Forensic Science Policy & Management: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ufpm20

Do Observer Effects Matter? A Comment on Langenburg, Bochet, and Ford

R. Koppl[a], D. Charlton[b], I. Kornfield[c], D. Krane[d], M. Risinger[e], C. Robertson[f], M. Saks[g] & W. Thompson[h]

[a] Syracuse University, Syracuse, New York
[b] Surrey and Sussex Forensic Services, UK
[c] University of Maine, Orono, Maine
[d] Wright State University, Dayton, Ohio
[e] Seton Hall University, Newark, New Jersey
[f] University of Arizona, Tucson, Arizona
[g] Arizona State University, Phoenix, Arizona
[h] University of California Irvine, Irvine, California

Published online: 19 Aug 2015.

To cite this article: R. Koppl, D. Charlton, I. Kornfield, D. Krane, M. Risinger, C. Robertson, M. Saks & W. Thompson (2015) Do Observer Effects Matter? A Comment on Langenburg, Bochet, and Ford, Forensic Science Policy & Management: An International Journal, 6:1-2, 1-6, DOI: 10.1080/19409044.2014.995385

To link to this article: http://dx.doi.org/10.1080/19409044.2014.995385

Forensic Science Policy & Management, 6(1-2):1–6, 2015

ISSN: 1940-9044 print / 1940-9036 online
DOI: 10.1080/19409044.2014.995385

Do Observer Effects Matter? A Comment on Langenburg, Bochet, and Ford

R. Koppl,[1] D. Charlton,[2] I. Kornfield,[3] D. Krane,[4] M. Risinger,[5] C. Robertson,[6] M. Saks[7] and W. Thompson[8]
[1]Syracuse University, Syracuse, New York
[2]Surrey and Sussex Forensic Services, UK
[3]University of Maine, Orono, Maine
[4]Wright State University, Dayton, Ohio
[5]Seton Hall University, Newark, New Jersey
[6]University of Arizona, Tucson, Arizona
[7]Arizona State University, Phoenix, Arizona
[8]University of California Irvine, Irvine, California

ABSTRACT We identify methodological problems in Langenburg et al. (2014), which undermine its conclusions about the size of the observer effect problem and the importance of sequential unmasking as a solution. The scoring method of Langenburg et al. (2014) appears to be subjective. The classification of cases is not congruent with the three keys to observer effects in forensic science: the analyst's state of expectation, the analyst's state of desire, and the degree of ambiguity in the evidence being examined. Nor does the paper adequately support its claim, "[I]t has been asserted that the high context/high interaction cases are essentially where there is the most danger of bias." While the paper tends to minimize concern over observer effects, the evidence in it seems to support the view that fingerprint analysts look to contextual information to help them make decisions.

KEYWORDS Observer effects, bias, sequential unmasking, error

DO OBSERVER EFFECTS MATTER?

Langenburg et al. (2014) have provided a service by presenting data on latent print work at the Minnesota Bureau of Criminal Apprehension. They uncover potentially useful facts such as the rate at which latents were recovered from plastic bags. Thus, the paper succeeds in its stated goal of providing "casework statistics" (p. 16) from Minnesota's Bureau of Criminal Apprehension Latent Print Unit (BCA-LPU). Additionally, "A portion of the . . . paper was dedicated to the exploration of possible bias effects from significant interaction between the forensic analyst and the case investigator, or from analyst exposure to contextual information about the case" (p. 16). Langenburg et al. drew several conclusions in this later part of their paper, including "there is usefulness in sequential unmasking." We commend this affirmation of a role for sequential unmasking, even though Langenburg et al. see the potential as more limited than we do.

In spite of this point of agreement, we believe that the portion of the paper on bias was flawed and cannot be used to infer that the risk of observer effects is lower in fingerprint analysis than in other expert domains. The study is of limited value, therefore, in judging the proper scope for sequential unmasking or other measures that are meant to limit bias or its bad effects. The paper suffers

Received 1 October 2014; accepted 2 December 2014.

Address correspondence to R. Koppl at 721 University Avenue, Syracuse, NY 13244. E-mail: rkoppl@syr.edu

from some basic methodological flaws and, consequently, it may not be possible to draw policy-relevant inferences from its analysis.

Langenburg et al. distinguish between “interactions” and “contextual information” as sources of bias. They use a two-dimensional classification in which cases are rated on the “level of interaction” of the analyst with case investigators and on the “amount” of domain-irrelevant context information the analyst was exposed to. The level of interaction might be “high,” “moderate,” or “none/minimal,” and the amount of domain-irrelevant context information might be “high,” “moderate,” or “none/minimal.” Langenburg et al. seem to think that the risk of a bias-induced error grows as we move from none/minimal to high on either dimension. Thus, they reason, a bias-induced error is most likely in cases with high interaction and high context information, and error is least likely in cases with none/minimal interaction and none/minimal context information. They say, “Two subsets of those data were compared: the cases where there was high context information and high interaction (high context/high interaction; N D 18) versus the cases where there was no context information and no interaction (no context/no interaction; N D 466). The reason for doing so is that it has been asserted that the high context/high interaction cases are essentially where there is the most danger of bias—that the analyst is receiving significant non-domain information and cues from investigators” (p. 30).

As a bottom line, the authors report that 142 of 650 (22%) “no context/no interaction” cases resulted in an individualization, whereas 25 of 121 (21%) “high context/high interaction” cases resulted in an individualization (pp. 30–31). Langenburg et al. infer, “Essentially, there was no difference in the rate of identification between these two subgroups.” But the 95% confidence interval for the first rate minus the second rate is about −7% to 8%. (Our calculation assumes that we may model the underlying processes as independent Bernoulli trials.) While they could correctly point out that this result shows that we cannot rule out the null hypothesis of equal rates between the two groups, their evidence does not rule out substantial observation effects.

Yet, there are more fundamental problems. First, there is a methodological question of scoring. Although the authors acknowledge that the scoring of context and interaction was a “judgment call,” they do not report whether raters were blinded to the dependent variables under study when they rendered those judgments. Nor do the authors report any measures of inter-rater or intra-rater reliability for these assessments. It is thus hard to know if these measures have any validity, even if the underlying concepts were clear.

Second, there is a problem of “confounding.” As Vandenbroucke (2002) explains, the word “confounding” is generally “used for one particular form of the confusion of two effects: the confusion due to extraneous causes, i.e., other factors that really do influence” the processes under study (p. 219). By the authors’ own account, the paper’s “high context/high interaction” cases have a disproportionate number of homicides, while the “no context/no interaction” cases have a disproportionate number of property crimes. To compare their individualization rates thus seems an apples and oranges comparison.

One might, perhaps, use multivariate regressions to control for such confounds. Or one might attempt some sort of matching or subset analysis. We wonder, however, whether the dataset used is big enough to allow the fruitful use of such techniques. Even then, one might worry about an unavoidable risk of surveys of the type made by Langenburg et al.: omitted variables that correlate with the observed variables and thus skew the analysis.

Such considerations support the view that controlled experiments may be the best way to get at the issues in question. There have been such experiments using trained, working, professional fingerprint examiners (Dror and Charlton 2006; Dror, Charlton, and Peron 2006). These studies support the view that fingerprint examiners, like all other humans, are subject to observer effects. The effect sizes in these studies are large enough to make bias by domain-irrelevant information a source of concern for anyone interested in avoiding false convictions.

Langenburg et al. find that the rate of identification is about the same in the cases classified as high context/high interaction and those classified as no context/no interaction. They say, “This is not compelling evidence that analysts are highly motivated to find only evidence to support the police theory and are being influenced by interactions with police and prosecutors” (p. 31). But by their own account, as we have noted, the cases dubbed “high context/high interaction” include a disproportionate number of homicides and those dubbed “no context/no interaction” include a disproportionate number of property crimes. They say,

> There was a difference in the rates of exclusions to suspects: there were nearly 3 "exclusion" decisions per latent print for high context/high interaction cases, whereas there was only 1 "exclusion" decision for every 4 latent prints in no context/no interaction cases. Proportionately, there were 12 times as many exclusions of suspects in high context/high interaction cases as there were in no context/no interaction cases. This is likely due to the higher number of suspects against which to compare in homicide cases in the high context/high interaction cases compared to the large number of property crimes, where there is usually no suspect provided about half the time, dominating the no context/no interaction cases. (p. 31)

Thus, Langenburg et al. seem to have found that in cases such as murder we usually have multiple suspects and a larger numbers of latents to consider. In those cases, they report, examiners pay lots of attention to the case file and have frequent interactions with police investigators. By contrast, in the typical property crime, there are no suspects and only a few latents to consider. In the first class, more latents are judged to be "identifiable" than in the second class. We also get more exclusions in the first class than in the second class. Of the latents judged useable in each class, the ratio of identifications is about the same.

All of this is supposed to support the view that bias is somehow a smaller problem than advocates of sequential unmasking believe. But their own evidence says that when examiners have a greater chance of making an identification that might later be shown to be incorrect, they are more likely to interact with police investigators and to acquire case information. This result suggests to us that fingerprint analysts look to contextual information to help them make decisions.

Although the foregoing is sufficient to understand why the Langenburg et al. paper is *not* contrary to the emerging consensus that observer effects are a real and substantial problem, it is also important to address several foundational methodological and conceptual issues for the sake of future research and to advance the literature on observer effects in forensic science. Fundamentally, we think that classification scheme is inadequate for their stated purpose. It does not adequately reflect the three keys to observer effects in forensic science: the analyst's state of *expectation*, the analyst's state of *desire*, and the degree of *ambiguity* in the evidence being examined. Krane et al. (2008) say, "Observer effects are rooted in the universal human tendency to interpret data in a manner consistent with one's expectations. This tendency is particularly likely to distort the results of a scientific test when the underlying data are ambiguous and the scientist is exposed to domain-irrelevant information that engages emotions or desires" (p. 1006).

Risinger et al. explain, "At the most general level, observer effects are errors of apprehension, recording, recall, computation, or interpretation that result from some trait or state of the observer" (2002, p. 12). The relevant "state of the observer" may be a state of expectation. Domain-irrelevant information can create an expectation that a given pair of known and unknown prints have a common source or that they do not have a common source. The relevant "state of the observer" may be a state of desire. As Risinger et al. (2002, p. 24) note, "[W]here an observer has strong motivation to see something, perhaps a motivation springing from hope or anger, reinforced by role-defined desires, that something has an increased likelihood of being 'seen.'" Risinger et al. (2002) also note the importance of ambiguity. "Of course, where the evidence is clear, the cognitive biases, which operate best on ambiguity, can be overridden. Conversely, observer effects are most potent where ambiguity is greatest, when an observer's judgment is most likely to succumb to expectation, subjective preference, or external utility" (Risinger et al. 2002, p. 16). Whitman and Koppl (2010) have a model of Bayesian decision-making in which ambiguity increases the chance of a bias-induced error. Expectation, desire ("subjective preference or external utility"), and ambiguity are the three key factors producing observer effects in forensic science decision making.

Langeburg et al. do not provide citations that support their statement, "*it has been asserted* that the high context/high interaction cases are essentially where there is the most danger of bias" (p. 30, emphasis added). When first drawing the distinction between interactions and context information they cite three works authored or co-authored by Itiel Dror (Kassin, Dror, and Kukucka 2013; Dror 2013; Dror and Hampikian 2011). But the word "interaction" appears in only one of the cited works, and then only in the phrase "social-interactional context," which was used once in the description of someone else's study of confirmation bias (Kassim, Dror, and Kukucka, p. 44). At least one important article, however, does say that "interactions" between examiners and investigators create a risk of observer effects. Risinger et al. (2002) note the danger of such interactions and once use that word in that connection (p. 37). In elaborating on the problem, Risinger et al. quoted Evan Hodge:

> [The examiner] gave in to investigative pressure. We all do this (give in to investigative pressure) to one extent or another. A hot case comes in, the investigators want to wait, want to look over your shoulder, want to see the ident, help you shoot the gun, etc. Do you take shortcuts? Do the words "the commissioner, or the director, or the captain wants to know right now" affect you? Of course they do, don't kid yourself. (Hodge 1988, p. 292 as quoted in Risinger et al. 2002, p. 38)

Hodge's article has been cited elsewhere in the literature on observer effects in forensic science (Koppl 2005, p. 261; Kelly and Wearne 1998, p. 17). It is therefore true that "interactions" between examiners and investigators have been an object of concern, but this concern needs to be understood in the context of the scientific literature on observer effects. The root concern is not "interaction," *per se*, but the ambiguity of the evidence and an analyst's states of expectation and desire.

An interaction between an analyst and an investigator may produce a change in the analyst's state of expectation, her state of desire, or both. In this sense, the category "interaction" is a mélange of expectation and desire. An interaction may produce only a change in expectation if the analyst is already motivated to support the police theory. It may produce only a change in desire if the analyst, for example, feels the sort of social pressure Hodge (1988) warned of. Finally, of course, an interaction may produce a change in both the analyst's state of desire and her state of expectation. Similarly, "context" as defined by Langenburg et al. may change an analyst's state of expectation, state of desire, or both.

We are not aware of anything in the literature on observer effects in forensic science that asserts that the combination of "high context" with "high interaction" is "essentially where there is the most danger of bias" (p. 30). Rather, the key factors are expectation, desire, and ambiguity—not interaction and context. As we have seen, the sequential unmasking letter of Krane et al. (2008) does suggest that a combination of "expectation" and "desire" may be more likely to generate a bias-induced error than either in isolation. Thus, *this* combination has been viewed with particular concern in the literature on observer effects in forensic science. But if "context" and "interaction" as defined by Langenburg et al. are both mélanges of "expectation" and "desire" as defined in the literature on observer effects, then it seems doubtful what causal significance should be imputed to the combination of "context" and "interaction."

Langenburg et al. measure "interaction" principally by the number of communications between an analyst and an investigator. To be categorized as "high interaction" the record must reveal "at least 3 phone calls, at least 3 email exchanges, or attendance at the crime scene" (p. 18). While these factors may be generally correlated with an analyst's state of desire, state of expectation, or both, it must also be considered that a single intense interaction could easily have a greater effect than countless insubstantial ones.

Langenburg et al. measure "context" by the number of "case details" to which the analyst was exposed. But the number of details does not have to be high to induce a high state of expectation in an analyst. Some of the language used by Langenburg et al. suggests that they considered not only the number of details, but also whether the details given were "significant" (p. 19). Unfortunately, context was uniformly "high" in the one example they give to illustrate the difference between high and low context cases. They say,

> Typically, we opted for a higher level of interaction/context information if there was any doubt. For example, if the case only had a short note such as "we are looking for subject's prints on the gun," this was designated as a "minimal to no context" case. If the officer wrote (and this would be extraordinary and did not occur in these samples) "we are looking for subject's prints on the gun— we know he did it and he's a bad person who needs to come off the streets," then this would be categorized as a "high context" case even though it is a single, short statement made to the laboratory. (p. 22)

In a private communication, Langenburg has explained to us that police investigators must include an "evidence submission form" with all forensic evidence sent to the crime lab. This form includes the names and birth dates of both victims and any suspects. If there is no suspect, as when drugs are found at the side of the road, the lab may not take the case in. Analysts will generally be aware of the information on the submission form because they are required to check it for possible errors. Langenburg has explained to us that these completed forms were not generally considered a source of potentially biasing information in the study. An egregious remark in the context box could cause the case to be classified differently, but not remarks indicating that results are needed by a certain date or to look for latent prints first on this object, then on the other, and so on. In our experience, the information routinely supplied in evidence submission forms will often, indeed usually, contain potentially biasing information.

Indeed, in the very case used by Langenburg as an exemplar of "minimal to no context" the analyst has been told, as it were, the "right" answer. Certainly, the known and unknown prints must be considered together at some point. But the principles behind sequential unmasking as articulated by Krane et al. (2008) suggest that the crime-scene latent be submitted for characterization before any known prints are available to the analyst. Analysts should first determine whether the detail observed is sufficient to make the crime scene item a potential source of useful information before any known prints are available to her. (This proposal addresses the initial "analysis" stage of the ACE-V method of fingerprint examination.)

The general literature on observer effects (reviewed in Risinger et al. 2002) as well as the literature specific to fingerprint analysis (Dror and Charlton 2006; Dror, Charlton, and Peron 2006) together suggest that the simultaneous presentation of the latent and one known print creates a risk of error through observer effects. Thus, the example of "minimal to no context" given by Langenburg et al. is, instead, a case in which there is an appreciable risk of bias. As Whitman and Koppl (2010, p.70) have noted, "The authorities—police and prosecutors—implicitly convey information to forensic examiners by their very decision to submit samples for testing." Saying the suspect is a bad person is not always necessary to create a state of expectation in an analyst.

Let us consider the analysts' state of desire. The fingerprint examiners of the BCA-LPU are employed in a law-enforcement agency. This alone has the potential to create a desire to help law enforcement officers, and this desire, in turn, has the potential to bias decision-making. As the National Academy of Sciences has noted, "Forensic scientists who sit administratively in law enforcement agencies or prosecutors' offices, or who are hired by those units, are subject to a general risk of bias" (NAS p. 6–2). The administrative position of fingerprint examiners in the BCA-LPU is invariant across all cases studied by Langenburg et al. It is thus possible, even likely, that these examiners generally had a state of desire capable of producing observer effects. It seems questionable, then, whether the analyst's state of desire is highly correlated with either "context" or "interaction" as defined by Langenburg et al., even though one or more case details or pressure from an investigator could enhance an analyst's state of desire (Charlton et al. 2010).

Let us now consider the analysts' state of expectation. If the analysts are typically presented the known and unknown prints together, then they will typically have at least some degree of expectation that one or more latent prints have the same source as a print from the "subject." Thus, the analyst's state of expectation, like her state of desire, may not be highly correlated with either "context" or "interaction" as defined by Langenburg et al., even though one or more case details or an interaction with an investigator could enhance an analyst's state of expectation.

Importantly, the classification employed by Langenburg et al. does not appear to consider in any way the ambiguity of the latent prints submitted for evaluation. Dror et al. (2005) distinguish the "bottom up" information given by the latent or other evidence from domain-irrelevant information, which they dub "top down." They say, "weakening the bottom-up information may allow the top-down component more room to influence the process" (p. 803). When an evidence sample is unambiguous, context information is less likely to induce an error.

In the end, fingerprint analysts must *decide* when to declare an individualization, when to declare an exclusion, and when to declare that no reliable judgment can be made. Presumably, analysts want to moderate what we might call "reversal risk," the risk that a decision will later be determined to have been mistaken. Charlton et al. (2010) conducted a survey of fingerprint examiners and concluded in part, "[T]here was an expression of fear and consequence in making an erroneous match" (p. 391). They also found "a desire to avoid ambiguity" (p. 390). Domain-irrelevant information may help them reduce subjective doubt about which decision has the lowest reversal risk. If fingerprint examiners are decision makers who are similar to decision makers in other areas, including other expert domains, then the potential of domain-irrelevant information to help resolve ambiguity and subjective doubt may lead them to more energetically seek out and respond to domain-irrelevant information when reversal risk is greater. From this point of view, the differences Langenburg et al. found between "high context/high interaction" cases and "no context/no interaction" cases seem consistent with the view that domain-irrelevant information may be creating observer effects in the BCA-LPU.

REFERENCES

Charlton, D., P. Fraser-Mackenzie, and I. E. Dror. 2010. Emotional experiences and motivating factors associated with fingerprint analysis. *Journal of Forensics Sciences* 55(2): 383–393.

Dror, I. E. 2013. The ambition to be scientific: Human expert performance and objectivity. *Science & Justice* 53(2): 81–82.

Dror, I. E., and D. Charlton. 2006. Why experts make errors. *Journal of Forensic Identification* 56: 600–616.

Dror, I. E., D. Charlton, and A. Peron. 2006. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International* 156: 174–178. http://dx.doi./10.1016/j.forsciint.2005.10.017

Dror, I. E., and G. Hampikian. 2011. Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice* 51(4): 204–208.

Hodge, E. 1988. Guarding against error. *Journal of Association of Firearm and Tool Mark Examiners* 20: 290–293.

Kassin, S. M., I. E. Dror, and J. Kukucka. 2013. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition* 2(1): 42–52.

Kelly, J. F., and P. Wearne. 1998. *Tainting Evidence: Inside the Scandals at the FBI Crime Lab*. New York, NY: The Free Press.

Langenburg, G., F. Bochet, and S. Ford. 2014. A report of statistics from latent print casework. *Forensic Science Policy & Management* 5(1–2): 15–37

Risinger, M., M. J. Saks, W. C. Thompson, and R. Rosenthal. 2002. The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review* 90(1): 1–56.

Vandenbroucke, J. P. 2004. The history of counfounding. *Soz Praventiv Med*. 47(4): 216–224.

Whitman, D. G., and R. Koppl. 2010. Rational bias in forensic science. *Law, Probability & Risk* 9(1): 69–90.