J Forensic Sci, July 2003, Vol. 48, No. 4
Paper ID JFS2002428_484
Published 19 May 2003
Available online at: www.astm.org

Commentary on: Srihari SN, Cha S-H, Arora H, Lee S. Individuality of handwriting. J Forensic Sci 2002; 47(4):856–72.

Sir:

Although the proposition had never been tested in any systematic way, many forensic document examiners have long assumed that all writing is "unique"—that "no two people write alike and no one person writes the same way twice." A recent article by Sargur Srihari and his colleagues (1) reported a study which aimed to prove, for the first time, the truth of at least half of that proposition: that the writing of every person is distinguishable from that of every other person. Srihari et al. drew handwriting exemplars from 1568 individuals, a sample they sought to make representative of the U.S. population. The study comprised two major analyses, each involving about 700 writers. Computer algorithms were used to extract features from scanned images of the handwriting. Attributes of the handwriting were compared at different levels (document, paragraph, word, character) to try to distinguish writers from each other.

As explained below, the study not only failed to achieve its goal but, due to shortcomings of its design, was incapable of achieving the goal. Nevertheless, the Srihari et al. study is an important start to a very different and far more worthwhile line of research.

To begin with, it should be obvious to any reader that, far from proving that all writing in the human universe is unique, even within the four corners of their study Srihari et al. were unable to distinguish each writing from every other writing. Their findings:

> Based on a few macro-features that capture global attributes from a handwritten document and micro-features at the character level from a few characters, we were able to establish with a 98% confidence that the writer can be identified. Taking an approach that the results are statistically inferable over the entire population of the U.S., we were able to validate handwriting individuality with a 95% confidence. By considering finer features, we should be able to make this conclusion with a near 100% confidence.

In other words, the authors were able to accurately declare writings to have come from the same or different writers 98% of the time; they estimated that extrapolated to the entire population they could do so 95% of the time, and they thought that eventually they could do so "near" 100% of the time.

Thus, the study itself did not demonstrate unique individuality, and the authors did not expect to be able to do so with complete reliability in the future. At most, the study supports the conclusion that "most" or "the great majority" of writers are distinguishable from other writers. But that is not the same as uniqueness. In the quest for proof of a claim so extreme and absolute and essential to individualization as uniqueness, "near" is not enough.

Though the data were unable to carry the study to its stated goal ("establishing the individuality of handwriting"), Srihari et al. did not consider the whistle to have blown the ball dead where the data stopped. In the article's Conclusion the authors pick the ball up and carry it across the goal line anyway, simply by speculating: "[T]he objective analysis that was done should provide the basis for the conclusion of individuality when the human analyst is measuring the finer features by hand." But such speculation assumes human examiners add value to the computerized results, a proposition that is far from obvious, as we shall see.

Suppose the Srihari et al. study had done better and *had* been able to distinguish each writer from every other writer in its sample. Would that prove uniqueness for the entire universe of writings? It still would not have succeeded—if for no other reason, because of a series of design choices, each of which made the study a weaker test of the hypothesis of "uniqueness" than it might have been.

First, consider the structure of the writer sample. Srihari et al. obtained handwriting exemplars from 1568 individuals with the aim of obtaining a sample that was "as representative of the U.S. population as possible." That the sampling plan had no hope of achieving that goal is obvious from the description of the sampling: the researchers ". . . obtained samples by contacting schools in three states (Alaska, Arizona, and New York) and communities in three states (Florida, New York, and Texas) through churches and other organizations." In their article, Table 2 confirms that they missed their statistical benchmarks by a wide margin.

The goal of broad representativeness, however, was wrongheaded to begin with, given the purposes of the study. Maximizing writer diversity to the extent of the U.S. population would have *reduced* its ability to convincingly show distinguishability among writers. The study inadvertently sought to exploit regional, cultural, educational, and whatever other group differences exist. Put in more forensically familiar terms, the study inadvertently emphasized class characteristics when it should have been testing for the existence of individualizing differences within classes. By seeking to maximize the diversity of the sample of 1568, the researchers made the task of distinguishing writers misleadingly easy.

An analogy to eyewitness lineups makes this clear. Are we more convinced of the identification accuracy of a witness when a lineup contains foils who are tall and short, fat and thin, hairy and cleanshaven, light and dark complected? Or by an identification from a lineup where the suspect and the foils all look quite similar to each other? By analogy, a study of handwriting individuality would be far more convincing if the writers in a sample had all grown up in the same neighborhood, gone to the same school, and been taught to write by the same teachers.

A better sampling design for the purposes of a study of this kind would have been to gather a representative sample of *clusters* of writers from around the country, with each cluster composed of highly *similar* writers. That would have tested the degree to which highly similar writers can be distinguished from each other, and would have replicated that finding among numerous groups of such writers.

Second, consider the size of the writer sample and the problem of extrapolation to the universe of writers. Adequacy of sample size depends on the research question and the nature of the phenomenon under examination. For true randomized experiments, samples of 10 or 20 per condition can be sufficient. For national studies of public opinion, requiring confidence intervals of plus or minus a few percent, sample sizes of 1500 typically are adequate. For epidemiological research using prospective designs, sample sizes in the tens of thousands often are necessary to detect the hypothesized relationships. What sample size is necessary for a study aimed at establishing even the near unique individuality of all writers in the U.S.?

The smaller the sample, the less likely it is that indistinguishably similar handwriting, if it exists, can be found. As the size of the sample increases, the chances of encountering writing which is indistinguishably alike increases. For example, if we can distinguish the writing of every person in a seminar of ten students, that 100% accuracy would prove little about handwriting distinguishability or individuality. If we could accurately distinguish every one of ten thousand writers from each other, that would be more impressive. A million, even better. But, of course, even that would tell us only about variability, and would confirm a shrinking probability of making false positive errors due to coincidental matches. But Srihari et al.'s actual findings show, as we would expect, that the accuracy of drawing distinctions among writings drops as the data are extrapolated from sample to population.

The logic of drawing inferences about populations from samples, and the role that sample size plays in drawing those inferences, is well understood by statisticians and taught in every undergraduate statistics course. But the statistics of that process is inapplicable to the problem of trying to infer the existence of unique individuality among every member of the universe. In the context of a claim of unique individuality, the path from sample to universe remains uncharted. Interestingly, Srihari et al. are wise enough to note of their sampling that they could "not know it was a perfect sample without measuring the whole population." The same is almost certainly true for determining whether a population fits precisely the model supposed by the hypothesis of uniqueness: it is hard to imagine how one can conclude with certainty that every member of a population is different from every other member "without measuring the whole population." Indeed, the question of unique individuality, by its very nature, is undermined by reliance on statistical inference. If we are willing to settle for probability calculations, then we have abandoned the claim of unique individuality in favor of a claim concerning low probability of coincidental matches. That might be a very sensible move to make (2). But it is not the question that Srihari et al. posed for themselves, and it is not the question they purport to answer.

Third, consider the size of the *writing* sample. Just as the size of the sample of writers affects inferences about the ability to distinguish among them, so does the amount of handwriting sampled. In Srihari et al., the writing exemplars consisted of a specially created 156 word exercise designed to include all letters, numbers and punctuation in order to collect a wide range of writing, including distinctions between capitals and small letters, at the beginning and end of words, and involving certain combinations of particular interest. Again, this artificially maximizes the ability to distinguish writers. Were the writing sample to decrease in the amount of writing diversity (and begin to approximate more forensically typical writing), the ability to distinguish writers would decline further.

Fourth, consider the size of the intra-writer sample. Each writer was asked to provide three exemplars of the writing exercise in order to provide an estimate of the intra-writer variation. In the hypothetical n-dimensional space of handwriting attributes, the range of the cluster of points representing each writer's natural variation will expand as the number of intra-writer samples increases and as the circumstances of the occasion of each writer's giving of the sample increase, so that the intra-writer distribution of one writer is more likely to overlap with the clusters of other writers. In short, had the number of intra-writer samples increased, the risk of mistaking one person's writing for that of someone else would also have increased.

Fifth, consider the type of writing. The writing exemplars used by Srihari et al. apparently consisted entirely, or nearly entirely, of cursive writing. As the type of writing changes from a (lengthy) cursive document to (a few) printed words or letters or numbers, or a signature, the potential for distinguishing among writings of each type is likely to change. Forensic document examiners disagree among themselves, for example, about whether it is harder or easier to identify the authors of hand printing compared to cursive writing (3). No systematic empirical research exists to resolve that controversy. Thus, the findings of Srihari et al., whatever they are, cannot be generalized to hand printing or numbers or signatures. Separate studies would need to be done for those. In an important sense, the issue is not about the population of writers but populations of writings.

Finally, the Srihari et al. study involved no humans examiners. This is important for several reasons. Even were it shown that a computer algorithm looking at the attributes it was programmed to look at could distinguish every writing on earth from every other writing, the forensically relevant question is how well human examiners can make distinctions among the same writings. We know from numerous studies that human examiners sometimes cannot in fact distinguish one person's writing from another (e.g., 4–6).

Srihari et al. suggest that human examiners should be able to do better than their project's computer programs because human examiners look at more and finer features. This is a speculation which assumes that more information will lead to more accuracy for heuristic humans in the same way that it does for algorithmic computers. Findings from cognitive science show a different picture: as humans are given more information the precision of their judgments increases, but only up to a point. After that point they become overloaded and their decision-making deteriorates. In other words, while the relationship between information and accuracy is positive and monotonic for a mathematical model in a computer, the relationship is an inverted-U function for a human. Related research in cognitive science sought to develop quantitative systems that would establish floors of statistical accuracy below which human decisionmakers could not fall. But the floors turned out to be ceilings: humans could not perform complex but disciplined decision-making tasks as accurately as computers did. Humans have limited cognitive capacity which prevents them from processing much of the richness of the available information. Instead, they rely inconsistently on a small number of factors to which they apply nonoptimal weights, and then employ cognitive shortcuts (e.g., 7–9).

In summary the Srihari et al. study not only failed to demonstrate that every writer in its sample could be distinguished from every other writer in its sample, it failed under conditions highly favorable to its success. Had the study employed smaller and more forensically typical writing samples of more difficult types of writing, and drawn from a larger and more appropriately structured sample of writers, the failure would have been even more pronounced.

But the real value of the line of work begun by Srihari et al. will not be to pursue an empirical proof of uniqueness with greater rigor. The Srihari et al. research has something far more valuable to offer. Srihari et al.'s research begins to lay the foundation for a modern and scientifically defensible system of handwriting comparison. Such a computer-assisted system of a handwriting comparison would perform more consistently and predictably than human examiners, using a finite set of objective (observable, measurable, definable) procedures, and would facilitate the computation of the probability of error in the resulting identification or exclusion. That would be an extremely important direction in which to take future stages of this research.

References

1. Srihari SN, Cha S-H, Arora H, Lee S. Individuality of handwriting. J Forensic Sci 2002;47:856–72.
2. Saks MJ, Koehler JJ. What DNA "Fingerprinting" can teach the law about the rest of forensic science. Cardozo Law Review 1991;13: 361–72.
3. Conway JVP. The identification of handprinting. J Crim Law Criminology Police Sci 1955;45:605–12.
4. Harris J. How much do people write alike: a study of signatures. J Crim Law Criminology 1958;48:647–51.
5. Sita J, Found B, Rogers D. Forensic handwriting examiners' expertise for signature comparison. J Forensic Sci 2002;47(6):1117–24.
6. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. J Forensic Sci 1997;42(5):778–86.
7. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974;185:1123–31.
8. Bodenhausen GV, Lichtenstein M. Social stereotypes and information-processing strategies: the impact of task complexity. J Personality Social Psych 1987;52:871–80.
9. Meehl P. Clinical Versus Statistical Prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press, 1954.

Michael J. Saks, Ph.D.
Professor of Law
Professor of Psychology
Arizona State University
Tempe, AZ 85287-7906