.

Comparison of document examiners' opinions on original and photocopied signatures

Bryan Found[1,2], Doug Rogers[1], and Allan Herkt[3]

Abstract: *There is a lack of empirical evidence concerning document examiners' ability to perform handwriting comparisons on photocopied writings. This study aimed to compare the accuracy of examiners' opinions on 260 original questioned signatures and on the same signatures that had been photocopied. Six of the examiners from the Document Examination Section of the New Zealand Police participated in the study, which comprised two trials. Each trial was constructed according to the accepted process of comparing a group of known (specimen) signatures with a group of questioned signatures where the writer was known to the experimenters but not to the document examiners. One trial contained originals of the specimen and questioned signatures and the other comprised photocopies of the same specimen and questioned signatures. No errors regarding authorship were made for original or photocopied signatures, and there were no instances where an identification/elimination opinion was reversed between a photocopy and its original. Only 2.3% of opinions relating to an original signature differed in any way from that offered for its photocopy. The high correct rates for questioned genuine signatures were similar for original (100%) and photocopied signatures (98%). The correct opinion rate regarding the process of production of original and photocopied, simulated, questioned signatures combined was 99.7%. The results provide evidence that examiners are able to make comparisons on a complex signature with the same accuracy and similar sensitivity when using either originals or photocopies.*

Reference: Bryan Found, Doug Rogers, Alan Herkt. (1995, Vol.14 – reprinted and reformatted). Comparison of Document Examiners' Opinions on Original and Photocopied Signatures. J. Forensic Document Examination, Vol 29, pp. 83 - 89.

Keywords: Original signatures, photocopied signatures, Examiners' opinions.

1. Introduction

Document examiners may be requested to perform handwriting examinations on documents that are photocopied. As the photocopying process produces handwriting that contains less feature information than original handwriting, many examiners are hesitant to express authorship opinions on this type of material. However, a number of authors (Hilton, 1982; Ellen, 1989; Morton, 1989), while strongly emphasising major restrictions when expressing opinions regarding non-original writings (see Discussion), consider that fruitful comparisons can often be made. Hilton (1982), regarding the examination of non-original writing, wrote that "...general handwriting can often be tentatively and sometimes be positively identified" (p. 384) and that this condition also holds for signatures. This author does, however, recognise that "Some workers refuse to examine all copies, but the practical examiner recognises that it is necessary to rely on copies at times" (p. 385). Along similar lines Ellen (1989) has written "Although some of the detail will not be apparent, in many examples of good quality photocopies there will be adequate material

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.

Document Examination Section, Wellington Central Police Station, Victoria Street, PO Box 693, Wellington, New Zealand.

for a useful comparison to be made" (p. 62), and that "It is possible to identify photocopied writing as having been made by a known writer" (p. 62). Morton (1989) presented a study on non-original signatures and handwriting reproduced using seven plain paper photocopiers. The original images were produced using combinations of four paper types and different writing implement classes (ballpoint, roller ball and fiber tip pens). This author concluded that "most of the copiers reproduced the signatures, genuine and forged, well enough for a fruitful examination" (p. 464).

Despite the perceptions of these authors, there is a lack of studies that provide evidence regarding examiners' abilities to express comparison opinions on non-original writings.

A detailed study regarding experts' assessments of line quality features in non-original signatures was presented by Dawson and Lindblom (1998). These authors investigated the extent to which the photocopying to which the photocopying inhibits the ability of experts to assess a variety of line quality features, and whether the non-original features impacted on the assessment of overall line quality. These authors surveyed document examiners from a number of countries who provided comparative line quality feature assessments between non-original and corresponding original signature groups. In all, seventy-two genuine and forged signatures were evaluated by the examiner group (one questioned signature and ten specimens per person) These authors found that although not all line quality features were correctly identified by the examiners, this did not result in significant inaccuracies in the overall assessment, as evidenced by an accuracy rate of 95.8%. This study provides and interesting backdrop for the experiment described here.

In our study we aimed to investigate the skill of forensic document examiners in provided opinions regarding the process of production and authorship on both non-original and original signatures. The non-original signatures were second-generation photocopies of the original signatures.

2. Method

2.1 Participants

Six document examiners employed at the Document Section of theNew Zealand Police undertook the study. They provided informed consent for the results to be published, while maintaining anonymity of their results.

2.2 Material studied

The study comprised two trials. Each trial was constructed according to the accepted process of comparing a group of known (specimen) signatures with a group of questioned signatures, where the writer was known to the experimenters but not to the examiners. One trial contained originals of the specimen and questioned signatures and the other comprised photocopies of the same specimen and questioned signatures.

All original writings were made using the same make of blue ball point pens and using the same make of writing material. All writings in the study were performed on a backing-pad often A4 sheets of paper.

2.3 Signatures provided by the specimen writer

The specimen writer was selected from the academic staff at La Trobe University. This writer was provided with all of the materials required to form the specimen material. The specimen writer, each day, was required to write 21 normal signatures, 6 disguised signatures and 6 signatures which might appear to be forgeries (auto-simulations). This was repeated for seven days.

2.4 Construction of the specimen signature group

The specimen group comprised 21 of the normal signatures taken from seven days. These signatures were attached to backing boards (3 to a board) for use in the trial.

2.5 Generation of forged signatures not written by the specimen writer

Two 'forgers' were selected from the academic staff at La Trobe University. These individuals had both been used by the authors as forgers in previous

studies. Each of the forgers were provided with 9 normal signatures from the specimen group described in the previous section. Each of the forger's specimen signature group represented 3 signatures from each of 3 days of specimen writings (forger A's specimen group was taken from the specimen writers' day 1, 3 and 5 signatures, and forger B's specimen group was taken from the specimen writers' day 4, 6 and 7 signatures. Forgers were instructed to produce only 'free-hand' (i.e., not traced) simulations in this trial.

Each day for a seven-day period, the forgers practised simulating the specimen signature 15 times and then performed 12 simulations from which the trial set would be constructed. In all, 105 practices and 84 at tempted simulations were made by each of the forgers over the seven day period.

2.6 Construction of the questioned signature group

The questioned group contained the following types of signature:

- 50 genuine signatures (these comprised ten signatures from days 1 and 7 and six signatures from each of the other five days of writing).
- 168 simulated signatures (84 simulations from each of the two forgers, which comprised all simulation attempts from each of the seven days).
- 21 disguised signatures written by the specimen writer (these were disguised signatures 4, 5 and 6 from each of the seven days).
- 21 auto-simulations (these were auto-simulated signatures 4, 5 and 6 from each of the seven days).

The 260 questioned signatures were given a random number and attached to backing boards (3 to a board).

The boards containing the specimen and questioned signatures were copied on a Canon photocopier onto A4 sheets of paper, which were again photocopied. The photocopied signatures used in the trial were, therefore, second generation copies of their original form.

3. Procedure

The document examiners were initially provided with the photocopies of the specimen and questioned signature groups and with an answer booklet. Ten months later, following the return of the first answer booklet, they were provided with the originals of the specimen and questioned signature groups and with the second answer booklet. For each trial, examiners were informed that the date range over which the specimen material was taken was around the time that the questioned signatures were written. They were then asked to compare each questioned signature independently with the specimen signature group and to express an opinion using the answer booklet provided. The answer booklet comprised 260 lines, each line corresponding to one of the questioned signatures. On each line were the numbers 1 to 7. Each number was a code representing one of the seven possible opinions. For each questioned signature, examiners were required to circle a number that corresponded to their opinion. The answer (opinion) codes (1 to 7) corresponded to the following explanations.

1. There is evidence that the questioned signature was produced using a disguise/simulation process. There is evidence that the questioned signature was written by the writer of the signature specimens.

2. There is evidence that the questioned signature was produced using a disguise/simulation process. There is evidence that the questioned signature was not written by the writer of the signature specimens.

3. There is evidence that the questioned signature was produced using a disguise/simulation process. No opinion can be expressed as to whether or not the writer of the signature specimens wrote the questioned signature.

4. There is evidence that the questioned signature was not produced using a disguise/simulation process. There is evidence that the questioned signature was written by the writer of the signature specimens.

5. There is evidence that the questioned signature was not produced using a disguised/simulation process. No opinion can be expressed as to

whether or not the writer of the signature specimens wrote the questioned signature.

6. No opinion can be expressed as to whether the questioned signature was produced using a disguise/simulation process. There is evidence that the questioned signature was written by the writer of the signature specimens.

7. No opinion can be expressed as to whether the questioned signature was produced using a disguise/simulation process. No opinion can be expressed as to whether or not the writer of the signature specimens wrote the questioned signature.

In addition, on each of the 260 lines of the answer booklet there were the letters 'm' and 'vs'. Examiners were requested that if their opinion was related to identification or elimination (responses 1, 2, 4 or 6), they should indicate the strength of that opinion by circling either 'm' which refers to a moderate strength ('indications') opinion, or 'vs' which refers to a very strong opinion.

The above answers represent the range of opinions that could be expressed by examiners. It is noted that the statement 'There is evidence that the questioned signature was not produced using a disguise/simulation process caused concern amongst some examiners and after discussion was generally taken to mean that 'There is no evidence that the questioned signature was produced using a disguise/simulation process'.

Following completion of the first trial (comprising photocopies), answer booklets were returned to the investigators for analysis. The subjects did not review their answers prior to the undertaking of the second trial (comprising the originals), which they received 10 months after returning the answers to the first trial. They were not provided with any results until all analyses for both trials were finalised.

4. Analysis

Examiners' authorship responses (opinion units) were marked as correct, erroneous or inconclusive. These marks were then analyzed to produce scores for each of the different questioned signature types [genuine, disguised, auto-simulation and simulation (forgery)]. The scores are presented as numbers of opinions or as percentages, which represent opinion rates. The following definitions of the score categories are used in subsequent results tables in this report

 # Correct
The number of authorship opinions that were correct.
 # Error
The number of authorship opinions that were erroneous.
 # Inconclusive
The number of authorship opinions that were inconclusive.
% Correct
The number of correct authorship opinions divided by the total number of authorship opinions (expressed as a percentage).
% Error
The number of erroneous authorship opinions divided by the total number of authorship opinions (expressed as a percentage).
% Inconclusive
The number of inconclusive authorship op1mons divided by the total number of authorship opinions (expressed as a percentage).
% Correct called
The number of correct authorship opinions divided by the sum of the correct and erroneous authorship opinions (expressed as a percentage).
% Error called
The number of erroneous authorship opinions divided by the sum of the correct and erroneous authorship opinions (expressed as a percentage).

The 'called' scores do not include inconclusive opinions and, therefore, equate to a number that reflects the opinion rate when an examiner is expressing an opinion that is other than inconclusive.

Opinions regarding process are ones that relate to whether or not the signatures were considered to be the product of a disguise and/or simulation process. Examiners' process opinions were recorded and analysed. They have been reported in the Results where relevant.

Signature Type	Opinions									
	# correct		#Inc		% Correct		% Inc		% Correct called	
	Phc	Or	Phc	Or	Phc	Or	Phc	Or	Phc	Or
Genuine	147	150	3	0	98	100	2	0	100	100
Simulated	0	0	504	504	0	0	100	100	N/a	N/a
Autosim	0	0	63	63	0	0	100	100	N/a	N/a
Disguised	59	61	4	2	93.7	96.8	6.3	3.2	100	100

Autosim = Auto-simulation
The strength of identification opinions is not shown in this table.

TABLE 1. Scores for examiners pairs' opinions regarding the authorship of photocopied (Phc) and original (Or) signatures for each of the questioned signature types.

5. Results

For each trial, three answer booklets were submitted. These booklets were the agreed opinions from two examiners where a peer review process had been used. The same pairings of examiners were used for each trial. Each pair carried out the trials independently of the other pairs.

The group results for authorship opinions on both original and photocopied signatures are shown in Table 1. There were no errors made by this group for original or photocopied signatures. There were no instances where an identification/elimination opinion was reversed between a photocopy and its original. In fact, no elimination opinions were given. There were only three inconclusive opinions regarding genuine signatures, all on the photocopied signatures. The remaining opinions on genuine signatures were all correct. For all the simulations not written by the specimen writer, an inconclusive opinion regarding authorship was given. In all but two of these, for both original and photocopied simulations, examiners gave opinion code 3 (described in the Method) indicating that there was evidence of the simulation process but they were not prepared to exclude the specimen writer as having made them. In the two other instances, examiners were inconclusive regarding process (one in stance for originals and one for photocopies). The results for Auto simulations were similar. All opinions regarding authorship were inconclusive. However, in all but one of these types of signatures, examiners gave opinion code 3 indicating that there was evidence of the simulation process. The one instance where there was an inconclusive opinion regarding process for auto-simulated signatures concerned a photocopy.

Most authorship opinions relating to disguised signatures were that the writer of the specimens wrote the signatures. This suggests that the disguise process adopted by the specimen writer was not particularly effective. The difference between authorship opinions for original and photocopied signatures for this type of questioned signature, although small, was proportionally greater than for other types of questioned signature. In addition, as described below, half of the differences in opinion between an original signature and its photocopy were in the strength of the opinion relating to authorship for this type of signature (which is not shown in Table 1).

6. Consideration of differences between opinions for original signatures and their photocopies

For a numerical comparison of opinions regarding an original signature and its photocopy, we have used the term coupled opinion unit. A coupled opinion unit is the two opinions expressed by an examiner pair regarding one signature (the original and its photocopy - coupled signatures). Thus there were 780 coupled opinion units expressed by the group (260 signatures per trial by 3 examiner pairs). A coupled opinion unit could be either concordant (where the opinions were the same for the original signature and its photocopy) or discordant (where the

opinions differed for the original and its photocopy).

Three types of discordant, coupled opinions were given by the examiners. They occurred when there was:

- an authorship opinion for one of the coupled signatures and an inconclusive opinion for the other signature.
- a 'very strong' authorship opinion given for one of the coupled signatures and a 'moderate' strength opinion for the other signature.
- an opinion that there was evidence of a simulation process for one of the coupled signatures and an inconclusive opinion regarding process for the other signature.

For the whole group of examiners, the difference in authorship opinions between an original signature and its photocopy was very small. The total number of the three types of discordant opinion units was 18 (2. 3% of the 780 coupled opinion units). Thus 762 of the opinions expressed on the photocopied signatures were the same as the opinions expressed on the originals. For opinions in the direction of identification, when we ignored the strength of the opinions indicated by the examiners, only five of the 780 coupled opinion units were discordant (0. 6%). All five discordant opinions occurred when an examiner pair had given an inconclusive opinion regarding the authorship of a photocopied signature, but gave an opinion that the original signature was written by the writer of the specimens. Three of these signatures were genuine, and two were disguised.

There were 10 discordant opinions that were due to a difference in the strength of authorship opinions. Of these, there were nine instances where an examiner pair had given a moderate opinion that the photocopied signature was written by the writer of the specimens, but gave a very strong opinion that the original signature was written by the writer of the specimens. Eight of these signatures were, in fact, attempts at disguise by the specimen writer, and one was a normal, genuine signature. There was one instance where an examiner pair had given a very strong opinion that the photocopied signature was written by the writer of the specimens, but gave a moderate opinion that the original signature was written by the writer of the specimens. This signature was in fact an attempt at disguise by the specimen writer.

Three of the discordant opinions were related to the process of signature production. One was where, for an auto-simulated signature, an examiner pair had given the opinion that a photocopied signature was the product of a simulation process, but gave an inconclusive opinion regarding the process of production of the original signature. The other two discordant opinions were for the same signature that was, in fact, simulated by someone other than the specimen writer. In one instance, an examiner pair had given the opinion that the photocopied signature was the product of a simulation process, but gave an inconclusive opinion regarding the process of production of the original signature. In the other instance, an examiner pair had given an inconclusive opinion regarding the process of production of the photocopied signature, but gave the opinion that the original signature was the product of a simulation process.

7. Discussion

The results clearly indicate that this group of examiners are able to make comparisons on a complex signature with the same accuracy and similar sensitivity when using either originals or photocopies. No errors regarding authorship were made for original or photocopied signatures, and there were no instances where an identification/elimination opinion was reversed between a photocopy and its original. The high correct rates for questioned genuine signatures were similar for original signatures (100%) and photocopied signatures (98%). While none of the examiners were prepared to eliminate or identify the specimen writer as having written the simulations or auto-simulations, 99.7% of their opinions were that the original and photocopied signatures were produced using a simulation process. The remaining opinions (of which there were three) regarding the process of production of these simulated signatures were inconclusive.

In terms of the Dawson and Lindblom (1998) study, our findings illustrate that when using

photocopies, examiners can translate observaions regarding non-original line quality characteristics and address whether the observed characteristics are consistent with a genuine writing act or with an act of simulation.

The total number of discordant opinions (18 opinions or 2. 3%) was very small. The majority of the discordant opinions (10 of 18) were due to a difference in strength of identification opinions. Eight of these involved signatures where the specimen writer had attempted to disguise her signature, and examiners provided a moderate opinion that the photocopied signature was written by the specimen writer, but a very strong opinion that the original was written by the specimen writer. This seems to suggest that there was information missing in the photocopy that, in the original, provided the examiners with extra confidence regarding their opinion. In addition, there were proportionally more discordant authorship opinions for 'disguised' signatures than for the other questioned signature types.

Although it may be attractive to consider that the small number of discordant opinions expressed by the group is directly attributable to the original/non-original nature of the images, this must be taken in light of the time delay variable. There was a 10-month time difference between when the examiner group submitted their first opinions on the photocopied signatures and their final opinions on the originals of these signatures. It may have been that at least some of the discordance was due to longitudinal inter-examiner opinion variation where the extent to which examiner opinion changes over time is essentially unknown. We feel that the effect of this variable is likely to be negligible due to each opinion unit being the agreed opinion of two examiners.

This study does have certain limitations. The sample size is small and it is not possible to say that the results for this group of examiners are representative of what would be found for document examiners in general. In addition, the results for this group may be different for less complex signatures, for extended text, or for a more limited writing sample. The quality of the photocopy will obviously affect the results.

Despite these limitations, it can be said that this study provides support for the perceptions of those authors (Hilton, 1982; Ellen, 1989; Morton, 1989) who consider that in certain circumstances examiners can express fruitful comparison opinions on non-original writings. This should, however, be considered in relation to the major restrictions when expressing opinions regarding non-original writing highlighted by all of these authors. This was appropriately summarised by Ellen (1989) who wrote "Care must be taken to distinguish between the writing and the document on which it appears to have been written. The writing could be genuine but the document may not. The photocopy could be a composite of two or more documents, and so the writing appears in a context different from that in which it was written" (p. 62-63). It is clear that any opinion expressed regarding the authorship of non-original questioned writing should carry with it some explanation of the limitations imposed on the examination. Huber and Headrick (1999) wrote that "Findings must be so worded ... that they clearly indicate: 1. The identification is of a writing on a document of which the material at hand purports to be a trustworthy reproduction," and "2. The findings are subject to confirmation of their existence as original writings, upon examination of the original document."

References

Dawson, G.A., & Lindblom, B.S. (1998). An evaluation of line quality in photocopied signatures. Science & Justice, 38, 189-194.

Ellen, D. (1989). The Scientific Examination of Documents: Methods and Techniques, Chichester, Ellis Horwood.

Aknowledgements

We acknowledge the following examiners David Boot, Trish James, Gordon Sharfe, John Walker, and Delwynne Walsh, for their participation in this project.

Journal of forensic document examination (Online)
ISSN 0895-0849