# 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)

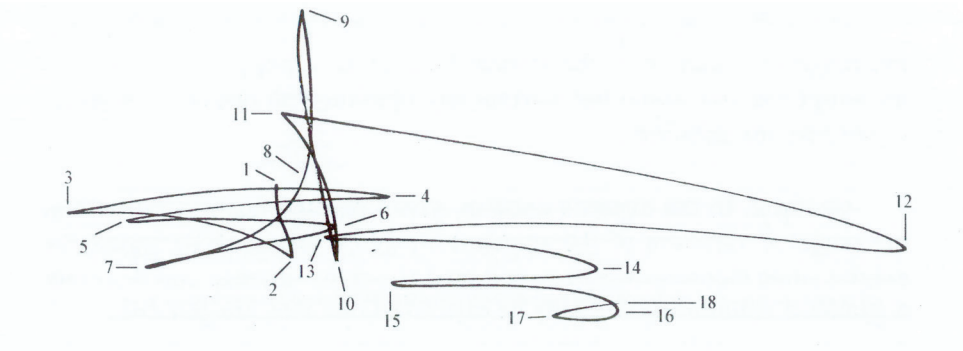


FIGURE 1. Example of a signature illustrating the application of the method used to manually count the number of turning points associated with each signature (TP=18).

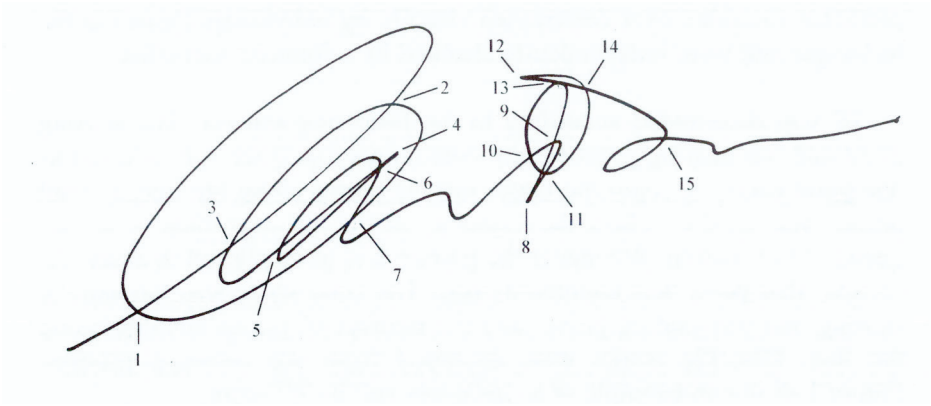


FIGURE 2. Example of a signature illustrating the application of the method used to manually count the number of intersections and retraces associated with each signature (INTRT=15).

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.