Reviewer Response

Susan Vanderplas, Heike Hofmann, Alicia Carriquiry

General Comments

We thank both reviewers for their insightful comments. While we respond point-by-point below, the reviews of this paper are careful, considerate, and helpful, and will serve to improve this paper. We regret that due to the space constraints PNAS imposes on brief reports, we could not address all of the suggestions in the paper, but in many cases we would have liked to include the extra nuance you (quite correctly) identified in your reviews.

Reviewer 1

- **Point 1**: Existing alternative solutions to multiple comparisons problems.
  We appreciate the reviewer's helpful suggestions about additional situations which may be similar and/or which may help to resolve the multiple comparison problem. One extremely unique factor of wire cuts that is not true of duct tape fits, bullets, or cartridge cases is that there is not any external structure which can assist with the multiple comparison problem. In duct tape, the tape has edges (and internal fibers) that provide structure. In bullets, there are lands and grooves which provide some structure and a way to sequence the lands according to the physical bullet orientation; this reduces a n^2 comparison problem to an n comparison problem, where n is the number of lands (though of course, database searches make even this a challenge). In cartridge cases, the extractor marks are of limited width and comparable size to the mechanism itself. In contrast, with wire cuts, there is a blade cut and a wire, and the wire could have been cut at any point along the blade; in most cases there is not any structure beyond the striation marks themselves. Thus, wire cuts serve as a uniquely problematic example within the wider issue of multiple comparisons in forensics. The Prusinowski paper is excellent, but the holistic judgement process demonstrated in e.g. Fig 7 of the manuscript is predicated on edge alignment as a first step; no physical analogue to this exists for wire cuts. While the Prusinowski paper's method is one way to avoid cross-correlation, it inherently relies on an assumption which is not true in most striation marks: that the length of the striated region is the same for the evidence as for the comparison sample. We also

note that previous papers by Prusinowski (e.g. 2020) use cross-correlation functions to describe the alignment between two samples.

- **Point 2** - Psychological concepts relating to visual comparison and decision making.
  The psychological concepts relating to holistic judgments are difficult to apply to wire cuts and even bullets, because examination of the striations visually precludes examination of the entire entity, either due to the length (e.g. the blade cut is many times longer than the wire diameter) or due to curvature (in the case of bullets). While it is possible that a talented cognitive psychologist could provide an improved examination procedure that would systematize this sort of comparison and enable holistic assessments, we aren't aware of any such procedures implemented in labs at the moment that are aimed at addressing this problem; in addition, as we are limited to 3 pages, it seems quite impossible to do the topic any justice. We would be most interested to collaborate with Reviewer 1 on an exploration of the zoomed-in vs. holistic cognition of striated mark comparisons, if they are a cognitive psychologist, as two of the three authors of this paper are statisticians who occasionally dabble in psychology-adjacent disciplines such as perception of statistical graphics.

- **Point 3** - Clarity, precision, and minor issues.

  - We have addressed the first bullet point by changing the language to emphasize that the issue is both multiple comparisons and large databases (which lead to multiple comparisons).

  - While we agree that the partial fingerprint was a contributing factor, given that we only have 3 pages, we have chosen to focus primarily on the database size factor, which is most relevant to this paper. It is certainly true that there are some analogies between partial fingerprints and e.g. stranded wires which may each be considered "partial" due to the nature of how patterns may be transferred from the tool to the wire fragment, but this distinction isn't necessary to make the main point of the paper.

  - We have added a description of what cross-correlation does, but the reference to the Vorburger paper will have to suffice for the definition because there is a lot of notation necessary to provide a definition of the theoretical quantity, and the computational implementations used in most algorithm implementations are even more complex to define and discuss.

  - Brief reports are limited to 15 references, so we did not have the option to additionally include citations to the following algorithms/papers, which is far from an exhaustive list of papers using the cross-correlation function in forensics to align items or evaluate fit:

    - Chu, W., Song, J., Vorburger, T., Yen, J., Ballou, S., & Bachrach, B. (2010). Pilot Study of Automated Bullet Signature Identification Based on Topography

Measurements and Correlations*†. Journal of Forensic Sciences, 55(2), 341–347. https://doi.org/10/bpkwtx

- Hare, E., Hofmann, H., Carriquiry, A., & others. (2017). Automatic matching of bullet land impressions. The Annals of Applied Statistics, 11(4), 2332–2356. https://doi.org/10.1214/17-AOAS1080
- Krishnan, G., & Hofmann, H. (2019). Adapting the Chumbley Score to Match Striae on Land Engraved Areas (LEAs) of Bullets,. Journal of Forensic Sciences, 64(3), 728–740. https://doi.org/10.1111/1556-4029.13950
- Ma, L., Song, J., Whitenton, E., Zheng, A., Vorburger, T., & Zhou, J. (2004). NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets. Journal of Forensic Sciences, 49(4), 1–11. https://doi.org/10/cdsbv8
- Chen, Z., Chu, W., Soons, J. A., Thompson, R. M., Song, J., & Zhao, X. (2019). Fired bullet signature correlation using the Congruent Matching Profile Segments (CMPS) method. Forensic Science International, 305, 109964. https://doi.org/10/gn649n
- Ott, D., Thompson, R., & Song, J. (2017). Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests. Forensic Science International, 271, 98–106. https://doi.org/10/gn649s
- Prusinowski, M., Brooks, E., & Trejos, T. (2020). Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. Forensic Science International, 307, 110103. https://doi.org/10.1016/j.forsciint.2019.110103
- Song, H., & Song, J. (2022). Virtual image standard (VIS) for performance evaluation of the congruent matching cells (CMC) algorithms in firearm evidence identifications. Journal of Forensic Sciences, 67(4), 1417–1430. https://doi.org/10.1111/1556-4029.15026
- Song, J., Chu, W., Tong, M., & Soons, J. (2014). 3D topography measurements on correlation cells—A new approach to forensic ballistics identifications. Measurement Science and Technology, 25(6), 064005. https://doi.org/10/gn65br
- Song, J., & Song, H. (2023). Reporting likelihood ratio for casework in firearm evidence identification. Journal of Forensic Sciences, 68(2), 399–406. https://doi.org/10.1111/1556-4029.15186
- Song, J.-F., Vorburger, T. V., Ma, L., Libert, J. M., & Ballou, S. M. (2005). A Metric for the Comparison of Surface Topographies of Standard Reference Material (SRM) Bullets and Casings. NIST. https://www.nist.gov/publications/metric-comparison-surface-topographies-standard-reference-material-srm-bullets-and
- Tong, M., Song, J., Chu, W., & Thompson, R. M. (2014). Fired Cartridge Case Identification Using Optical Images and the Congruent Matching Cells (CMC) Method. Journal of Research of the National Institute of Standards and Technology, 119, 575. https://doi.org/10/gn65bm
- Venkatasubramanian, G., Hegde, V., Padi, S., Iyer, H., & Herman, M.

    (2021). Comparing footwear impressions that are close non-matches using correlation-based approaches. Journal of Forensic Sciences, 66(3), 890–909. https://doi.org/10/gpjnz3

    - Wen, Z., Curran, J. M., & Wevers, G. (2023). Shoeprint image retrieval and crime scene shoeprint image linking by using convolutional neural network and normalized cross correlation. Science & Justice, 63(4), 439–450. https://doi.org/10.1016/j.scijus.2023.04.014
    - Zemmels, J., VanderPlas, S., & Hofmann, H. (2023). A Study in Reproducibility: The Congruent Matching Cells Algorithm and cmcR Package. The R Journal, 14(4), 79–102. https://doi.org/10.32614/RJ-2023-014
    - Zhang, N. F. (2021). Statistical models for firearm and tool mark image comparisons based on the congruent matching cells (CMC) method. Forensic Science International, 326, 110912. https://doi.org/10.1016/j.forsciint.2021.110912
    - Zhang, H., Zhu, J., Hong, R., Wang, H., Sun, F., & Malik, A. (2021). Convergence-improved congruent matching cells (CMC) method for firing pin impression comparison. Journal of Forensic Sciences, 66(2), 571–582. https://doi.org/10/gpjn2f

- We have changed the terminology to use d for the wire cross-section length, and b for the blade cut length. Hopefully this is somewhat more intuitive.
- We would love to use a more verbose explanation with less sophisticated vocabulary, but we are working within the confines of the 3 page limit imposed by Brief Reports - this requires linguistic brevity which sometimes necessitates a more terse explanation.
- The AFTE guidelines (emphasis added) are provided below. As interpretation is subjective and based on the individual's training and experience according to the AFTE guidelines, the rules used to evaluate the evidence are (it stands to reason) also subjective and based on individual training and experience, e.g. with the best agreement demonstrated between toolmarks produced by different tools, and consistent with the agreement demonstrated by toolmarks produced by the same tool.

  The three principles of the AFTE Theory of Identification as it Relates to Toolmarks are:

  1. The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the unique surface contours of two toolmarks are in sufficient agreement.
  2. This sufficient agreement is related to the significant duplication of random toolmarks as evidenced by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows. Specifically, the

relative height or depth, width, curvature and spatial relationship of the individual peaks, ridges and furrows within one set of surface contours are defined and compared to the corresponding features in the second set of surface contours. Agreement is significant when it **exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools** and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that sufficient agreement exists between two toolmarks means that the agreement is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility.

3. Currently the interpretation of individualization/identification is **subjective in nature**, founded on scientific principles and **based on the examiners training and experience.**

- We have added additional information about data pooling - the calculations were indeed weighted by sample size.

Reviewer 2

- We have attempted to streamline the number of terms used for error rates, using false discovery rate and familywise false discovery rate consistently throughout the paper. In the conclusion, we reference error rates more generally, but have clarified that we mean both false discovery and false elimination errors.
- There are two separate questions in the second comment:
  - Implicit multiple comparisons in different toolmark disciplines.
    While brevity forced us to distinguish implicitly between multiple comparisons in bullets or cartridge cases and multiple comparisons in wire cuts - the latter are a much more extreme case than the former, as there is some external structure present in both cartridge cases and bullets which limits the scale of comparisons necessary. For instance, in bullets, there are the land-groove transition areas, and we know that each land must occur in sequence; these physical constraints reduce the potential number of comparisons over the surface of the bullet. In wire cuts, however, there is an extremely limited external structure that would assist with identifying which part of the blade was used to cut the wire - the only indicators are the striation marks themselves and the potential configuration of the blade - whether two cutting surfaces or 4 contributed to the wire cut, and how those surfaces might be aligned. This magnifies the problem (which is still very present in bullets) to an extreme which is useful for presentation in a brief report such as this one - the problem itself is relatively simple to explain and sufficiently extreme

to illustrate the multiple comparison problem and motivate its importance when examining the scientific basis of forensic pattern comparisons.

  - Suggesting that the establishment of firearms examination bona fides is a done deal.

    While we do not agree that forensic firearm examination has been scientifically shown to be reliable to the levels necessary in the 2009 NAS report and 2015 PCAST report, within the category of striated toolmark evidence, firearm examination is certainly the most well-developed and well-studied discipline. We are well aware of the statistical issues with firearm studies, having published several papers ourselves highlighting these problems; with that said, we must also acknowledge that some more recent studies have fewer design flaws than historically well-cited studies. Even with the problems shared by these more recent studies (participant self-selection, lack of reporting of participant dropout, lack of casework level realism, participant blindness to the study, breadth of firearm/ammunition selection), the study design is sufficient to provide some information about error rates. This information may not be generalizable to a population or useful for statistical inference, but the studies we cite are sufficient to provide descriptive statistics; these statistics can be aggregated to provide a useful hypothetical error rate that represents a "ballpark" estimate of striated error rates. We have worked through the math with a whole host of different error rates and the conclusions are similarly dramatic even if the error rate changes by a factor of 10. Thus, our decision to ground the hypothetical error rate in estimates from firearm studies is motivated primarily by an attempt to be charitable and suggest that the multiple comparison problem is still a problem even if we accept the low error rates in better-designed (but still suboptimal) published studies.

- Putative tool - we have changed the language, using 'recovered tool' to indicate a tool recovered from the suspect.

- Figure 1 clutter and "two sides"

  We've cleaned up the figure as much as possible. We agree that it is cluttered, but we needed to increase the information density of the picture in order to meet the space constraints of PNAS brief reports. We have simplified the tool labels, using *b* for the length of the blade cut and *d* for the diameter of the wire. Hopefully these are easier to read.

  The tool shown in Figure 1 has a single blade with two sides (which are shown as inset figures between the legs of the tool). In other tool configurations, it is possible to have two blades that meet at the tip, and thus to have 4 cutting surfaces, 2 of which may be represented on each half of the cut wire. As we mention ("number of surfaces"), this setup may make additional comparisons necessary, further compounding the situation we illustrate in the paper. Examiners may also take multiple blade cuts at different angles, which only makes the problem more serious. We have written this paper to account for the simplest possible wire cut solution with a tool that has minimal cutting surfaces, and our results are still catastrophic when we try to enforce even moderate controls on false discovery errors.

- "on average"
  Strictly speaking, because striated surfaces tend to have periodic oscillations due to the striations, non-overlapping comparisons may have some dependence. Striations are typically random, and so the expected value of this dependence is zero, even if for a single instance of a striated blade, there may be non-zero correlation between non-overlapping comparisons. We've examined and established this result empirically, but wanted to be technically precise while maintaining readability for an audience who are not all statistically sophisticated enough to identify the nuances of dependence structures in this type of data.

- "family-wise error rate"
  We've added a citation to Tukey's 1953 JASA paper on multiple comparisons.

- "two components."
  The problem is that with a subjective set of evaluation rules that are examiner-dependent, the two sources cannot be separated. We address this a few sentences later when we say "Assuming that lab procedure errors are not a factor in studies, we use reported error rates from three open-set studies of striated evidence"...that is, we are explicitly assuming that the procedure errors are not a factor in the studies we are using.

- Relevance of the results of firearms studies.
  We agree that the firearms studies have their weak points, which we've briefly summarized in a previous comment response. We decided after some discussion that using an error rate grounded in some empirical study was more concrete (and thus easier to relate to) than using an example error rate not connected to reality (which, statistically, would have worked just as well without the baggage). There are certainly pros and cons to each approach, but we wanted to make this paper as concrete as possible with the hopes of motivating action on the issue of multiple comparisons more generally and wire cut forensic analysis specifically.

- We've removed the reference to Type II studies, as it is a distraction from the main point of this paper.