JOURNAL OF FORENSIC SCIENCES

*J Forensic Sci*, 2017
doi: 10.1111/1556-4029.13640
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

*Jeremy R. Hadler*,[1] *Ph.D.; and Max D. Morris*,[1] *Ph.D.*

# An Improved Version of a Tool Mark Comparison Algorithm

**ABSTRACT:** Chumbley et al. (2010) described a statistically based algorithm for comparing pairs of tool marks. They presented empirical evidence that the algorithm produces well-separated similarity score values for "matching" and "non-matching" pairs of tool marks. However, the algorithm has two substantial weaknesses. First, it is "uncalibrated" in the sense that error rates can be determined only through empirical investigation. Second, it relies on a randomized test and can lead to different similarity scores when the algorithm is repeatedly applied to the same pair of tool marks. We present an improved version of the procedure, which eliminates the randomized scores and yields more consistent and predictable error rate control. This is accomplished by replacement of a random sampling step from the original algorithm with a deterministic process. We demonstrate the improved algorithm and compare its performance to the original by applying to known "matching" and "non-matching" pairs of tool marks.

**KEYWORDS:** forensic science, striae, profilometry, *U*-statistic, correlated data, screwdriver

Our subject is the comparison of two linearly striated tool marks, such as those left by a screwdriver on the surface of softer metal, with the goal of determining whether they were produced by a common tool. Expert forensic examiners perform such comparisons visually in the laboratory, aided by instruments such as a comparison microscope that allows side-by-side viewing of striae in different specimen orientations. While there is considerable empirical evidence that experienced examiners can effectively identify pairs of "*matching*" and "*non-matching*" striated tool marks in many circumstances (e.g., (1)), the legal imperative to develop analytical methods for which error rates can be determined, at least in principle, has led to research in automated comparison methods applied to digitized versions of the tool marks, for example digital images or representations produced by profilometry. (Somewhat more specifically, we use the term "match" to denote a pair of tool marks that were produced using the same tool, and "non-match" to denote a pair of tool marks that were produced by different tools.) In a recent report by the President's Council of Advisors on Science and Technology (2), encouragement was given to research in directions that would "convert firearms analysis from a subjective method to an objective method," with specific emphasis on algorithm development. One such algorithmic method was described by Chumbley et al. (3), who described an algorithm that quantifies the similarity of two tool marks using the Mann–Whitney *U*-statistic (4), a popular nonparametric statistical index.

In the context of forensic analysis, the procedure in (3) has two appealing properties:

1 Motivated by the process of visual tool mark comparison, the algorithm first identifies the segments of the two tool marks that "best match" or are most similar according to a correlation measure. The tool marks are aligned or registered according to these "best match" segments, and additional pairs of tool mark segments which *should* then physically correspond, are also compared to determine whether the apparent match is consistent over a greater portion of the tool mark. These two steps are called *optimization* and *validation*, respectively, by (3).

2 The overall index of similarity for the two tool marks is a statistic based only on the similarity of segments compared in the validation step. In comparison, a procedure that focuses on the degree of similarity between "best match" segments would be very difficult to calibrate, because some pairs of segments can easily be very similar simply by chance, even when the two tool marks are actually nonmatching. The *U*-statistic, applied to comparisons made only in the validation step, is nonparametric (e.g., does not depend on any specific assumptions about statistical distributions), and under some conditions, can be calibrated (i.e., critical values determined for formal statistical hypothesis testing with controlled error rates) without "tuning" the procedure for the specific characteristics of the data.

In (3), this method is tested in tightly controlled laboratory studies, using tool marks produced with sequentially manufactured screwdrivers. They showed that their method did indeed effectively separate pairs of tool marks produced with a common screwdriver, from pairs produced using different screwdrivers. By "effectively separate," we mean the *U*-statistics for the matching pairs were consistently larger than those produced by the nonmatching pairs. However, the critical values that are suggested by simple probability assumptions do not always separate the observed *U*-statistics, that is calibration of the method is not "automatic" due to statistical dependencies among the comparisons made in the validation step, substantially complicating the

[1] Department of Statistics, Iowa State University, Ames, IA 50010.

Received 7 April 2017; and in revised form 29 July 2017; accepted 17 Aug. 2017.

process of determining error rates. In addition, this algorithm depends on a *randomized* sampling process for selecting the tool mark segments compared in the validation step. This process has the unfortunate side-effect of leading to results that are not identical when two tool marks are repeatedly compared. That is, in some cases, the procedure could indicate that a pair of marks match, but on repeated application could indicate that the same pair is a nonmatch.

In this paper, we propose a modification to the methodology in (3) that does not depend on a randomized test and that reduces the lack of independence between the segments selected in the validation step. This is accomplished by completing the validation step in a deterministic (nonrandom) manner rather than the original randomized process and comparing the *U*-statistics to critical values computed from the standard normal distribution, after a tool mark normalization procedure. We finish by verifying that our proposed method results in a reference distribution that can be considered standard normal while, in general, the original method (3) does not. First we briefly discuss some of the details of the original algorithm to give context to our proposed modification.

## Method of Chumbley et al.: a Review

Here, we briefly describe the algorithm introduced in (3) for comparing two digitized tool marks, which take the form of a sequence of numerical values representing depth of the marked surface at each of a linear set of equally spaced "pixel" locations. (For our work with screwdriver marks, these digitized tool marks are generally several thousand pixels in length.) Given these digitized versions of the tool marks, the algorithm proceeds through *optimization* and *validation* steps.

### Optimization Step

Given two digitized tool marks, $x_1(s); s = 1, 2, \ldots, S$ and $x_2(t); t = 1, 2, \ldots, T$ indexed by pixel locations $s$ and $t$, where $S$ need not equal $T$, the basic measure of "similarity" used by (3) is the ordinary (Pearson product-moment) correlation (or cross-correlation) between segments of $n$ consecutive pixels taken from $x_1$ and $x_2$, where $n$ is a user-specified parameter of the algorithm. $n$ is typically much smaller than the tool mark lengths—generally on the order of 10% of $S$ and $T$—so that these comparisons are made between relatively small "windows" taken from the two tool marks. Denote by $\rho_n(s,t)$, the correlation between the data from $n$ consecutive pixels beginning at pixel $s$ in tool mark $x_1$ and pixel $t$ in $x_2$, that is $\rho_n(s,t) = \mathrm{corr}(x_1(s,\ldots,s+n-1), x_2(t,\ldots,t+n-1))$. The optimization step consists of computing the correlation for all pairs of length-$n$ segments to find the pair that are the "best match" by this criterion; that is $(s^*,t^*)$ is determined such that $(s^*,t^*) = \underset{s,t}{\mathrm{argmax}}\, \rho_n(s,t)$.

The aim of this step is to empirically find the most likely physical *alignment* of the two tool marks, if they were indeed made by the same tool. That is, aligning pixel $s^*$ from $x_1$ with pixel $t^*$ from $x_2$ mimics the process of alignment of specimens under a comparison microscope in such a way that two tool marks appear most similar at this point. However, even with tool marks that do not have a common source, these best-match segments for tool marks generally produce a large correlation by chance, simply because so many (on the order of $S \times T$) correlations are computed. As a result, the overall index of the likelihood of a match is not the maximized correlation itself, but a function of correlations computed in the *validation* step that relies on the alignment or registration just identified.

### Validation Step

The logic of the validation step is the following: If the two marks exhibit a large maximized correlation at $(s^*,t^*)$ because they really were made by the same tool, then shifting to segments on both tool marks an equal distance and direction away from the segments of maximized correlation should still result in a relatively large correlation. If the two marks exhibited a large maximum correlation by chance, then in general such a shift should not result in a relatively large correlation value (again, except by chance). Following the analogy to visual comparison using a comparison microscope, this step is intended to mimic the examination of agreement between tool marks over a wider span of length, once they have been aligned in the apparently best orientation.

To apply this idea, a series of fixed length segments are chosen at random but equal distances and direction away from the optimal segment in each mark, and the correlation is recomputed for each of these segment pairs. These are referred to as the set of 'same-shift' correlations. Additionally, a second set of correlations is computed for segments a random distance and direction away from the optimal segments, but chosen independently for each of the marks, creating a set of "different-shift" correlations. The assumption is that if the tool marks really are a match, the same-shift set of correlations correspond to physically equivalent shifts along the two marks, and so should be systematically larger than those in the different-shift set (for which there should be no physical correspondence in any case). Same-shift and different-shift pairs are depicted graphically in Fig. 1 as recreated from (5).

More specifically, the validation step compares the set of correlations $\rho_m(s^*+\delta, t^*+\delta)$ for randomly selected $\delta$, to the set of correlations $\rho_m(s^*+\delta_1, t^*+\delta_2)$ for independent, randomly selected values of both $\delta_1$ and $\delta_2$. Note that the subscript here is now m denoting that these segments may be a different length than the segments used in the optimization step ($m$ is typically less than $n$). If the two marks were made by the same tool then the set of $\rho_m(s^*+\delta, t^*+\delta)$ should be systematically larger than the set of $\rho_m(s^*+\delta_1, t^*+\delta_2)$. To determine if these values are in fact significantly larger, a Mann–Whitney *U*-statistic is applied to the two sets of correlations computed in the validation step. If the resulting *U*-statistic is large, we reject the null

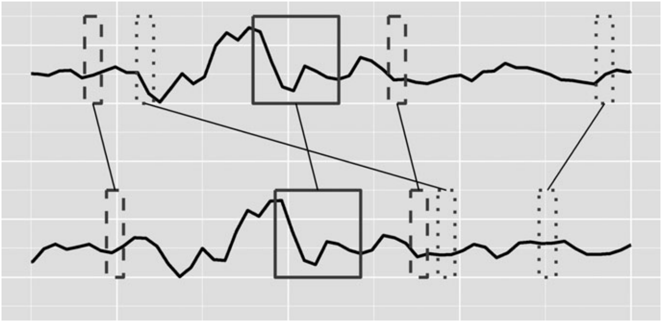

FIG. 1—*The pair of windows selected in the optimization step of the Chumbley algorithm are represented by the large solid rectangles. Examples of same-shift window pairs are given by the smaller dashed rectangles and different-shift window pairs are given by the smaller dotted rectangles.*

hypothesis that the tool marks were not made by the same tool and conclude that the tool marks were made by the same tool. The Mann–Whitney procedure is used because it is nonparametric, that is, it does not require the two samples of correlations to be drawn from normal (or any other specific) distributions. Under the assumption that each value in the two samples is *statistically independent* of the others, the asymptotic distribution of the *U*-statistic is known, and is not dependent on the underlying distributions of the two samples. This allows specification of a *critical value* for this statistic from first principles, that is, a value above which the null hypothesis of nonmatch can be rejected with a predetermined false-positive error rate.

## Proposed Method

As just noted, the asymptotic (limiting large-sample) distribution of a *U*-statistic is known under the ideal statistical assumption of independence. Unfortunately, independence between computed correlations cannot be guaranteed in this application because (1.) there is some degree of serial correlation among surface depth values recorded for a tool mark, and (2.) when segments are chosen randomly (in both the same-shift and different-shift samples), some groups of pixels are often chosen more than once, leading to a further lack of independence between computed correlations involving common data values. Empirical studies have shown that the U-statistics computed for nonmatching tool marks as per the original method in (3) do not have the theoretical distribution that would follow if the data were independent. That is, while the method effectively separates otherwise similar matching and nonmatching pairs of tool marks, no universal "first principles" critical value can be determined for hypothesis testing for comparing a single pair of marks.

### *Tool Mark Normalization*

Our proposed alteration of this method differs only in the validation step; that is, the optimization step is carried out exactly as described in the last section. However, prior to application of the validation step, we apply a tool mark normalization procedure. A number of uncontrolled factors can influence a tool mark, including the amount and direction of applied force (6). Also, some aspects of laboratory analysis, such as specimen fixturing, can contribute variability to the recorded digitized marks. To account for (remove) at least some of this nuisance variation, we apply a coarse lowess smooth (7) to the instrument recording and treat the residuals of this smooth as the normalized version of the tool mark (see also (8)). The effect of this is to remove very long-scale trends in each recording, as might be expected when the fixtured specimen is not perfectly level relative to the profilometer's reference plane. The result of this procedure is shown in Fig. 2.

### *Validation Step*

As with the original method, our proposed validation step involves the selection of same-shift and different-shift samples, but these are now selected by a deterministic rule rather than through random sampling. By explicitly specifying which segments we sample, we control the amount of overlap in sampled segments, reducing the largest source of dependence among computed correlations. In addition, because the sampling is deterministic, the randomized nature of the analysis is eliminated so that repeated analysis of the same pair of digitized tool marks (for fixed values of *n* and *m*) always leads to the same conclusion.

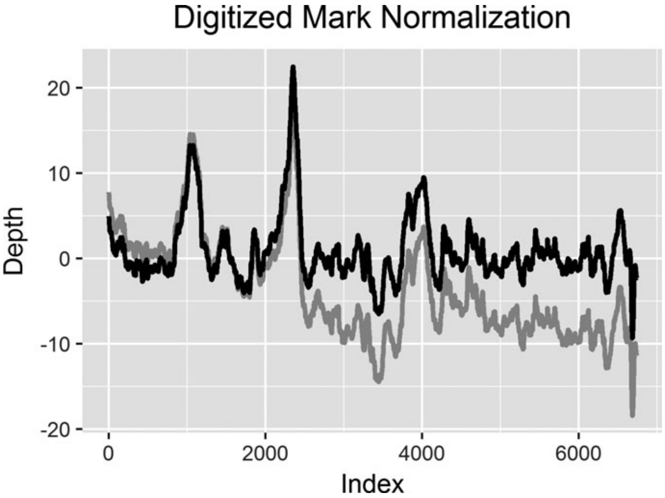

FIG. 2—*A digitized tool mark (gray) and its normalized version (black) obtained as the residuals of a coarse lowess smooth applied to the digitized mark using 25% of the data for the smooth at each location.*

More specifically, our proposition is to choose the same-shift validation window locations in such a way that $t - s = t^* - s^*$ for each computed correlation (as is true for the original method), but under the restriction that there are no pixels in common to any pair of computed correlations by separating *s*-values and *t*-values in increments of *m*. We require our different-shift set of correlations to be indexed by pixel pairs of form $(s^* + \delta, t^* - \delta)$ for positive or negative integer values of δ, again by increments of m in the value of δ so that no pair of correlations involve the same pixels. As in the original algorithm (3), neither of these sets of correlations will be allowed to include any pixel locations contained in the best-match segments identified in the optimization step.

To illustrate this change in segment selection, the correlations, $\rho_m(s, t)$, can be collected into a matrix

$$\begin{bmatrix} \rho_m(1,1) & \rho_m(2,1) & \cdots & \rho_m(S-m+1,1) \\ \rho_m(1,2) & \rho_m(2,2) & \cdots & \rho_m(S-m+1,2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_m(1,T-m+1) & \rho_m(2,T-m+1) & \cdots & \rho_m(S-m+1,T-m+1) \end{bmatrix} \tag{1}$$

and this matrix can be plotted as an image. This is represented in Fig. 3 where an example of the original same- and different-shift sets of correlations (randomly selected) are shown on the left, and our proposed sets of correlations are shown on the right. To be clear, the symbols in these plots denote the lower endpoints for the windows used in the correlation computations and not the entire windows. It should also be noted that occasionally $(s^*, t^*)$ selected in the optimization step will occur in one of the corners of the matrix of correlations, preventing the construction of either the same- or different-shift correlations. However, if $(s^*, t^*)$ occurs in the lower-left or upper-right corner of the matrix (upper left or lower right in Fig. 3) this is most likely an indication that the tool marks were not made by the same tool as the marks have been shifted to opposite ends of each other.

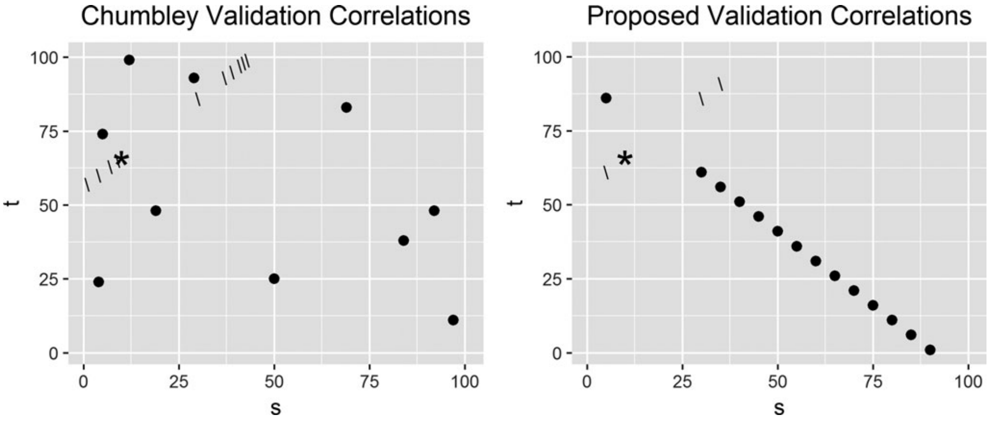

FIG. 3—*In these graphics, the* s*-axis corresponds to columns of Matrix 1, the* t*-axis to its rows, and the entries at each (*s,t*) correspond to the correlation between the two segments starting at* s *and* t*. In both of these images, the * represents the optimal correlation as determined by the optimization step. The image on the left depicts the selection of correlations for the validation step in (3). The same-shift correlations represented by the symbols running from lower left to upper right are randomly chosen along the same diagonal as (*s*,t**) so that there may be pixels common to multiple correlations. The different-shift correlations are randomly chosen from anywhere in the image and are denoted by the bullets. The image on the right depicts our proposed modification where the same-shift correlations are still chosen along the same diagonal as (*s*,t**) so that there is no overlap in the pixels used in each correlation, given by the slashes. The different-shift correlations are chosen in an analogous manner, but along the antidiagonal, represented by the bullets.*

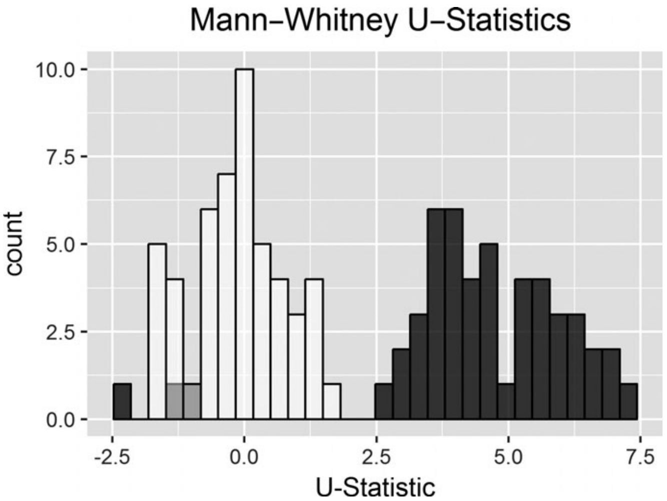

FIG. 4—*$\tilde{U}$-statistics associated with the 100 total sets of matching and nonmatching pairs of tool marks. The black histogram represents $\tilde{U}$-statistics from known matching tool marks and the white histogram for known nonmatching tool marks.*

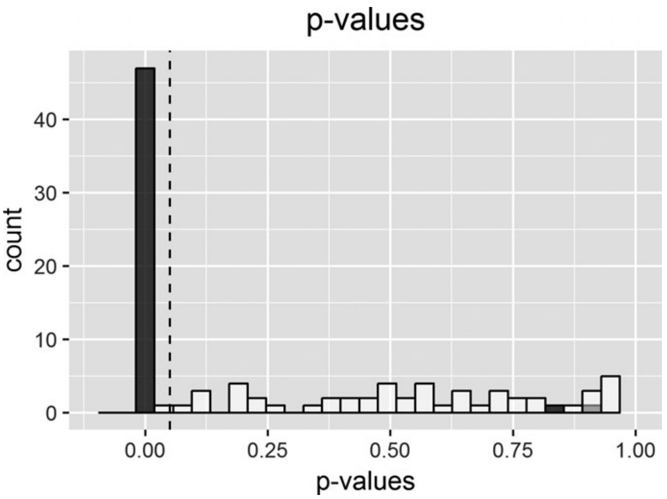

FIG. 5—*The distribution of* p*-values for known matches (black) and nonmatches (white) from the assumption of a standard normal reference distribution. The dashed line corresponds to a significance level of $\alpha = 0.05$.*

Given same-shift and different-shift correlations selected in this manner, we compute a standardized version of the Mann–Whitney $U$-statistic, which we denote here by $\tilde{U}$. (Details for how this is computed are given in the Appendix) In this modified form, we rely on the random variation in tool marks, rather than the randomized choice of correlations, as the basis for applying the $U$-statistic. Our deterministic selection of correlations for the same- and different-shift sets of correlations does not completely account for the nonindependence between the underlying segments because the data value associated with each pixel is used once in a same-shift correlation and once in a different-shift correlation. However, it does eliminate the reuse of any pixels from either mark in multiple same-shift correlations and multiple different-shift correlations. In the section to follow, we will see that this leads to a distribution very close to standard normal for values of $\tilde{U}$ computed from pairs of tool marks produced using different screwdriver, suggesting that critical values based on assumed independence can be used to at least approximately control the false-positive error rate. Because this is a one-tailed test (i.e., $\tilde{U}$ values associated with matching pairs tend to be larger than those associated with nonmatching pairs), the 95th percentile of the standard normal distribution (1.645) serves

TABLE 1—*The classification rates corresponding to the* p*-values in Fig. 5 where the rows correspond to the known status of the tool marks and columns represent the model conclusion. Of 50 pairs of known matching marks, we misclassified three pairs as nonmatches for false-negative error rate of 6% and 0 pairs of known nonmatching tool marks were identified as being matches.*

| | Conclusion | |
|---|---|---|
| Classification | Match | NonMatch |
| Truth | | |
| Match | 47 | 3 |
| NonMatch | 0 | 50 |

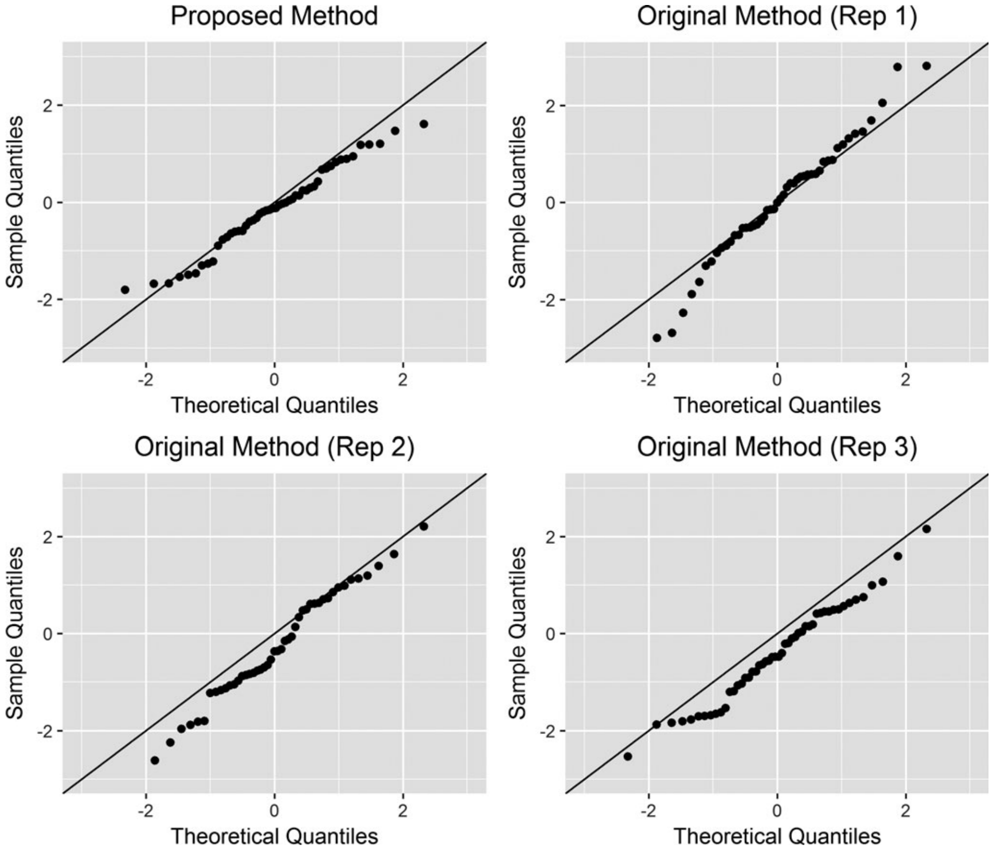

FIG. 6—Q–Q *plots comparing the $\tilde{U}$-statistics for 50 known nonmatching marks using our proposed method and three replications of the original method in (3) compared to standard normal distributions. Applying the Kolmogorov–Smirnov test for a standard normal distribution, we find that the $\tilde{U}$-statistics for the proposed method and the first rep of the original method are not significantly different from standard normal (*p*-values of 0.5102 and 0.8814, respectively). Conversely, the second two reps of the original method result in $\tilde{U}$-statistics that are significantly different than those that would correspond to a standard normal distribution (*p*-values of 0.0278 and 0.0281, respectively).*

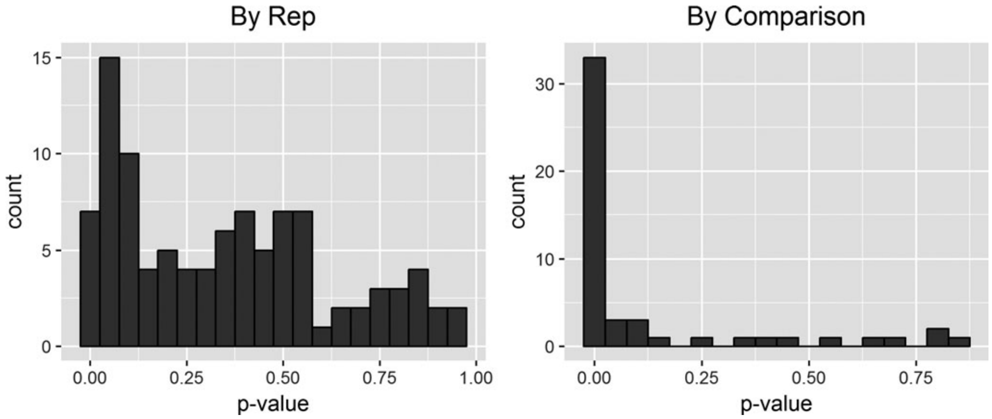

FIG. 7—*Based on applying the original method in (3) to the 50 pairs of known nonmatching tool marks, we conduct the Kolmogorov–Smirnov test for a standard normal distribution within reps (across the 50 comparisons) and within comparisons (across the 100 reps). The red vertical line represents a significance level of $\alpha = 0.05$, where we reject the assumption that the $\tilde{U}$-statistics follow a standard normal distribution if the* p-*value falls to the left of this line. Within reps (left histogram), we reject standard normality 14% of the time, and within comparisons (right histogram), we reject standard normality 72% of the time.*

as the critical value for a false-positive error rate of 0.05, and the 99th percentile (2.326) serves as the critical value for a false-positive error rate of 0.01. Alternatively, a *p*-value can be computed for any observed $\tilde{U}$ statistic by determining the probability that a standard normal random variable will be larger, for example, *p*-values smaller than 0.05 indicate rejection of the null hypothesis of nonmatch if a false-positive error rate of 0.05 is acceptable.

## Application

The data for this study were obtained from 50 sequentially manufactured screwdriver tips. The tool marks we consider were created with the screwdriver tips held above the surface at 30°, 40°, and 50°. (The study for which these tool marks were made was designed to investigate the effect of tool incidence angle on matching accuracy; we do not consider this effect here.) The known matching pairs of tool marks were formed by randomly selecting two tool marks made by the same tool at the same angle while the known nonmatching pairs of tool marks were formed by randomly selecting tool marks that were created at the same angle using different screwdrivers. We have selected 50 pairs of matching and nonmatching tool marks such that no tool mark is used more than once. For more information about the data collection process see (9).

Computing $\tilde{U}$ for each of these 50 pairs of known matching and nonmatching tool marks results in the values displayed in Fig. 4 where we have used an optimization window size of $n = 500$ and a validation window size of $m = 50$. It is visually clear that when the two tool marks match the $\tilde{U}$-statistic tends to be relatively large, and when they do not match, the value tends to be relatively small. Further, the sample of 50 values of $\tilde{U}$ obtained from nonmatching marks (on the left in the figure) resembles a standard normal distribution, and the Kolmogorov–Smirnov test for a standard normal distribution for this sample yields a $p$-value of 0.5102, implying we cannot reject the assumption of a standard normal distribution.

$p$-Values associated with each $\tilde{U}$ represented in Fig. 4, calculated under the assumption that the nonmatch distribution is standard normal, are displayed in Fig. 5. As should be the case, the $p$-values for nonmatching pairs (white histogram) are approximately uniformly distributed between zero and one, while the $p$-values for matching pairs are concentrated on small values (suggesting that they would be declared to be "matching"). Table 1 displays the results of classification based on a false-positive rate of 0.05, or critical value of $\tilde{U}$ of 1.645, showing that three of 50 matching samples (6%) are misclassified as nonmatching, and 0 of 50 nonmatching samples are misclassified as matching (Fig. 5 seems to show a nonmatching $p$-value below the vertical cutoff, but this is a visual side-effect of discretizing the continuous $p$-values).

While we have shown that our proposed method results in a distribution of $\tilde{U}$-statistics that can be considered standard normal and the resulting error rates for this small validation study are close to the specified error rate, we have not explicitly shown the lack of standard normality for the original method (3). As the original method is a randomized test, we apply the original method 100 times to each of the 50 known nonmatching tool marks and record the $\tilde{U}$-statistics. The $\tilde{U}$-statistics for three of these replicates as well as those from our proposed method are compared to a standard normal distribution using $Q$–$Q$ plots in Fig. 6. The $p$-values from the Kolmogorov–Smirnov test for a standard normal distribution show that our proposed method and one of the replicates (replicate 1) result in $\tilde{U}$-statistic distributions that can be considered standard normal, while the other two replicates (replicates 2 and 3) cannot. This verifies that in general the original method (3) does not result in a standard normal reference distribution.

In Fig. 7, we apply the Kolmogorov–Smirnov test for a standard normal distribution to each of the 100 replicates (across the 50 nonmatch comparisons) for the original method and also to each of the 50 known nonmatches (across the 100 replicates) and plot the resulting $p$-values. Of the 100 replicates, standard normality was rejected 14% of the time for tests performed at the $\alpha = 0.05$ level, while within each of the 50 known nonmatch comparisons, and standard normality was rejected 72% of the time.

## Summary and Discussion

We have proposed a modification to the method of (3) that eliminates the random nature of the U-statistic computed between same-shift and different-shift sets of correlations and reduces the nonindependence in these correlations due to overlapping tool mark segments. Using a normalized version of the digitized tool mark, our proposed solution is to select the set of same-shift and different-shift segments used in the validation step in a predetermined pattern so that there is no overlap between the segments chosen for either sample. By fixing the segments that are chosen, the randomized nature of the original procedure is also eliminated. While there is still some lack of independence *between* the correlations in the same- and different-shift sets of correlations, this appears to have less influence on the overall comparison procedure, with the reference distribution appearing to be standard normal. We have shown that, in general, this standard normality assumption does not hold across multiple applications of the original method (3). Our proposed alterations to the algorithm in (3) are simple and, in our relatively small validation experiment, lead to controlled false-positive error rates.

Finally, we acknowledge on referee's preference for statistical methods based on likelihood ratios and agree that the procedure described here is not of this class. Statistical methods for tool mark comparisons based on likelihood ratios have recently been advocated by a number of authors, for example (10–12), and we agree that they hold substantial promise in some cases. Computing a likelihood ratio requires the calculation of the probability of observing the result of a forensic examination under two different hypotheses or assumptions—specifically that the comparison in question is between objects that actually do "match" and that it is between objects that actually do not. In contrast, the method described in (3) and modified here is based on an assumption of statistical independence between two tool marks if they are produced by different tools, without reference to specific distributions. Both approaches to comparing competing hypotheses are well established in the statistics literature, and each has its advantages. Estimates of in-use false-positive and false-negative error probabilities require empirical studies for either kind of test and are valid only for the populations of material and conditions employed in those studies. This latter topic is certainly critical, but beyond the scope of the present paper.

The R code used to apply this methodology is being developed into an R package called *toolmaRk*, and we hope to have it available through CRAN in the near future. We would like to thank Heike Hofmann, Professor of Statistics at Iowa State University, for her hard work compiling the code and developing this package.

## References

1. Smith TP, Smith GA, Snipes JB. A validation study of bullet and cartridge case comparisons using samples representative of actual casework. J Forensic Sci 2016;61(4):939–46.
2. President's Council of Advisors on Science and Technology. Report to the President: forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: Executive Office of the President, 2016.

3. Chumbley LS, Morris MD, Kreiser MJ, Fisher C, Craft J, Genalo LJ, et al. Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm. J Forensic Sci 2010;55(4):953–61.
4. Conover WJ. Practical nonparametric statistics, 2nd edn. Danvers, MA: John Wiley & Sons, 1980.
5. Lock AB, Morris MD. Significance of angle in the statistical comparison of forensic tool marks. Technometrics 2013;55(4):548–61.
6. http://projects.nfstc.org/firearms/module13/fir_m13.htm (accessed March 2015).
7. Cleveland WS. Lowess: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 1981;35(1):54.
8. Bachrach B, Jain A, Jung S, Koons RD. A statistical validation of the individuality and repeatability of striated tool marks: Screwdrivers and tongue and groove pliers. J Forensic Sci 2010;55(2):348–57.
9. Faden D, Kidd J, Chumbley LS, Morris MD, Genalo L, Kreiser J, et al. Statistical confirmation of empirical observations concerning tool mark striae. AFTE J 2007;39(3):205–14.
10. Riva F, Champod C. Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases. J Forensic Sci 2014;59 (3):637–47.
11. Bunch S, Weavers G. Application of likelihood ratios for firearms and toolmark analysis. Sci Justice 2013;53:223–9.
12. Baiker M, Keereweer I, Pieterman R, Vermeij E, van der Weerd J, Zoon P. Quantitative comparison of striated toolmarks. Forensic Sci Int 2014;242:186–99.

Additional information and reprint requests:
Jeremy R. Hadler, Ph.D.
Department of Statistics
Iowa State University
Ames
IA 50010
E-mail: jrhadler@iastate.edu

## Appendix

Given two sets of correlations, those corresponding to same-shifts and those to different-shifts, we compute a standardized version of the Mann–Whitney *U*-statistic. Let $n_\mathrm{s}$ and $n_\mathrm{d}$ be the number of same-shift and different-shift correlations computed, and $N = n_\mathrm{s} + n_\mathrm{d}$. Additionally, let $R_\mathrm{s}(i)$ and $R_\mathrm{d}(j)$ be the ranks associated with the combined vector of correlations for the same-shift and different-shift correlations, for $i = 1,2,\ldots,n_\mathrm{s}$ and $j = 1,2,\ldots,n_\mathrm{d}$. Then the Mann–Whitney *U*-statistic is given by:

$$U = \sum_{i=1}^{n_s} R_s(i)$$

and its standardized version which accounts for the possibility of rank ties is as follows:

$$\tilde{U} = \frac{U - M}{\sqrt{V}}$$

With

$$M = n_s\left(\frac{N+1}{2}\right)$$

$$V = \frac{n_s n_d}{N(N-1)}\left[\sum_{i=1}^{n_s} R_s(i)^2 + \sum_{j=1}^{n_d} R_d(j)^2\right] - \frac{n_s n_d (N+1)^2}{4(N-1)}$$

where *M* and *V* are the mean and variance of the unnormalized *U*-statistic. Under assumptions that include statistical independence of all computed correlations, the large-sample distribution of $\tilde{U}$ is normal with mean zero and variance one (or standard normal) (4).