5-2018

Assessing the complexity of handwritten signatures

Hal S. Stern
University of California, Irvine

Miriam Angel
Los Angeles Police Department

Melvin Cavanaugh
Los Angeles County Sheriff's Department

Eric L. Lai
University of California, Irvine

Shuying Zhu
University of California, Irvine

Assessing the complexity of handwritten signatures

Abstract

The complexity of a handwritten signature is a key factor in the ability of forensic document examiners (FDEs) to determine whether a questioned signature is genuine or a simulation. Examiners are likely to offer stronger and more reliable conclusions in comparisons involving high complexity signatures. Comparisons of low complexity signatures introduce uncertainty and are likely to result in weaker conclusions or perhaps an opinion that the evidence is inconclusive regarding authorship. It is thus of great interest to explore the reliability and reproducibility of perceptions of signature complexity. This article reports on a study which collected signatures from 123 individuals and obtained subjective assessments of the complexity of the signatures on a three-point scale and a five-point scale from five FDEs. The degree of correspondence of the complexity ratings was evaluated by considering each pair of examiners separately. There was exact agreement on the five-point scale for about 45% of signatures and the examiners differed by more than one point on the five-point scale for approximately 9% of signatures. A statistical model treating the scores as continuous measurements estimates the intraclass correlation, the expected correlation of complexity scores from two different FDEs, at 0.65. Data on intra-rater reproducibility of complexity assessments was obtained by acquiring additional evaluations of complexity on seven of the signatures from the same FDEs. These data suggest the expected intra-examiner correlation of complexity scores aross the two occasions is 0.67, slightly higher than the inter-examiner correlation. These results regarding repeatability and reliability of subjective assessments of complexity provide important information for researchers developing objective measures of complexity to enhance the signature examination process.

Keywords

reliability, repeatability, reproducibility, analysis of variance

Disciplines

Forensic Science and Technology

Comments

This is an article published as Stern, Hal S., Miriam Angel, Melvin Cavanaugh, Shuying Zhu, and Eric L. Lai. "Assessing the complexity of handwritten signatures." *Law, Probability and Risk* 17, no. 2 (2018): 123-132. Posted with permission of CSAFE.

Assessing the Complexity of Handwritten Signatures

Hal S. Stern

*Department of Statistics, University of California, 2216 Bren Hall, Irvine, CA 92697**

Miriam Angel

Forensic Science Division, Los Angeles Police Department,
1800 Paseo Rancho Castilla, Los Angeles, CA 90032

Melvin Cavanaugh

Scientific Services Bureau, Los Angeles County Sheriff's Department,
1800 Paseo Rancho Castilla, Los Angeles, CA 90032

Eric L. Lai

Department of Statistics, University of California, Bren Hall, Irvine, CA 92697

Shuying Zhu

Department of Statistics, University of California, Bren Hall, Irvine, CA 92697

Revised March 12, 2018

Abstract

The complexity of a handwritten signature is a key factor in the ability of forensic document examiners (FDEs) to determine whether a questioned signature is genuine or a simulation. Examiners are likely to offer stronger and more reliable conclusions in comparisons involving high complexity signatures. Comparisons of low complexity signatures introduce uncertainty and are likely to result in weaker conclusions or perhaps an opinion that the evidence is inconclusive regarding authorship. It is thus of great interest to explore the reliability and reproducibility of perceptions of signature complexity. This article reports on a study which collected signatures from one hundred and twenty-three individuals and obtained subjective assessments of the complexity of the signatures on a three-point scale and a five-point scale from five FDEs. The degree of correspondence of the complexity ratings was evaluated by considering each pair of examiners separately. There was exact agreement on the five-point scale for

*Corresponding author. Email: sternh@uci.edu

about 45% of signatures and the examiners differed by more than one point on the five-point scale for approximately 9% of signatures. A statistical model treating the scores as continuous measurements estimates the intraclass correlation, the expected correlation of complexity scores from two different FDEs, at 0.65. Data on intra-rater reproducibility of complexity assessments was obtained by acquiring additional evaluations of complexity on seven of the signatures from the same FDEs. These data suggest the expected intra-examiner correlation of complexity scores aross the two occasions is 0.67, slightly higher than the inter-examiner correlation. These results regarding repeatability and reliability of subjective assessments of complexity provide important information for researchers developing objective measures of complexity to enhance the signature examination process.

Keywords: Reliability, repeatability reproducibility, analysis of variance.

1. Introduction

Forensic document examiners (FDEs) are often called on to compare a writing sample (perhaps a signature) of unknown or disputed authorship with samples of known authorship to determine whether the questioned sample was written by the author of the known samples. Examiners will typically report their conclusion using terminology such as: "It is my opinion that the author of the known samples wrote the questioned sample," "It is highly probable that the author of the known samples wrote the questioned sample," "The author of the known samples probably wrote the questioned sample," "There are indications that the author of the known samples wrote the questioned sample," and similar statements on the negative side of the scale when it appears unlikely that the author of known samples wrote the questioned sample. Examiners may also give an inconclusive opinion where they cannot determine whether the writer of the known samples also wrote the questioned sample. The comparisons and resulting opinions rely on the notion that, given a large enough sample of writing, it is unlikely that two different individuals will produce writing that has the same combination of characteristics such as character construction, spacing, beginning and ending strokes, pen lifts, and size. The complexity theory of handwriting analysis (see, e.g., Dewhurst et al., 2007) suggests that more complex handwriting is considered more difficult to simulate and less likely to match another author by chance. As a result, FDEs are more likely to give a strong opinion when the questioned writing is considered high complexity and a weaker or inconclusive opinion when it is deemed low complexity. Studies exploring the relationship

of handwriting complexity to FDE performance and research investigating automatic ways of assessing complexity (e.g., the studies described below) typically use subjective assessments of complexity as a starting point. The purpose of the present study is to better understand the subjective assessment of complexity in the context of handwritten signatures.

The importance of signature complexity can best be seen in the study of handwriting examiners' expertise published by Sita et al. (2002). Examiners were given test packages comprised of a series of genuine or simulated signatures, along with a set of relevant exemplars, and asked to record for each signature their opinion as to whether the signature was genuine, the signature was simulated, or the evidence was inconclusive. Sita et al. (2002) carried out a number of comparisons of the results, e.g., comparing FDEs with lay people and comparing performances of FDEs on different types of cases. Most relevant for this article is that they used formulas developed by Found and Rogers to estimate complexity (Found and Rogers, 1996; Found et al., 1998) and used the resulting estimates to assess the impact of complexity on performance. They found that FDEs were correct in 95.6% of their called opinions (i.e., excluding inconclusive opinions) for high complexity signatures and 90.8% of their called opinions for medium complexity signatures. There were also many more inconclusive opinions rendered for the medium complexity signatures.

Given the key role played by signature complexity in handwriting analysis, it is natural that many researchers have attempted to develop objective, feature-based methods for assessing complexity. An interesting starting point in the discussion of measures of signature complexity is the work of Brault and Plamondon (1993). They used information obtained dynamically during the signing process (i.e., using the coordinates of the points made in recording the signature) to derive a measure of the difficulty one would have in reproducing the signature. The measure incorporates the length of the various strokes that comprise the signature as well as information regarding the curvature of the strokes. The measure of difficulty was compared in a small study to subjective assessments of complexity as well as rankings of complexity derived from a pattern recognition algorithm and was found to be a promising start. Found and Rogers (1996) and Found et al. (1998) take a statistical approach to assessing signature complexity. They used subjective assessments of complexity on a three-category scale as the target and built a statistical discriminant model based on various features of the signature to classify signatures into complexity groups. Examiners were initially asked to classify signatures as low complexity (too simple to express

an opinion other than inconclusive), medium complexity (would allow only a weak or moderate opinion regarding authorship) and high complexity (would likely allow a strong opinion regarding authorship). Then Found and Rogers defined a number of features of the signature, including total line length, number of turning points, number of line intersections including retraced line sections, and number of pen lifts as potential predictors. Many of these features are related to measures used by Brault and Plamondon (1993). They found that linear disciminants based on the number of turning points and the number of intersections that included retraced line segments best predicted the FDEs' subjective assessments of complexity. The analysis was carried out separately for a number of different examiners with resulting rates of agreement (between the model estimate and the subjective assessment) ranging from 57% to 83%. The data were subsequently pooled to produce a final algorithm which produced approximately 70% agreement on a validation sample. Not surprisingly, the large majority of the inconsistencies in complexity assessment involved classification into neighboring categories. Dewhurst et al. (2007) found that the results of the discriminant model were consistent with FDE judgments in actual casework in the sense that stronger opinions were reported in cases with signatures that were classified as high complexity than in cases with signatures that were classified as lower complexity. Alewijnse et al. (2011) validated the discriminant-based model approach in a new data set. They also gathered data using subjective assessments of complexity on a 0-100 scale and built a multiple regression model that explained those scores using the same kind of features. Angel et al. (2017) used the same signatures studied in this article to relate subjectively assessed complexity to kinematic features measured while the signatures were being written. They considered three types of signatures: text-based (containing easily recognized characters), stylized (lacking easily identifiable characters), and mixed (containing at least one legible character and the rest illegible) signatures and found that subjective complexity assessments differed depending on signature type. The kinematic features, similar to those considered by Brault and Plamondon (1993), explained between 70 and 80% of the variation in the subjective assessments of complexity.

It is clear from the literature reviewed above that: (1) researchers have made good progress in relating subjective assessments of complexity to more objective feature-based characterizations of the signatures and (2) signature complexity is related to FDE performance. There has however been little formal investigation of subjective assessments of complexity and in particular the reliability of such assessments. For the most part, authors have eliminated signatures from their studies when there is too much variation in

the subjective assessments of different examiners. Such variation may in itself be important information about the complexity of the signature. This paper reports on a study that gathered subjective assessments of signature complexity for a set of one hundred twenty-three signatures (the same data used in Angel et al. (2017)) and characterizes the variability, repeatability, and reliability of subjective assessments of complexity.

2. Methods

Signatures. Signatures were obtained from 123 individuals recruited from the Forensic Science Division of the Los Angeles Police Department and the Scientific Services Bureau of the Los Angeles County Sheriff's Department as described by Angel et al. (2017). Signers ranged in age from 21 to 70 years and included 56 males and 67 females. The group included 10 left-handed writers and three signers self-identified as ambidextrous. Each signer provided five signatures in a single sitting. Participants signed in ink on paper placed on top of a digitizer tablet that was connected to a laptop running software that collected dynamic data while the signatures were being written. The use of an inking pen provided original inked signatures that could be used for complexity assessments. The dynamic data were used by Angel et al. (2017) but are not used here and thus not described further.

Subjective assessments of complexity. Five experienced forensic document examiners, a mix of private and government examiners, agreed to provide complexity assessments for the 123 signatures. The FDEs averaged 28.8 years of experience (with a standard deviation of 8.9 years). Examiners were sent a multi-page document containing 300 dpi images of the five signatures from a single writer on each page. The examiners were asked to judge the complexity of signatures on both a three-point scale and a five-point scale where complexity was defined in terms of how difficult it would be to simulate the signature without detection by a forensic document examiner. The three-point scale provided the following options for answering the question of how hard it would be to simulate a particular signature: 1=fairly easy; 2=medium; 3=difficult. The five-point scale provided the following options: 1=easy, 2=fairly easy, 3=medium, 4=difficult, 5=very difficult. Two scales were used because feedback from pilot studies suggested that the three-point scale, which has been used often in the literature, may not provide sufficient options to reflect a rater's complexity perceptions.

All five examiners rated the complexity of the signatures provided by each of the 123 writers, one

assessment per examiner per signer. Thus we have five complexity assessments for each of the 123 signatures. In addition, duplicates of seven of the signature pages were sent to the examiners to evaluate intra-rater reliabaility.

Overview of statistical methods. Basic descriptive statistics (mean, standard deviation, distribution of scores) were calculated for each examiner. The agreement of complexity assessments was studied for each pair of examiners. Measures of similarity include the Pearson correlation of scores, the percent of signatures for which examiners agree, and the percent of signatures for which examiners disagree by more than a single category. A two-factor analysis of variance model was used to estimate the reproducibility / reliability of the subjective assessments across examiners. Reliability is summarized by the intraclass correlation coefficient (Shrout and Fleiss, 1979). Standard errors for the ICC are obtained via the delta method (Casella and Berger, 2002). The number of signatures with duplicate assessments available to provide information about repeatability of intra-examiner assessments is quite small (n=7). We summarize repeatability by pooling correlations that are estimated separately for each of the individual examiners.

3. Results

Descriptive statistics. We initially describe results for the analysis of the 123 signatures, ignoring the duplicates for the seven cases that were assessed twice. We return to the duplicates below. Tables 1 and 2 present summary results of the complexity assessments for each forensic document examiner. On the three-point scale (Table 1), mean scores range from 2.20 to 2.50 and the standard deviations are approximately 0.7. On the five-point scale (Table 2), mean scores range from 3.33 to 3.81 and the standard deviations are approximately 1.1. There is considerable variation in the distribution of scores used. For example, document examiner 1 rarely uses the very difficult category on the five-point scale where document examiner 3 used the category for 33% of the signatures.

Table 1. Summary results of examiner complexity assessments using the three-point scale.

FDE	Pct of scores 1	2	3	Mean	S.D.
1	17	40	43	2.26	0.73
2	15	30	55	2.41	0.73
3	12	25	63	2.50	0.71
4	16	48	36	2.20	0.70
5	19	28	53	2.33	0.79

Table 2. Summary results of examiner complexity assessments using the five-point scale. (Percent of scores may not add to 100 because of rounding.)

FDE	Pct of scores 1	2	3	4	5	Mean	S.D.
1	2	17	34	38	8	3.33	0.94
2	2	16	18	40	24	3.66	1.09
3	2	11	24	29	33	3.81	1.07
4	8	14	22	42	14	3.40	1.14
5	5	15	24	31	24	3.54	1.16

Table 3 presents summaries of pairwise analyses of the complexity assessments. Each pair of examiners is considered separately. For each pair we report the Pearson correlation of the assigned scores, the percent of cases on which the two examiners agree, and the percent of cases on which the two examiners provide scores that differ by more than one category. As reported by others (e.g., see Bollen and Barb, 1981) correlations are lower when there are fewer categories. The median correlation is .62 for the three-point scale and the median correlation is .68 for the five-point scale. Examiners agree exactly 63% of the time on the three-point scale, differ by a single category 33% of the time and differ by more than one category 2% of the time. Precise agreement happens less often (45%) and differences by more than one category occur 9% of the time on the five-point scale.

Table 3. Summary results of pairwise analyses of complexity assessments.

	Three-point scale			Five-point scale		
FDEs	Correlation	%Agree	%Differ> 1	Correlation	% Agree	% Differ> 1
1,2	.75	73	1	.75	50	5
1,3	.52	54	3	.62	33	11
1,4	.67	66	0	.69	50	6
1,5	.64	68	3	.68	53	11
2,3	.61	64	2	.73	51	6
2,4	.66	61	0	.74	46	6
2,5	.63	69	4	.67	48	8
3,4	.55	54	2	.64	34	13
3,5	.58	63	4	.65	44	10
4,5	.61	60	2	.64	43	12

Reliability. A common way to describe agreement among the full set of judges (rather than examining them two at a time) is through the use of the intraclass correlation coefficient (see, e.g., Shrout and Fleiss, 1979). The intraclass correlation can be easily estimated through an analysis of variance that incorporates FDE effects and signature effects. Both FDEs and signatures are modeled as random effects because we are interested in generalizing the results of the study to FDEs and signers beyond those participating in the study. The observed score is assumed to be the sum of an overall mean, a contribution due to the specific signature (some signatures are more complex than others), a contribution due to the FDE (some examiners give higher scores than others on average), and an additional residual term (sometimes referred to as the error term) that accommodates unspecified variation (e.g., variability that might be expected in the score assigned by an individual examiner in repeated trials with the same signer). See the appendix for more details on the statistical modeling. We focus on the results for the five-point scale as the analysis of variance model is most appropriate for continuous data. The results turn out to be similar for the complexity scores on the three-point scale. The estimated variation due to signatures is estimated at 0.79, the estimated variation due to examiners is estimated at 0.036, and the estimated residual variation is estimated at 0.38. These terms allow us to estimate the correlation that would be expected when the same signature is assessed by two examiners. The estimated intraclass correlation is 0.65, and an approximate 95% confidence interval for the intraclass correlation is $(0.58, 0.72)$. Cicchetti

(1994) provides guidelines for interpreting ICC measures, with values below 0.4 considered poor, values between .4 and .59 considered moderate, values between .6 and .75 considered good, and values above .75 considered excellent. The FDE assessments of signature complexity are considered to exhibit good reliability according to that scale.

Repeatability. The LAPD/LASD signature data provides a unique opportunity to assess the repeatability of signature complexity assessments by the same examiner through the seven sets of duplicate assessments. There are 35 instances where the same signature was rated for complexity by the same examiner on more than one occasion (5 examiners provided two sets of complexity scores for 7 signers). It appears based on this small data set that repeated assessments of a signature by the same examiner at two different times exhibit approximately the same amount of variation as assessments of a signature by two different examiners. We report on several analyses that support this claim. First, on the five-point scale, the duplicate assessments match exactly 43% of the time, differ by one category in 51% of the pairs and differ by more than one category in 6% of the cases. These results match those for pairs of examiners reported in Table 3. A second analysis focuses on the observed correlations of each examiner's scores for the seven signers for which we have duplicate measurements. The correlation of the duplicate assessments by the five examiners are respectively .42, .64, .88, .40, .79; these can be pooled using an approach based on the Fisher z transformation (Snedecor and Cochran,1989, Chapter 10) which yields an aggregate estimate of 0.67. This estimate is not very precise because of the limited amount of data (approximate 95% confidence interval is $(.36, .85)$). A third approach fits an analysis of variance model, similar to the one described above, to just the duplicate signatures (7 signers, 5 examiners, 2 repetitions) and uses the results to compute an alternative intra-examiner ICC. (See the appendix for more details.) This yields a lower estimate of the correlation for repeated assessments by an individual (0.57) but again is not very precise (approximate 95% confidence interval is $(.28, .85)$).

Interestingly, our earlier analysis of variance model, that was used above to estimate the inter-examine intraclass correlation coefficient from the examiner assessments of complexity of the 123 signatures (i.e., without the duplicates), can be used to extrapolate what we might expect in the case of duplicate assessments by a single examiner. To make this extrapolation, we note that for repeated measurements by the same examiner we would have the same signature effect, the same FDE effect, and a new contribution from the random variation term, and then we use the estimated variance parameters from the analysis of

variance model to calculate the implied correlation (see the appendix for more details). This calculation suggests that the correlation of repeated measurements made by the same examiner at different times would be 0.68 with an approximate 95% confidence interval of $(.62, .74)$. This is in excellent agreement with the repeatability estimate obtained by pooling individual examiner correlations, .67, and in good agreement with the repeatability estimate obtained from the analysis of variance of the duplicate assessments, .57. In summary, all of the analytic approaches support the finding that repeated assessments by the same examiner exhibit nearly as much variation as repeated assessments by different examiners.

4. Discussion

Complexity of handwriting, here signatures, is believed to be a key determinant of the ability of forensic document examiners to determine whether a questioned signature is genuine or simulated. Much research in the field has used subjective assessments of signature complexity as a starting point for the development of more objective methods of measuring complexity (Found and Rogers, 1996; Found et al., 1998; Alewijnse et al., 2011; Angel et al., 2017). The present study takes a close look at subjective assessments of signature complexity in a data set comprised of complexity assessments of 123 signatures provided by five experienced document examiners. We find that the reliability of such judgments across examiners is good, with the inter-examiner correlation of scores (on a five-point scale) estimated at 0.65 (approximate 95% confidence interval $(.58, .72)$). A similar analysis of the data from the three-point scale (not reported above) yields estimated inter-examiner correlation of scores of 0.61. Alewijnse et al. (2011) asked examiners to rate signature complexity on a scale ranging from 0-100. They report reliability of 0.67 which is in excellent agreement with the findings here. The concordance of the various estimates is reassuring that the results reported here are not specific to these data.

A small set of signatures for which duplicate assessments were provided allows for the study of intra-examiner repeatability of complexity assessments. Interestingly, repeated measurements by the same examiner appear to exhibit almost as much variability as is seen in repeated measurements by different examiners. The estimated repeatability of complexity assessments (intra-examiner) is 0.67, while the estimated reliability of complexity assessments (inter-examiner) is 0.65.

Recent evaluations of current practices regarding forensic evidence (National Research Council, 2009;

President's Council of Advisers on Science and Technology, 2016) have called for increased study of the science underlying forensic comparisons. These reports have recommended research into the repeatability and reliabaility of measurements as well as the validity of the resulting forensic comparisons. In the case of handwritten signatures, it has been found that the complexity is a key factor that should be incorporated in such studies. To date, complexity has been subjectively assessed by document examiners or estimated using statistical models developed to reproduce such subjective assessments. When subjective assessments have been used to develop objective measures of complexity, it has been common to rely on a small number of assessors (typically three), and to remove signatures from consideration when the assessors exhibit too much variation. The current study contributes to the literature in two ways. First, our study demonstrates that subjective assessments of complexity have good, but not excellent, reliability. The estimated reliability from our study can be used to inform decisions about how many examiners ought to be included in future studies of subjective assessments of complexity. For example, the approach of Spearman and Brown (Spearman, 1910; Brown, 1910) enables one to infer the reliability that would be expected for an average assessment by a team of examiners. Second, our study included all signatures (even those where examiners disagreed greatly about complexity) and thus more precisely reflects reliability of assessments across the full population of signers. Methods being developed to provide objective complexity scores should use samples that are representative of the population.

Given the important role that signature complexity plays in the comparison process, it is important that there be continuing research into the measurement of complexity. Objective measures (i.e., those that don't rely on subjective assessment) would be helpful; the recent work of Neumann et al. (2016) provides an example of such an approach in the context of friction ridge examinations. Examining the results of FDE examiner performance studies may aid in the development of measures of complexity. Signatures for which all examiners provide strong (and accurate) opinions regarding authorship are clearly of high enough complexity to enable these opinions; signatures for which there is variability among the opinions expressed by examiners about authorship, or for which many examiners provide inconclusive opinions, could be considered medium or low complexity. Data of this type can help researchers build better complexity measures. The development of reliable, feature-based (and hence objective) measures of complexity and additional studies that relate signature complexity to examiner performance will provide improved scientific support for the examination of handwritten signatures. The present study provides

useful benchmark data regarding the subjective assessments of complexity which play a key role in the process of generating objective measures of complexity.

Acknowledgements

The authors are grateful to the editors of the special issue and reviewers for helpful comments. This research was partially funded through Cooperative Agreement # 70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

Appendix: Assessing reliability through the intraclass correlation coefficient

When the same items, signatures in the case of this article, are assessed by a number of evaluators, a common approach to assessing reliability is through a two-factor analysis of variance of the scores. If Y_{ij} is the complexity score assigned by the jth FDE ($j = 1, \ldots, J = 5$ in the study reported here) to the ith signature ($i = 1, \ldots, I = 123$ in the study reported here), then the complexity scores can be modeled as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where μ denotes the overall mean complexity score, α_i measures the effect of the ith signature (i.e., its tendency to receive higher or lower scores than the average), β_j measures the effect of the jth FDE (i.e., the examiner's tendency to assign higher or lower scores than the average), and ϵ_{ij} models variation not accounted for by the other terms. The latter can include random variation and/or tendencies for an interaction between the FDE effect and a given signature (e.g., an examiner may have a particularly difficult time with a given type of signature). If the scores are considered a continuous measure, then it is common to model each of the contributing factors (except the overall mean) as a random factor with $\alpha_i, i = 1, \ldots, I$ assumed to be independent and normally distributed with mean 0 and variance σ^2_{sign}, $\beta_j, j = 1, \ldots, J$ assumed to be independent and normally distributed with mean 0 and variance σ^2_{fde}, and the ϵ_{ij} assumed to be independent and normally distributed with mean 0 and variance σ^2. Under this model, the estimated correlation of two assessments of the same signature by different examiners (Y_{ij} and $Y_{ij'}$) is

$$ICC(\text{inter} - \text{examiner}) = \text{Correlation}(Y_{ij}, Y_{ij'}) = \sigma^2_{sign}/(\sigma^2_{sign} + \sigma^2_{fde} + \sigma^2).$$

The same model can also be used to estimate the correlation that would be expected in two repeated assessments by the same examiner (Y_{ij} and Y^*_{ij} differing only in that they have different realizations of ϵ, the unaccounted for variation). This correlation is

$$ICC(\text{intra} - \text{examiner}) = \text{Correlation}(Y_{ij}, Y^*_{ij}) = (\sigma^2_{sign} + \sigma^2_{fde})/(\sigma^2_{sign} + \sigma^2_{fde} + \sigma^2).$$

The normal distribution is not a particularly good approximation to the five-point scale used to assess complexity here but the resulting correlation estimates are still informative about the reliability of the measurement process.

In the case where repeated assessments by the same examiner are available, it is possible to augment the model so that observations are denoted Y_{ijk} with $k = 1, \ldots, K$ denoting the repeated assessments of the ith signature by the jth document examiner ($K = 2$ in the study reported here). The additional data allows for the evaluation of the possiblity of an interaction between FDEs and signatures. The model can be modified to $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ where a non-zero value of the interaction term γ_{ij} is an indication of a tendency for the scores of FDE j on signature i to differ from what would be expected based only on the general effects of the signature and the examiner. This could be relevant if examiner assessments varied depending on particular characteristics of a signature. There is no evidence that the interaction term is needed to model the data in the study reported here.

References

Alewijnse, L.C., Van Den Heuvel, C.E., & Stoel, R.D. (2011). Analysis of signature complexity. *Journal of Forensic Document Examiners*, **21**, 37-49.

Angel, M., Caligiuri, M.P., & Cavanaugh, M. (2017). Kinematic models of subjective complexity in handwritten signatures. *Journal of the American Society of Questioned Document Examiners, Inc*, **20**(2), 3-10.

Bollen, K.A. & Barb, K.H. (1981). Pearson's R and coarsely categorized measures. *American Sociological Review*,**46**(2), 232-239.

Brault, J. & Plamondon, R. (1993). A complexity measure of handwritten curves: Modeling of dynamic signature forgery. *IEEE Transactions on Systems, Man and Cybernetics*, **23**(2), 400-413.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, **3**, 296-322.

Casella, G. & Berger, R.L. (2002). *Statistical Inference* 2nd ed. Duxbury/Thomson Learning: Pacific Grove, CA.

Cicchetti, D.V. (1994). Guidlines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, **6**(4), 284-290.

Dewhurst, T., Found, B. & Rogers, D. (2007). The relationship between quantitatively modeled signature complexity level and forensic document examiners' qualitative opinions on casework. *Journal of Forensic Document Examination*, **18**, 21-40.

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

Found, B., Rogers, D., Rowe, V. & Dick, D. (1998). Statistical modelling of experts' perceptions of the ease of signature simulation. *Journal of Forensic Document Examiners*, **11**, 73-99.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press.

Neumann, C., Armstrong, D.E. & Wu, T. (2016). Determination of AFIS "sufficiency" in friction ridge examination. *Forensic Science International*, **263**, 114-125.

President's Council of Advisors on Science and Technology (2016). *Forensic Science in Criminial Courts: Ensuring Scientific Validity of Feature Comparison Methods.* Washington, DC: Executive Office of the President.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rate reliability. *Psychological Bulletin*, **86**(2), 420-428.

Sita, J., Found, B. & Rodger, D.K. (2002). Forensic handwriting examiners' expertise for signature comparison. *Journal of Forensic Sciences*, **47**(5), 1-8.

Snedecor, G.W. & Cochran, W.G. (1989). *Statistical Methods*, 8th edition. Iowa State University Press: Ames, IA.

Spearman, C.C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, **3**, 271-295.