Hidden.multiple.comparisons.increase.forensic.error.rates.by.Vanderplas.
Carriquiry.Hoffman

One thing that strikes me in this manuscript is the number of terms for error rates. Here's a list, with the lines where they occur:

false positive rate	27-28; 234-235
false discovery rate	47, 51-52 and reference (1)
family-wise error rate	229-230, 267-268, 304
family-wise false identification error	273
examination-wide error rate	318
actual error rate	327
examination-wise error rate	340

There is very little attempt to define and distinguish among these. In particular, the reference to false discovery rate—the only one of these terms where the paper cited gives a definition—is irrelevant to the authors' argument, as I understand it, which is: given n separate comparisons to test the single hypothesis that *this* wire cutter gave rise to *this* cut in *this* wire, what would the chances be of making an error, if the probability of an erroneous assessment in a single comparison (of the marks on the wire to a given portion of the blade) were such-and-such—the latter being their implicit definition of the false positive rate.

So a little distinguishing clarity and better referencing might help. I suspect only two or three of the above list are necessary.

* * *

13-15 "….introduces complexities......better investigated..." This puzzles me. Are the authors saying there are no implicit multiple comparisons in bullet comparisons? Would seem to be contrary to practice of, e.g. searching for best alignment of grooves. Also, seems to suggest that establishment of firearms examination *bona fides* is a done deal. Do the authors mean to imply that?

58 "putative tool" surely it *is* a tool. Maybe better "suspect tool"

Fig 1 very cluttered figure. Was not very helpful to me in coming to conclusion as, e.g., what the different "sides" being spoken of were. The label above the upper rectangle overlapping the tool is illegible.

I take it the two “sides” refers to a single blade, but in the case of the wire to the marks on the two separate wire segments produced by the single cut. This could be made clearer in the text. (Maybe a drawn figure would help.)

182-183 “the number of surfaces” would it be just one surface, if one of the wire segments were missing? Or is something else meant?

209 “(on average)” I don’t understand what this phrase is adding, or what qualification is in mind

229 “*family-wise error rate”* key concept of the paper—maybe deserves a reference?

234-238 “two components” not sure this can’t be expressed better. Isn’t it really “factors leading to”—(a) the chance alignment of different items out there in the world, and (b) the possible “internal” procedural errors (like being misled by expectation bias and so on). Maybe would be good to make clear that you are focusing on the former in this article.

243-267 and Figure1. I am uneasy about the relevance of the results of the firearms studies. It seems to give an imprimatur to those studies, which certainly have their weak points. It’s not clear why anyone would be tempted to take their results as representing the sort of FPR that would attach to comparing one segment of a blade to the marks in a cut wire. As noted in lines13-15, the firearms exams themselves have implicit multiple comparisons. This attempt at realism seems misleading.

325-328 “Type II studies” Very loose wording. Koehler’s actual term is “Type II proficiency testing”—and he is very specific about this “Type II” activity differing from other “Type II” procedures. What is involved is specified, not in the paper the authors cite, but in one that paper itself references: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255

And Koehler’s focus is on (b) above, i.e. is orthogonal to the main concern of this manuscript (cf. 244). Shouldn’t this be made clear?