JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, May 2011, Vol. 56, No. 3
doi: 10.1111/j.1556-4029.2011.01713.x
Available online at: onlinelibrary.wiley.com

PAPER

QUESTIONED DOCUMENTS

Christopher P. Saunders,[1] Ph.D.; Linda J. Davis,[2] Ph.D.; and JoAnn Buscaglia,[3] Ph.D.

Using Automated Comparisons to Quantify Handwriting Individuality[*,†,‡]

ABSTRACT: The proposition that writing profiles are unique is considered a key premise underlying forensic handwriting comparisons. An empirical study cannot validate this proposition because of the impossibility of observing sample documents written by every individual. The goal of this paper is to illustrate what can be stated about the individuality of writing profiles using a database of handwriting samples and an automated comparison procedure. In this paper, we provide a strategy for bounding the probability of observing two writers with indistinguishable writing profiles (regardless of the comparison methodology used) with a random match probability that can be estimated statistically. We illustrate computation of this bound using a convenience sample of documents and an automated comparison procedure based on Pearson's chi-squared statistic applied to frequency distributions of letter shapes extracted from handwriting samples. We also show how this bound can be used when designing an empirical study of individuality.

KEYWORDS: forensic science, handwriting, handwriting individuality, writing profiles, random match probabilities, writer verification

Forensic document examiners (FDEs) perform handwriting comparisons for writer identification and verification. One underlying proposition of such comparisons is that "No two people write exactly alike" (1, p. 130); each individual incorporates distinguishable characteristics into his/her handwriting regardless of the particular style or system of handwriting learned originally. If one conceptually views an individual's entire body of natural handwriting as his/her writing profile, then this proposition can be restated as no two individuals have the same writing profile.

This proposition that each individual has a distinct writing profile has been debated (2). In general, the uniqueness of writing profiles cannot be validated empirically. To do so would require access to documents written by each individual in the relevant population and possibly a large number of documents from each individual—sufficient to completely characterize that individual's writing profile, including the natural variation in the individual's handwriting.

One approach to measuring the "*degree*" *of individuality* of writing profiles in a population is to consider the chance of observing two individuals with the same writing profile. The goal of this paper is to illustrate what can be stated empirically about the probability of observing two individuals with indistinguishable writing profiles in a given population using a database of handwriting samples and an automated comparison procedure that is restricted to two decisions: either a "match" decision (two writing samples were written by the same individual) or a "no-match" decision (two writing samples were written by different individuals).

For this study, we utilize an automated comparison procedure because it permits the rapid processing of large quantities of writing samples, not because it "models" the comparison techniques utilized by experienced FDEs. There are significant differences between such automated comparisons and comparisons conducted by FDEs, and because of these fundamental differences, any characteristics of automated comparisons described in this paper cannot be related to comparisons made by FDEs. For example, an automated procedure almost exclusively relies on a limited set of (quantifiable) features that can be extracted from scanned images of writing samples. On the other hand, there are some subjective characteristics and features exploited by FDEs that can be extracted only from the original documents, as well as quantifiable features comparable to those used by automated comparison procedures. Another difference is that an FDE is not limited to "match" (i.e., *identification*) or "no-match" (i.e., *elimination*) decisions. An FDE may reach a *no conclusion* opinion because the writing samples are not sufficient to conclusively determine whether or not the two writing samples were written by the same individual, or an FDE may offer qualified opinions regarding the likelihood that a particular writer prepared the questioned document (3). As mentioned by Morris (1), "For limited amounts of writing, the FDE may not be able to absolutely determine which was written by a particular writer (Crane 1999:39–45)" (p. 131).

[1] Document Forensics Laboratory (MS 1G8), George Mason University, 4400 University Drive, Fairfax, VA 22030.

[2] Department of Statistics (MS 4A7), George Mason University, Fairfax, VA 22030.

[3] FBI Laboratory, Counterterrorism and Forensic Science Research Unit, 2501 Investigation Parkway, Quantico, VA 22135.

[*] Presented at the 60th Annual Meeting of the American Academy of Forensic Sciences, February 18–23, 2008, in Washington, DC.

[†] This is Publication No. 09-13 of the Laboratory Division of the Federal Bureau of Investigation (FBI).

[‡] This work was supported in part under a Contract Award from the Counterterrorism and Forensic Science Research Unit of the Federal Bureau of Investigation's Laboratory Division. Names of commercial manufacturers are provided for information only and inclusion does not imply endorsement by the FBI. Points of view in this document are those of the authors and do not necessarily represent the official position of the FBI or the US Government. Co-author Christopher Saunders was also supported by an Intelligence Community (IC) Postdoctoral Research Fellowship, NGIA HM1582-06-1-2016.

Received 11 May 2009; and in revised form 29 Oct. 2009; accepted 17 April 2010.

An automated comparison procedure that is restricted to "match"/"no-match" decisions will not always produce the correct decision; it is subject to two types of errors:

- *False Match Error*: Two writing samples from different individuals may be declared to "match."
- *False No-Match Error*: Two writing samples from the same individual may be declared "no match."

One characteristic that contributes to such errors is the almost exclusive reliance of an automated procedure on a set of features that can be quantified; it ignores subjective characteristics that can be exploited by an FDE for identification and verification. Purdy (4, p. 71) in his discussion of modern developments in identification of handwriting states:

> The discriminating eye and sharp mind of the FDE are still required to see and correctly interpret physical evidence in handwriting cases. Only through logical reasoning and the application of scientific principles by a qualified expert can the authorship of a contested handwritten document be accurately established.

Another "source" of the errors associated with an automated comparison procedure is the natural variation observed in handwriting: "No one person writes, exactly, the same way twice" (1, p. 130). Each individual demonstrates some natural variation in his/her handwriting from one document to another. Lindblom (5, p. 13) attributes this variation to:

> ... no repeated act is always accomplished with exactly the same results, regardless of whether it is produced by a machine or human effort. For instance, an individual's handwriting is made up of a complexity of habitual patterns that are repeated within a typical range of variation around the master patterns....

Harrison (6) also describes the natural variation in handwriting as resulting from variations around a writer's "master pattern" (p. 299). Other researchers, such as Bulacu and Schomaker (7), have described a writer as a "stochastic generator of ink blobs shapes, or graphemes" (p. 703).

These characterizations of the natural variation in handwriting suggest conceptualizing a writing profile as a probability distribution across documents generated by that individual, rather than as a static characteristic of an individual, such as a fingerprint or DNA. (Although an FDE incorporates the totality of the information into the comparison, he/she may not explicitly construct probability distributions to characterize an individual's writing. The FDE's comparison includes the consistency and quality of the characteristics and the presence or absence of characteristics, not solely the frequency with which they occur.)

Viewing an individual's writing profile as a probability distribution suggests that the errors associated with an automated comparison of writing samples cannot be completely avoided, regardless of the comparison procedure used. One cannot reconstruct an entire probability distribution from a small number of samples from that distribution, but only observe some of its properties—namely how frequently certain characteristics occur. So, any writing samples with significant and similar probability of occurring under two distinct writing profiles will make it improbable that an automated procedure will correctly classify every pair of writing samples on the basis of a finite number of features extracted from each. Also, the stochastic nature of writing profiles implies that the "best" information any automated comparison procedure can provide, even one not limited to "match"/"no-match" decisions, is a likelihood that a specific individual wrote a given document.

Even though an automated comparison procedure cannot be made error free, it nevertheless can tell us something about the individuality of writing profiles via its ability to discriminate among writers within the relevant population. An automated comparison procedure's *discriminating power* can be characterized by its associated *random match probability* (RMP). The RMP of interest in handwriting analysis is the chance of randomly selecting two individuals from some relevant population and then randomly selecting two writing samples, one from each individual's available body of handwriting, that are declared to "match" on the basis of the chosen comparison procedure. The "smaller" the RMP, the "better" the automated comparison procedure is for identification and verification.

The RMP is one measure of the false match error rate of a comparison procedure. Intuitively, the RMP summarizes the result of all possible pairwise comparisons of writing samples known to be from different writers via the ratio of the number of comparisons in which the writing samples are indistinguishable to the total number of comparisons. So, the RMP can be viewed as an average probability—the rate of false match errors averaged over all possible writers that may be compared and all types of writing samples that may be available for comparison. Also, viewed conditionally on the size of the writing samples selected for comparison, it measures the effectiveness of the comparison procedure applied to a particular size of writing sample at distinguishing between individual writers.

The RMP is associated with Stoney's (8) first question: "What is the probability of encountering two corresponding objects (generally)?" (p. 475). Stoney rejects this question in relation to the evaluation of evidence, as it does not depend upon the specific evidence. He recommends against the use in court of values of average probabilities, such as the RMP, because they do not provide a measure of the value of the evidence in a particular case. However, he goes on to say that average probabilities do have relevance in deciding what techniques are best to use routinely across a variety of evidence and as a general measure of the worth of the type of evidence under consideration. Aitken (9, p. 52) similarly characterizes an RMP:

> If background data exist, the task of the forensic scientist in determining the rarity of any items with which he is presented is greatly eased. However, there are situations, as in hair analysis, where such background data do not exist. Some idea of the value of the evidence in these situations may be obtained by considering all possible pair-wise comparisons of items known to be from different sources. The ratio of the number of comparisons in which the items are indistinguishable to the total number of comparisons provides a measure of the value of the evidential process in general, though not of the value of the evidence in a particular case. The ratio is known as an average probability. The use of values of average probabilities in court is not recommended, but they do provide a general measure of the worth of the type of evidence under consideration.

Stoney's (8) fourth question: "What is the probability of encountering a corresponding object, given the crime object?" is the one he concludes to be the "fundamental relevant question in evaluating associative evidence" (p. 477). This probability treats the observed evidence as fixed and considers only the probability of encountering this specific evidence among randomly selected individuals.

Estimation of this probability and its relationship to likelihood ratios are the subject of ongoing research and future publications.

As mentioned previously, there is another type of error associated with automated comparison of writing samples. If two writing samples generated by the same writer are declared to "not match," then a false no-match error has occurred. The *random no-match probability* (RNMP) is one measure of the false no-match error rate of a comparison procedure. It is defined as the probability of randomly selecting a writer from the population and then selecting two writing samples at random from the selected individual's body of handwriting that fail to "match" on the basis of the chosen comparison procedure. Heuristically, the RNMP is related to the "expected" ability of the comparison procedure to correctly identify the writer of a questioned document.

Both the RMP and the RNMP depend upon:

- The automated comparison procedure used, specifically the method used to compare features and the set of features being compared.
- The relevant population of writers (more specifically, writing profiles) generating the writing samples being compared. Some individuals' writing profiles are harder to distinguish between than others.
- The size of the writing samples (measured say by the number of characters) being compared. Comparisons between larger writing samples from different writers should result in false match errors less frequently than when comparing smaller writing samples.

In this paper, we describe how the RMP associated with an automated comparison procedure links the concept of handwriting individuality to empirical studies of a comparison procedure. One approach to validating the concept of handwriting individuality is to measure the "degree" of individuality of writing profiles in a population using the chance of observing two individuals with the same writing profile. As discussed in the next section, this measure of the degree of individuality is bounded above by the RMP associated with any comparison procedure. The RMP can be estimated on the basis of quantifiable features extracted from a collection of writing samples from a sample of writers from the relevant population. Therefore, although an empirical study cannot "prove" the individuality of handwriting, it potentially can be used to show that the chance of observing two individuals with the same writing profile is very small.

This paper is organized as follows. First, we define another useful concept, *biometric individuality*, and relate it to both the chance of observing two individuals with the same writing profile and the RMP associated with a comparison procedure. As discussed in the next section, estimating the biometric individuality will be our ultimate goal because it is the "best" measure of the degree of individuality possible when using a given comparison procedure. Then, we propose an estimator of an upper bound on the biometric individuality. We illustrate the computation of this upper bound using a convenience sample of documents from 98 individuals. We end with an illustration of how the upper bound on biometric individuality can be used to design an empirical study to provide an upper bound on the degree of individuality of writing profiles of a specified size, assuming that writing profiles are indeed unique or at least as rare as the specified upper bound.

Methods

Biometric Individuality

In this paper, we are considering the degree of individuality of writing profiles in a population as measured by the probability of randomly selecting two (different) individuals with the same writing profile.

Jain et al. (10) relate the issue of individuality in biometrics to the question: "What is the probability that the biometric data originating from two different individuals will be sufficiently similar?" (p. 131). Although this question is somewhat vague—what does "sufficiently similar" mean in the context of biometric data—it has been used in fingerprint individuality studies (11), where the individuality of a population is defined as the chance of observing two indistinguishable fingerprints selected from a population at random.

For handwriting analysis, the question posed by Jain et al. (10) suggests defining the *biometric individuality* (of a population with respect to a comparison procedure) as the probability that two (different) randomly selected writers from the population have indistinguishable writing profiles with respect to the comparison procedure being used. Intuitively, two writing profiles being indistinguishable mean that one concludes that the handwriting of two writers looks the "same" after observing their entire body of handwriting. Specifically, two writing profiles being indistinguishable mean that the probability that two randomly selected writing samples "match" is the same whether: (i) the two writing samples are selected from the writing profile of the first individual, (ii) the two writing samples are selected from the writing profile of the second individual, or (iii) one writing sample is selected from each of the two writing profiles.

Biometric individuality is an important concept to study in relation to the individuality of handwriting because it represents the "best" measure of the degree of individuality possible when using a given comparison procedure. If handwriting is indeed unique in the sense that every individual has a unique writing profile, then the probability of randomly selecting two writers with the same writing profile is zero. However, practically, there might be some individuals whose handwriting cannot be distinguished using a given comparison procedure based on a given set of measured features. So, it would be useful to know how rare it is to encounter two writers with "practically" indistinguishable writing profiles, which is what the biometric individuality measures. Also, the biometric individuality is always greater than the probability of randomly selecting two writers with the same writing profile. So, any upper bound on it will also provide an upper bound on the rarity of matching profiles, which we are using as a measure of the degree of individuality of writing profiles in general.

Biometric individuality is also related to the RMP (Fig. 1). Biometric individuality refers to the probability of a random match of

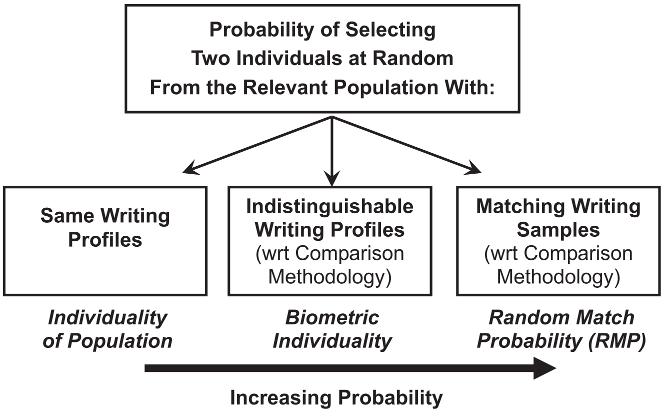
Probability of Selecting
Two Individuals at Random
From the Relevant Population With:
Same Writing Profiles
Indistinguishable Writing Profiles (wrt Comparison Methodology)
Matching Writing Samples (wrt Comparison Methodology)
Individuality of Population
Biometric Individuality
Random Match Probability (RMP)
Increasing Probability

FIG. 1—*Three probabilities related to measuring the individuality of writing profiles.*

writing *profiles*, whereas the RMP refers to the probability of a random match of writing *samples*. So, generally, the RMP is larger than the biometric individuality. However, as the size of the writing samples being compared increases, the RMP gets closer to the biometric individuality. In other words, once randomly selected writing samples are of sufficient size, there is negligible difference between the performance of the comparison procedure applied to distinguishing individual writing samples and individuality of writing profiles.

Therefore, the biometric individuality is "close" to the RMP associated with large writing samples, which can be estimated on the basis of a representative collection of large writing samples from a representative sample of writers. So, if one can identify a comparison procedure with a very, very small associated RMP when applied to large writing samples, such a comparison procedure can be used to provide an upper bound that is close to the true rarity of matching writing profiles. However, even when the writing samples are not "large" and regardless of the comparison procedure used, the RMP still provides an upper bound on the rarity of matching writing profiles in general.

Comparison Procedure

Bolle et al. (12) define a biometric matcher as a system that takes two biometric data samples and returns a score indicating their similarity. A biometric matcher can be used as the basis for comparing two writing samples. A matcher takes (the scanned images of) two writing samples, converts these writing samples to a set of features, and then computes a score based on these features, which measures the similarity of the handwriting in the two writing samples.

With the introduction of a threshold value, a biometric matcher becomes a comparison procedure producing "match"/"no-match" decisions. A pair of writing samples is declared to "match" if the similarity score exceeds the predefined threshold value. Otherwise, a "no-match" decision is declared for the two samples. An alternative to computing similarity scores is to determine distances or dissimilarities between features from two writing samples. Using a distance or dissimilarity score, two writing samples are said to "match" if the associated distance score is equal to or below a predefined threshold value, and "not match" if the distance exceeds the threshold value.

The Document Forensics Laboratory at George Mason University is currently investigating several biometric matchers. One of these matchers, which we will refer to as the *Chi-Squared Classifier*, uses a Pearson's chi-squared similarity score that was specifically designed for verification problems and that can take advantage of knowing what letter is being written, when such information is available. Details of calculating this similarity score associated with comparing any pair of writing samples (from the same writer or different writers) can be found in Saunders et al. (13).

The relative importance of the false match and false no-match errors can be used to choose a threshold value to use with a biometric matcher. For illustration in this paper, we choose the threshold to fix the rate of false no-match errors at some prespecified constant level, say 1%, based on the theoretical properties of the similarity score. The similarity score associated with the Chi-Squared Classifier is related to an approximate *p*-value (13). This implies that the distribution of similarity scores resulting from the comparison of two documents has approximately a uniform distribution when applied to two randomly selected writing samples from the same individual, regardless of the size of the two writing samples being compared. So, theoretically, the 1% RNMP threshold for the Chi-Squared Classifier is 0.01 regardless of the size of the writing samples being compared.

If it is not possible to determine theoretically a threshold based on controlling the level of the RNMP for a specific biometric matcher, one would need to estimate such a threshold. This estimation may be complicated if the biometric matcher responds to the quality and/or amount of information contained in the biometric samples being compared. Furthermore, if the estimation of the threshold to control RNMP is dependent on the RMP, such as when controlling the equal error rate, the standard error of the RMP will need to be incorporated into this estimation strategy.

Estimating RMPs

As discussed previously, the RMP for any comparison procedure provides an upper bound on the rarity of matching writing profiles in general. So, to bound the degree of individuality of writing profiles, one can compute a point estimate and an upper confidence bound for the RMP.

Suppose the data available for estimating the RMP consist of one writing sample per writer (possibly from combining multiple documents) from each of N randomly selected writers from the relevant population. Results of all $N\ (N-1)/2$ pairwise comparisons (i.e., for each pair of writers, compare their associated writing samples and record whether or not the pair "match") can be used to construct both a point estimate and an upper confidence bound for the RMP.

A natural point estimator of the RMP is the number of matches divided by the number of pairwise comparisons, which is $N\ (N-1)/2$. To construct an upper confidence bound for the RMP, one needs to know the standard error of the estimated RMP, which is not just a function of the number of matches among the pairwise comparisons. Further, the standard error is not of the "standard form": population standard deviation divided by the square root of "sample size" because the outcomes of pairwise comparisons are dependent. (The outcome of comparing Document 1 to Document 2 is related to the outcome of comparing Document 1 to Document 3, and so on.) So, our point estimate is based on averaging dependent "observations," instead of the more familiar averaging of independent "observations." However, because this estimator of the RMP is in a class of statistics called *U-statistics of degree 2*, we can use the general form of the standard error of a *U*-statistic (14) to construct a consistent estimator of the standard error of our estimator of the RMP, and thus a (Wald-type) 95% upper confidence bound for the RMP of the form: point estimate plus 1.645 times the standard error estimate (15).

The formula given in Wayman (15) for the upper confidence bound cannot be used when there are zero observed matches. For interval estimation of a proportion, Agresti and Coull (16) illustrate that an adjusted Wald interval obtained after adding two "successes" and two "failures" to the sample yields coverage probabilities close to the nominal confidence levels. We conducted a small simulation study to investigate the coverage probability when a similar type adjustment is made to the Wald interval given in Wayman (15). We added one "match" and one "no match" resulting in the formula for the 95% upper bound in the case of no observed matches to be: $4.65/[(N+1)(N+2)]$. Our preliminary investigations suggest that this adjustment yields coverage probabilities close to the nominal confidence levels.

Results

In this section, we use a set of handwriting samples collected by the FBI Laboratory to illustrate the statistics and probability inequalities discussed earlier. These writing samples, which we refer to as the FBI data set, form a *convenience sample*, not a random sample representative of some relevant population. The FBI data set is used in this study only to illustrate the proposed approach to bounding the probability of observing indistinguishable writing profiles in a given population using a database of handwriting samples and an automated comparison procedure, not to make a statement about the degree of individuality in any specific population.

FBI Data Set

Cursive documents were collected by the FBI Laboratory from 100 volunteers at the FBI, training classes, various forensic conferences, and from friends and family members over a 2-year period. Each volunteer was asked to provide multiple cursive samples of a modified London Business Letter ("London Letter"; [13]). For illustration of the proposed techniques in this paper, we required larger writing samples, so two documents were combined to produce writing samples of roughly 1000 characters for each writer. However, two of the volunteers failed to provide multiple cursive samples, so samples from only 98 volunteers are used in the following computations.

The particular text of the modified "London Letter" was selected because it gives a reasonable representation of the frequencies of lowercase letters in English writing and contains at least one instance of each uppercase letter and of each of the digits 0 through 9. The modifications, which were made by an FDE, consisted of the addition of two sentences at the end of the "London Letter" to incorporate some occurrences of specific letter combinations (e.g., "th," "qu," "ll").

Following is a brief description of how the writing samples were quantified; more details about the processing can be found in Walsh and Gantz (17). Scanned images of the writing samples were segmented into individual characters manually. Then, the pixels in each segmented character were "skeletonized," and an automated process was used to represent each "skeletonized" character by a mathematical graphical isomorphism whose internal structure can be enumerated by a code, referred to as an *isocode*. Mathematical graphs describe the "skeletonized" character in terms of nodes, edges, and their relative connectivity. The isomorphic part of the name refers to the fact that the edges of the graph can be "unbent," reoriented, or resized and leave the pattern of nodes, edges, and their relative connectivity unchanged. This proprietary method of handwriting quantification, developed by Gannon Technologies Group, was originally applied to optical character recognition for handwriting, but has since been shown to provide a powerful foundation for biometric identification using handwriting (17).

The process used to extract features from each of the writing samples ultimately then associates each segmented character in the modified "London Letter" with a letter and an isocode, thus reducing each document to the frequency of isocodes used to write each letter. This allows representing each writing sample as a cross-classified table of letter by isocode (13).

As mentioned previously, any number of quantification systems could be used to study handwriting individuality. Each of these methods will possess a different biometric individuality and in turn provide a different upper bound on the degree of individuality in a population.

Estimated Upper Bound

We applied the procedure described in the last section to the FBI data set to compute an upper bound on the RMP. Using the Chi-Squared Classifier (13) on the samples from 98 writers, there are no matches at the 1% RNMP threshold of 0.01, which implies a 95% upper confidence bound on the RMP of 0.00047.

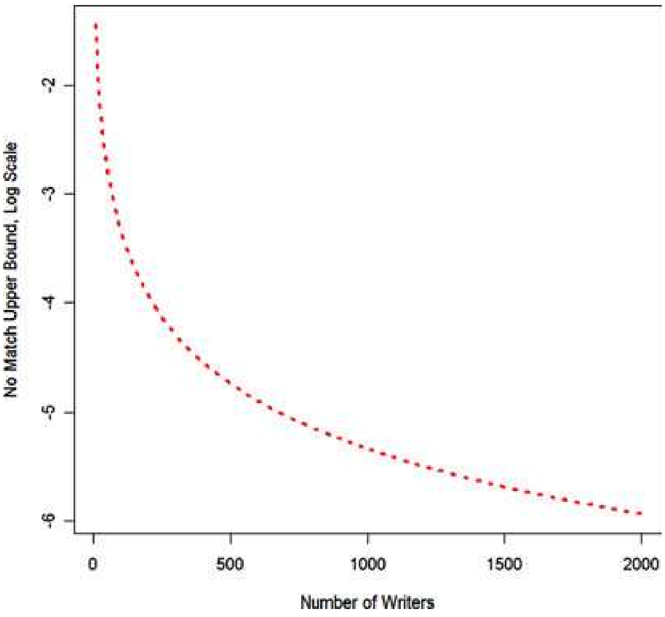

FIG. 2—*Smallest possible 95% upper bound on the RMP, as a function of number of writers.*

As pointed out previously, this upper bound on the RMP is also an upper bound on the biometric individuality associated with the Chi-Squared Classifier, and also an upper bound on the rarity of matching writing profiles. This bound is probably not as small as would be needed to support court testimony of an individualization made by an FDE. Additionally, owing to the previously discussed differences between such automated comparisons and comparisons conducted by FDEs, it does not "measure" the ability of the FDE to identify writers. However, the procedure presented above does suggest that one approach to designing an empirical study of the individuality of handwriting within a specific population is to fix the desired upper bound on the RMP and then select the number of writers in the study to produce the desired upper bound.

The smallest possible upper bound on the RMP occurs when there are no observed matches in a collection of writing samples from a large number of writers. In this case, the larger the number of writers in the study, the smaller the upper bound, as shown in Fig. 2. For example, the calculations graphed in Fig. 2 show that a sample of 2000 writers would yield a 95% upper confidence bound on the RMP on the order of 1 in one million, assuming no observed matches.

This upper bound, however, assumes the ideal scenario that there are no matches observed when the collected writing samples are compared pairwise. Furthermore, for a fixed size of writing samples and fixed RNMP threshold, the probability of observing a match goes up as writing samples from more writers are compared. So, one must also control the chance of observing a match in the $N\ (N-1)/2$ pairwise comparisons. Heuristically, this can be done by choosing an appropriate size for the writing samples (i.e., the number of characters in the writing sample). In other words, there must be a balance between the number of characters in the writing samples and the number of writers providing writing samples in the study.

The relationship between writing sample size and probability of a match is the focus of ongoing research at the George Mason

TABLE 1—*Size of writing sample (i.e., number of characters) needed for specified number of writers and probability of observing no matches.*

	Probability of No Matches At Least:		
Number of Writers (*N*)	50%	80%	95%
50	690	740	810
100	800	850	920
200	910	950	1020
500	1050	1090	1160
1000	1150	1190	1260
2000	1250	1300	1370

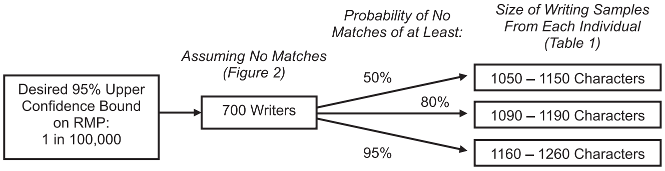
Desired 95% Upper Confidence Bound on RMP: 1 in 100,000
Assuming No Matches (Figure 2)
700 Writers
Probability of No Matches of at Least:
50%
80%
95%
Size of Writing Samples From Each Individual (Table 1)
1050 – 1150 Characters
1090 – 1190 Characters
1160 – 1260 Characters

FIG. 3—*Designing an individuality study.*

University Document Forensics Laboratory. For the Chi-Squared Classifier, Table 1 shows the minimum size of writing samples required for three specific probabilities of observing no matches when comparing writing samples pairwise from each of *N* writers. The values given in Table 1 are based on a very conservative probability inequality (Chebyshev's Inequality [18]) and the assumption that the "true" RMP is very small, if not zero.

As illustrated in Fig. 3, Table 1 and Fig. 2 can be combined to determine the number of writers and size of writing sample needed for a specified upper bound. For example, suppose we would like to obtain a 95% upper bound on the RMP of 1 in 100,000 for a RNMP threshold of 0.01. From Fig. 2, the 95% upper confidence bound for the RMP would be 1 in 100,000 for *c.* 700 writers, assuming no observed matches. Furthermore, from Table 1, we would need to have each individual submit a writing sample with at least the number of characters shown in Fig. 3 for various chances of observing no matches.

Discussion

The proposition that no two individuals have the same writing profile cannot be proven empirically. However, the RMP, which can be investigated empirically, provides both a measure of the degree of individuality relative to a specific comparison procedure of two writing samples (the so-called biometric individuality) as well as an upper bound on the underlying degree of individuality of writing profiles (as measured by the rarity of matching profiles). Therefore, although an empirical study cannot "prove" uniqueness of writing profiles, it potentially can be used to show that the chance of two writers having the same writing profile is very small.

The empirical estimation of a small probability, such as an RMP, in a population where individuals tend to have different (if not ultimately unique) writing profiles, is a difficult problem. The estimation problem is further complicated in the current scenario by the inherent natural variability in writing samples, thus necessitating collecting writing samples with a large number of characters from a large number of individuals. In such a scenario, there are several advantages to basing estimation of the RMP on pairwise comparisons. First, with the current availability of automation, the difficulty in handwriting studies is collecting the writing samples, not comparing two documents. So, all possible comparisons should be utilized to decrease the variability in estimates of the RMP. Second, the RMP estimated from pairwise comparisons is a natural *U*-statistic, and *U*-statistics are well studied. *U*-statistics more efficiently utilize the information from all pairwise comparisons of samples by modeling the dependency structure between individual comparisons. Thus, they can be used to construct an approximate upper confidence bound on the RMP based on using the information in all pairwise comparisons. Also, the upper bound used in this study can be used when there are no pairwise matches, a situation where many of the classical methods fail.

Finally, we have shown that the strength of the results from an empirical study depends on the size of the writing samples used and the number of individuals in the study. While the "true" RMP depends only on the size of the writing samples studied, our ability to estimate it is very dependent on both the size of the writing samples and the number of writers represented in the database. It is important to keep this interplay in mind when developing an empirical investigation of handwriting individuality.

Acknowledgments

The authors would like to acknowledge the help of Donald T. Gantz regarding clarification of the original manuscript, the computational support of Gannon Technologies Group, and the scientists from the FBI Laboratory for their reviews and comments.

References

1. Morris RN. Forensic handwriting identification: fundamental concepts and principles. London, UK: Academic Press, 2000.
2. National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: National Academies Press, 2009.
3. ASTM. Standard terminology for expressing conclusions of forensic document examiners (ASTM E1658). Philadelphia, PA: American Society for Testing and Materials, 2008.
4. Purdy DC. Identification of handwriting. In: Kelly JS, Lindblom BS, editors. Scientific examination of questioned documents, 2nd edn. Boca Raton, FL: CRC/Taylor & Francis, 2006;47–74.
5. Lindblom BS. What is forensic document examination? In: Kelly JS, Lindblom BS, editors. Scientific examination of questioned documents, 2nd edn. Boca Raton, FL: CRC/Taylor & Francis, 2006;9–17.
6. Harrison WR. Suspect documents, their scientific examination. Chicago, IL: Nelson-Hall, 1981.
7. Bulacu M, Schomaker L. Text-independent writer identification and verification using textural and allographic features. IEEE Trans Pattern Anal Mach Intell 2007;29(4):701–17.
8. Stoney DA. Evaluation of associative evidence: choosing the relevant question. J Forensic Sci Soc 1984;24(5):473–82.
9. Aitken CGG. Populations and samples. In: Aitken CGG, Stoney DA, editors. The use of statistics in forensic science. New York, NY: E. Horwood, 1991;51–82.
10. Jain AK, Ross A, Pankanti S. Biometrics: a tool for information security. IEEE Trans Inf Forensic Secur 2006;1(2):125–43.
11. Pankanti S, Prabhaker S, Jain AK. On the individuality of fingerprints. IEEE Trans Pattern Anal Mach Intell 2002;24(8):1010–25.
12. Bolle RM, Connell JH, Pankanti S, Ratha NK, Senior AW. Guide to biometrics. New York, NY: Springer, 2004.
13. Saunders CP, Davis LJ, Lamas AC, Miller JJ, Gantz DT. Construction and evaluation of classifiers for forensic document analysis, http://arxiv.org/abs/1004.0678 (accessed April 5, 2010).
14. Serfling RJ. Approximation theorems of mathematical statistics. New York, NY: Wiley, 1980.
15. Wayman J. Confidence interval and test size estimation for biometric data. In: Wayman J, editor. National biometric center collected works: 1997–2000. San Jose, CA: National Biometric Test Center, 2000; 89–99.

16. Agresti A, Coull BA. Approximate is better than ''exact'' for interval estimation of binomial proportions. Am Stat 1998;52(2):119–26.
17. Walsh MA, Gantz DT. Pictographic matching: a graph-based approach towards a language independent document exploitation platform. In: Lubbes K, Ronthaler M, editors. Proceedings of the 1st ACM workshop on hardcopy document processing; 2004 Nov 12; Washington, DC. New York, NY: Association of Computing Machinery, 2004;53–62.
18. Ross S. A first course in probability, 7th edn. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

Additional information and reprint requests:
Christopher P. Saunders, Ph.D.
Document Forensics Laboratory (MS 1G8)
George Mason University
4400 University Drive
Fairfax, VA 22030
E-mail: csaunde6@gmu.edu