Exhibit ‘F’

Forensic Science International: Digital Investigation 37 (2021) 301175

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

A hierarchy of expert performance (HEP) applied to digital forensics: Reliability and biasability in digital forensics decision making

Nina Sunde [a, *], Itiel E. Dror [b]

[a] University of Oslo, Norwegian Police University College, Norway
[b] University College London, UK

ARTICLE INFO

Article history:
Received 7 November 2020
Received in revised form
29 April 2021
Accepted 3 May 2021
Available online xxx

Keywords:
Bias
Reliability
Digital forensics
Forensic science decision making
Human factors
Contextual information
Expert judgment

ABSTRACT

In order to examine the biasability (impact of contextual information) and reliability (consistency) of digital forensic observations, interpretations, and conclusions, 53 digital forensics (DF) examiners analysed the same evidence file. For biasability, some DF examiners were provided with contextual information suggesting guilt or innocence, while a control group received no contextual information. As per biasability, the results showed that the DF examiners' observations were affected by the biasing contextual information. As per reliability, the results showed low reliability between DF examiners in observations, interpretations, and conclusions. For improving DF work, as well as for transparency, it is important to study and assess the biasability and reliability of their decision making.

1. Introduction

There is growing use of digital evidence in criminal investigations, a trend that is likely to increase with greater use of technological devices by the public and with new digital forensics (DF) capabilities. Although digital evidence is often perceived as objective and credible, there is a range of possible errors and uncertainties inherent in the evidence itself (Casey, 2002), as well as errors that can derive from the human factors involved in the processes in which the evidence is produced (Pollitt et al., 2018; Sunde and Dror, 2019). DF work involves many judgements and decisions that require interpretation and subjectivity (Sunde and Dror, 2019). This means that the quality of the outcome of the DF process – digital evidence – is dependent on cognitive and human factors, which can lead to bias and error.

DF is a relatively new and fast-changing domain in forensic science. The ever-changing landscape of technology, both of the evidence itself as well as the forensic tools to examine it, require constant adaptation and updating of methodologies and tools. This constant flux may have contributed to what may be described as a 'quality challenge' in DF. Page et al. (2019) referred to the quality regime of DF as a "wild west", where key components for quality management rarely seemed to be in place.

The DF domain faces challenges with rapid technology changes, increasing complexity, issues of accessibility, and increasing amounts of data. Research has attempted to shed light on how to produce high-quality digital evidence within this challenging context (e.g. Watkins et al., 2009; van Baar et al., 2014; van Beek et al., 2015; van Beek et al., 2020). However, the role of the human DF examiner has only been subject to research in a few studies (e.g. James and Gladyshev, 2013; Wilson-Kovacs, 2019). Specifically, performance and decisions making in terms of reliability (i.e., consistency, repeatability, reproducibility) and biasability have not been given sufficient attention. Understanding the human role, and the sources of bias and error (Dror, 2020) is crucial for developing effective quality measures and for transparency.

The research on expert performance and decision making in other forensic science domains has shown that even DNA analysts, fingerprint examiners, and forensic toxicology analysts can produce inconsistent results and are prone to bias (e.g. Dror and Hampikian, 2011; Dror et al., 2006).

To gain knowledge about potential bias and error in the DF process, it is therefore important to study DF decision making. This

* Corresponding author.
E-mail address: Nina.Sunde@phs.no (N. Sunde).

https://doi.org/10.1016/j.fsidi.2021.301175

knowledge is an essential foundation for transparency and designing effective measures that can minimize error, or detect them before they cascade further into the investigation process. Invalid digital evidence is a threat to the fair administration of justice and can cause wrongful convictions (Garrett and Neufeld, 2009; Garrett, 2021) or the release of guilty culprits.

1.1. Hierarchy of expert performance

The Hierarchy of Expert Performance (HEP) (Dror, 2016 – see Fig. 1) is a useful framework for exploring and quantifying expert performance in DF. This framework has been applied to examine forensic anthropology (Hartley and Winburn, 2021), forensic interviewing (Huang and Bull, 2020), and forensic psychology (Dror and Murrie, 2018). HEP covers three perspectives of decision making: first, reliability vs biasability; second, observations vs conclusions; and third, differences between (among and across) experts vs within experts (same expert, same evidence, at different times) (see Fig. 1).

Reliability and biasability are two fundamental properties of decision making. Reliability refers to the consistency, reproducibility, or repeatability of decisions, i.e., would the same observations and conclusions be made on the same evidence. Biasability refers to the effects of task-irrelevant contextual information and other biases (see eight different sources of bias, Dror, 2020) that can impact observations and conclusions. For example, knowing that the suspect was arrested, that they confessed or even their race may influence the experts' judgements and decisions. Biasability can be explored by assessing whether such contextual information impacts decision making. Reliability in decision making can be explored by assessing whether decisions are consistent when the decision-makers analyse the same evidence with the same information basis.

Reliability should not be confused with validity. The DF examiners may produce consistent results, but this does not necessarily mean that they are valid, since a reliable tool or a repeatable process may perfectly well produce consistent, but invalid, results. While reliability is concerned with consistency, validity is about the results being correct (Christensen et al., 2014). These concepts are connected, since if there is no reliability, there are no consistent results, and thus one cannot even consider validity (if the ground truth is unknown). Only when results are consistent, then one can examine whether they are valid, thus, reliability is a prerequisite for considering validity.

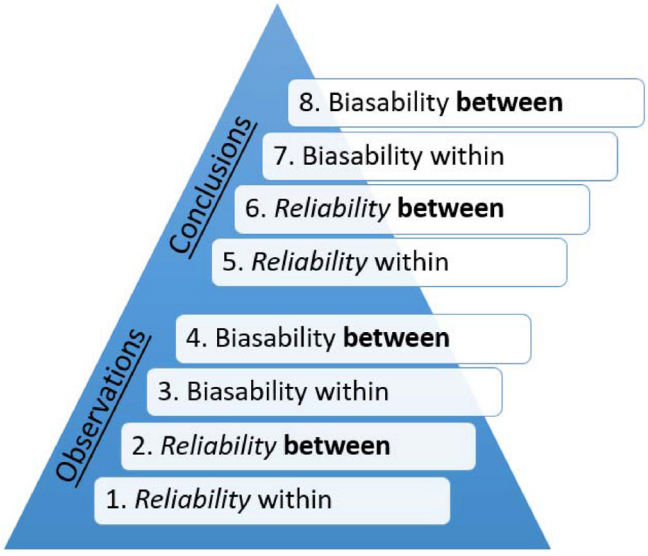

Fig. 1. Hierarchy of expert performance (Dror, 2016).

Reliability and biasability can be examined as per the conclusions reached (see levels 5–8 of Fig. 1), that is, whether they are reliable and are impacted by bias. However, expert decisions are underpinned by the observations, the perception of what the data are. HEP takes apart these different decision making components, which are often mixed (see levels 1–4 of Fig. 1). For example, differences in decisions may be incorrectly attributed to different ways of interpreting and concluding, when the different decisions actually emerge from inconsistencies in the observations of data. All of these comparisons and perspectives can be examined between examiners, or within the same examiner at different times.

While some of DF work is about presenting observed traces (observation) or conducting and presenting conclusions with formal evaluations of evidence in the light of propositions (conclusions), the result of DF work often involves presenting interpretations of the observed evidence/traces (what they mean) (Tart et al., 2019). These interpretations are inferences based on observations, measurements, or experimental testing where the DF examiner starts "connecting the dots", and prepares the ground for a conclusion. Concerning HEP, such interpretations of observed evidence are somewhere between observations and conclusions, which is between HEP levels 4 and 5 in Fig. 1. All three levels; observation, interpretation of observation, and conclusion were explored in the current study.

1.2. Contextual information and decision making

Bias may originate from many sources, and one of these is contextual information. For biasability, contextual information can be divided into two categories: *task-relevant* and *task-irrelevant.* To minimize contextual bias, task-irrelevant information should be excluded from the forensic decision making process. Task-relevant information is necessary for forensic decision making, but should - if possible - be managed and introduced at the right time to prevent it from biasing the decision making processes (see, for example, the Linear Sequential Unmasking (LSU) procedure, Dror, 2020). Due to the biasing nature of contextual information, the forensic examiner should be transparent about what they knew and at which point in time while conducting their forensic work.

Classifying what is task-relevant and task-irrelevant can be difficult, and could also vary from case to case. However, some information would be clearly task-irrelevant, such as whether the suspect is arrested, whether the suspect has confessed, and whether the police detective thinks they are guilty or innocent.

We conducted a background investigation to explore how DF work was commissioned to ascertain whether DF examiner would normally be exposed to contextual information. Research into forensic fingerprinting has shown that examiners are exposed to task-irrelevant contextual information through the submission forms, such as whether the suspect has previous criminal convictions (Gardner et al., 2019). We extend this study to the DF domain and collected 30 submission forms from different DF units/organisations from Europe and the United States.

The forms we obtained included a variety of information, such as names of suspects, victims, gender, addresses, information about the seized device, and the place/situation from which it was collected. Also, 22 (73%) of the forms had a section where the commissioning party was requested to provide 'information about the case' (or comparable phrasings). None of the forms explicitly stated not to provide task-irrelevant information or provided instruction about what information should be included (or not included) in this section. The 'information about the case' was often

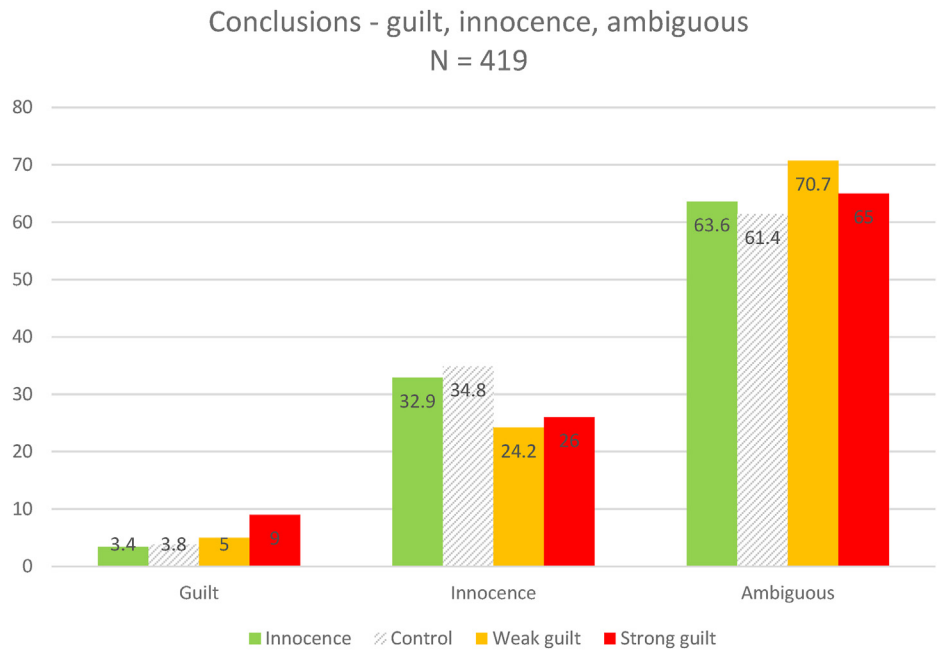

Fig. 2. Proportion of guilt, innocence, and ambiguous ratings for 10 traces by the groups.

a large portion of the form, allocating much writing space (and sometimes noting to add addition if the space was not sufficient).

Eleven (over a third) of the DF units informed us that instead of, or in addition to, using a form they would also directly communicate with the commissioning party to clarify the task and 'expectations'. Thus, the submission form seemed to function as a starting point for further dialogue concerning the task and introduced no restrictions or limitations about the information exchange. The pre-study suggest that it is common to provide contextual information when DF work is commissioned and that some of this information may be task-irrelevant.

The aim of the study reported below was twofold. First, to explore whether contextual information biases the DF decision making, and second, whether the DF examiners are consistent in their decision making. Specifically, the research questions we address in this study are:

1. Are DF examiners biased by contextual information when making observations, interpretations of observations, or in their conclusions during the analysis of digital traces?
2. Are DF examiners consistent with one another when making observations, interpretations of observations, or conclusions during the analysis of digital traces?

Concerning HEP, the current study specifically examined the reliability and biasability between examiners at observation and conclusion, levels 2, 4, 6, and 8 in HEP (see Fig. 1). To address the first research question concerned with biasability in DF decision making, we gave the same evidence file to the examiners and manipulated if and what contextual information was given. A control group received no contextual information, and the other groups received contextual information suggesting either strong guilt, weak guilt/ambiguous, or innocence. The strong guilt context was clearly task-irrelevant (the suspect has confessed to the crime), while the other contexts contained both task-relevant and irrelevant information (see Appendix 1 for details). We first examined whether the contextual information biased the DF examiner's observations (HEP 4). Second, we explored whether interpretations of the observations were biased by the contextual information. And third, we examined whether their conclusions were impacted by the contextual information (HEP level 8).

To address the second part of the study concerning reliability in DF decision making, we first examined how consistent DF examiners who examined the same evidence file were (within the same contextual information), in their observations of information (HEP level 2). Second, we explored how consistent they were in their interpretations of the observed information, and third, how consistent they were in their conclusions (HEP level 6).

2. Method

2.1. Participants

An invitation letter was sent out to DF examiners through the European Union Cybercrime Task Force (EUCTF) network, Interpol digital forensics expert network, as well as DF departments within the Norwegian Police. The invitation specified that the participants should be qualified and experienced with DF work and that analysis of digital evidence in criminal investigations should be their main professional task. The participants were informed that the research aimed at gaining knowledge about factors that may increase the quality of DF investigations and the reliability of digital evidence, and that the purpose of the experiment was to explore to what degree that participants arrived at similar results, and how the results were documented.

This provided us with 65 consenting participants. Nine did not complete the experiment, and 3 were excluded from the sample (2 due to the lack of sufficient expertise in performing analysis of evidence files, and 1 participant due to not being a practitioner). The data from the remaining 53 participants (82% of the initial participants) were included in the analysis.

The DF examiners came from 8 different countries: Norway (44), India (2), UK (2), Denmark (1), Finland (1) The Netherlands (1), Kenya (1), Canada (1), among these 9 women and 44 men. In terms of organisational level, 20 worked at a national/state level, 31 at a local level, and 2 in a private corporation/independent unit. Thirty-two had a civil educational background, 18 had a police educational background, and 3 had both. As per educational level, 1 had a Ph.D.,

15 had an MSc, 34 had a BSc, and 3 had other education without a degree.

In terms of experience with criminal investigation within law enforcement, the overall distribution was: 0–1 years: 32%, 2–4 years: 25%, 5–9 years: 32%, 10–15 years: 5%, above 15 years: 5%. In terms of experience with DF within law enforcement the overall distribution was: 0–1 years: 28%, 2–4 years: 32%, 5–9 years: 32%, 10–15 years: 8%, and above 15 years: 0%).

The main processing/analysis tool used during the experiment was according to the participants: Magnet Axiom: 53%, X-Ways: 28% EnCase: 6%, Autopsy 6%, Forensic Toolkit (FTK) 4%, The Sleuth kit (TSK): 2%, Other: 2%.

To achieve a good distribution of the background variables across experimental conditions, participants were quasi-randomly assigned to groups (Shadish et al., 2002). The participants were first matched in groups of four with participants with similar educational background and level, organisational level, and experience within law enforcement. They were then randomly assigned to an experimental condition. The demographics per group are shown in Table 1.

2.2. Material

2.2.1. Evidence file

To generate data for the experiment, the participants were tasked with analysing an evidence file. We used the evidence file from Digital Corpora (Garfinkel et al., 2009). The file was approximately 3 GB, and in E01 format (nps-2008-jean.E01 and nps-2008-jean.E02). The evidence file had an NTFS file system and Microsoft Windows XP operating system, and the timestamps in the file were from 2008. The owner was defined as Jean User, and contained the typical programs and files that would be present on a work computer, such as programs for handling documents, spreadsheets or presentations, e-mail, chat, internet browsing, etc.

2.2.2. Scenarios

A scenario explaining an incident of confidential information leakage was provided to the participants (see Appendix 1). Participants in the experimental groups received additional contextual information strongly indicating guilt (referred to as the Strong Guilt group), ambiguous and weak indications of guilt (referred to as Weak Guilt group), and innocence (referred to as the Innocence group). Also, we had a control group that only received the scenario, and no additional contextual information (Control group).

2.3. Procedure

After signing and returning an information letter and a consent form, the participants were given a background survey (Part 1 survey). They were free to choose the day for completing the experiment. They were allowed to work at their normal workstation, and with the processing/analysis software of their own choice. They were instructed to work alone, and not to consult colleagues during the experiment.

The day before the experiment day, they were given the link for downloading the evidence file, and they were informed that they could process the file in advance, but not start on the analysis until they had received information and templates the next morning. On the experiment day, they received the scenario and templates for the log and analysis reports on e-mail.

The participants received different contextual information, but this was not visible from the document title – which all contained the title “READ THIS description of the experiment”. They were reminded that they should work alone, and not communicate with anyone about the experiment. They were asked to make notes in the log while analysing, and the template contained fields for actions, justification/purpose, result, whether the result was bookmarked, and assessment of the result. They were also asked to note the starting time and completion time in the log, and if they were interrupted, also the periods they worked on the experiment.

The participants were asked to reserve 4–5 h on the analysis and writing the report. The analysis report template contained the heading ‘Analysis report’, and a line for the name of the report author and date. They were instructed to bookmark files of interest and export them to a pdf and return this document by e-mail together with the analysis report and log when the experiment was completed. After receiving these documents from the participants, they were provided with the final survey (Part 2 survey). In this survey, they were provided with a list of information that was located on the evidence file (see Appendix 2, A1, A3-A11), and asked whether they identified the particular trace and if so, to assess whether they thought it indicated guilt, innocence or was ambiguous for the suspect Jean. They were also asked if they identified traces in addition to the list that indicated guilt/innocence/ambiguity for Jean, with the same alternatives as described above.

The output of the experiment included in this article was the following:

- Part 1 Survey: Background information
- Analysis reports

Table 1
Demographics – percentage* distribution of background variables per group.

Demographics		Control N = 16	Strong Guilt N = 12	Weak Guilt N = 13	Innocent N = 12
Country*	Norway	75	83	85	92
	Other	25	17	15	8
Gender*	Men N = 44	81	100	85	67
	Women N = 9	19	0	15	33
Educational background*	Civil	63	58	54	66
	Police/both	38	42	46	33
Educational level*	PhD	6	0	0	0
	MSc	25	33	23	33
	BSc	69	58	69	58
	Lower	0	8	8	8
Organisational level*	National	31	25	54	42
	Local	63	67	46	58
	Private industry	6	8	0	0
Mean years of experience within law enforcement	With crim. inv.	5	6	5	6
	With DF	5	4	3	4

- Part 2 Survey: About how the analysis was performed and assessment of the findings

2.4. Analysis and quality control

The analysis was done in IBM SPSS Statistics Release 26.0.0.0.64 bit edition.

The observations of traces (see section 3.1.1 and Appendix 2) were coded in dichotomous variables "found" and "not found". The interpretations of observed traces (see section 3.1.2 and Appendix 2) were coded in dichotomous variables "corresponding interpretation" and "non-corresponding interpretation". The conclusions (see section 3.1.3 and Appendix 3) were coded as "not found", "guilt", "innocence" or "ambiguous".

Kruskal–Wallis test was used for the analysis of biasability (the difference between groups). Kruskal–Wallis is a non-parametric rank-based test used for small sample sizes which do not meet the assumption of a normal distribution and tests whether there is a significant difference between groups (Skovlund and Fenstad, 2001). The reported results are the H statistic, degrees of freedom, and the p-value; e.g. $H(3) = 9.438$, $p = .024$. Mean ranks are the average of the ranks for all observations within each sample, and are used for calculating the statistic (H). Mean ranks should be interpreted as when a group's mean rank is higher than the overall average rank, the observation values in that group tend to be higher than those of the other groups.

Krippendorff's Alpha coefficient was used for the analysis of reliability (the variation within groups) and was chosen as measurement since it is a generalisation of several known reliability indices (Krippendorff, 2011). It and enables the judgement of a variety of data with the same reliability standard, and applies to any number of observers, any number of categories, scale value or measures, any level of measurement, large or small sample sizes, and datasets with incomplete or missing data (Krippendorff, 2011). The method can also be used for where multiple coders are coding data (Nili et al., 2017). Values of $\alpha = 0.80$ or higher are considered a strong level of consistency, while values lower than $\alpha = .667$ are considered an inadequate level of consistency (Hayes and Krippendorff, 2007).

Krippendorff's Alpha coefficient is often used to assess agreement among coders in qualitative analysis and was used for the assessment of inter-coder agreement in the current study. The background information from Part 1 and assessment of findings from Part 2 (Section 3.1.3) was multiple-choice and required only minimal interpretation. The observations of traces (A) (Section 3.1.1) and interpretations of observed traces (B) (Section 3.1.2), required some interpretation and were coded as dichotomous variables (values: identified – not identified). To assess inter-coder agreement, 10% of the reports (A and B) were coded by a person who was not engaged with the study. Krippendorff's Alpha coefficient (Hayes and Krippendorff, 2007) was computed, with the result of $\alpha = 0.91$, which is considered a strong level of agreement (Hayes and Krippendorff, 2007).

3. Results and discussion of biasability and reliability

3.1. Biasability in DF examinations

The first research question was to explore whether DF examiners were biased by contextual information in their observations, interpretations of observations, or conclusions. We gave the same evidence file and baseline scenario to the participants, but different contextual information (indicating strong guilt, weak guilt, or innocence – see Appendix 1). A control group received only the baseline scenario and no contextual information. The results with regards to observations are presented in section 3.1.1, interpretations of observations in section 3.1.2, and conclusion in section 3.1.3. The results concerning biasability are discussed in section 3.2.

3.1.1. Biasability in observations of evidence (A)

The evidence file provided to the DF examiners was relatively small (3 GB), but there was a lot of information to explore. The DF examiners documented relevant findings in the analysis report, and 11 different traces were selected for comparison of the proportion of observations. Some of the traces were easier to find on the evidence file, such as e-mail content, documents, and chat. Other traces would require more in-depth examination, such as e-mail headers or discovering the mounting of USB. The traces A1-A11 are not very complex, and finding these traces would require basic DF skills.

None of the participants observed all 11 traces (A). Overall, fourteen (26%) of the participants found 1–4 traces, 35 (66%) found 5–8 traces and 4 (8%) found 8–10 traces. When comparing the groups, higher numbers were observed in the Guilt groups, then the Control group, and the lowest in the Innocence group. Specifically, the Weak Guilt group had the highest number of observed traces per participant ($M = 6.9$, $SD = 2.0$, range 4–10), followed by the Strong Guilt group ($M = 5.7$, $SD = 1.2$, range 3–7), the Control group ($M = 5.4$, $SD = 1.7$, range 3–9) and the Innocence group ($M = 4.5$, $SD = 1.8$, range 1–7).

Table 2 shows the proportion of participants per group that observed the trace, and a detailed description of the traces is provided in Appendix 2. The mean and SD refer to the proportions of observed traces within each of the groups. The possible number of observations was 145.75 per group (11 observations x 53 participants/4 groups = 145.75). To compare the number of observations between groups, they were adjusted for group size (Number of observations x .63 (Control group), x .83 (Strong Guilt group), x .77 (Weak Guilt group), x .83 (Innocence group).

The Innocence group observed the least number of traces, indicating that they were biased to find less evidence. The Guilt groups had the highest number of traces, indicating that they were biased to find more evidence. The Control group was between the Innocence and Guilt groups. However, there was very little difference between the Strong Guilt group and the Control group in the proportion of observations, which indicates that the Strong Guilt context (the suspect had confessed) did not bias the participants to observe more traces than an examination without such context. The Weak Guilt group observed significantly more traces, suggesting that the ambiguous Weak guilt context (wage conflict where the suspect had taken side with the workers) biased the Weak Guilt group to find more traces. These differences were statistically significant ($p < .05$) as reflected by a Kruskal–Wallis test between groups: $H(3) = 9.438$, $p = .024$, with mean ranks of 36.62 for the Weak Guilt group, 28.08 for the Strong Guilt group, 24.84 for the Control group, and 18.38 for the Innocence group.

Summarised, a statistically significant effect from contextual information was observed on the proportion of observations. The Weak Guilt context led to most observations, and Innocence context led to fewer observations. The Strong guilt context did not have a significant effect on observations compared to the Control group.

3.1.2. Biasability in interpretations of observed evidence (B)

During an examination of digital evidence, DF examiners use their expertise to interpret and derive meaning from the observed traces. These interpretations are presented in analysis reports as various forms of opinion statements, which in this paper are

Table 2
Proportion of DF examiners per group that observed the traces (A).

Observations of traces	Control (N = 16)	Strong Guilt (N = 12)	Weak Guilt (N = 13)	Innocence (N = 12)	Total (N = 53)	Range
(A1)	.69	1.00	.85	.50	.76	.50
(A2)	.94	1.00	.92	.92	.94	.08
(A3)	.19	.33	.23	.08	.21	.25
(A4)	1.00	.92	1.00	.83	.94	.17
(A5)	.44	.58	.62	.25	.47	.37
(A6)	.88	.92	1.00	.92	.93	.12
(A7)	.38	.08	.54	.00	.26	.54
(A8)	.19	.00	.39	.00	.15	.39
(A9)	.19	.08	.31	.08	.17	.23
(A10)	.25	.33	.39	.42	.34	.17
(A11)	.25	.42	.69	.50	.45	.44
Mean	.49	.52	.63	.41	.51	
SD	.32	.39	.28	.36	.32	
Total observations	86	68	90	54		
Observations adjusted for group size	54	56	69	45		

referred to as interpretations of observed traces, or just interpretations. Examples from the current study were e.g. “m57plan.xls was not available in the image file” or “the two spreadsheets had similar information but different file names”. These are often not the final conclusion, but would often underpin or justify the final conclusion. Seven different interpretations of observed traces (B) were selected for comparison and verified by a DF expert with no role in the study. If the information was missing, incorrect, or incomplete, it was coded as a “non-corresponding interpretation”. A detailed description of the criteria for being coded as a corresponding interpretation is provided in Appendix 2.

The possible number of interpretations was 92.75 per group (7 interpretations x 53 participants/4 groups = 92.75). To compare the number of observations between groups, they were adjusted for group size (Number of interpretations x .63 (Control group), x .83 (Strong Guilt group), x .77 (Weak Guilt group), x .83 (Innocence group), see Table 3 for details. The groups were comparable and the variation between them was quite small. A Kruskal–Wallis test performed at the total of observations showed no significant effect ($p < .05$) from contextual information between groups: $H(3) = 0.502$, $p = .918$.

3.1.3. Biasability in conclusions about guilt/innocence/ambiguity (HEP 8)

After writing the report from the analysis and submitting the result, the participants were provided with a survey about traces (A) from the experiment. They were asked to assess whether they identified the trace, and if so, whether it indicated guilt, innocence, or was ambiguous for the suspect. Traces that was identified by a minimum of 35% of the DF examiners were included here: A1, A3-A11, see detailed description in Appendix 2 and 3.

Of a total of 529 ratings of the traces, 110 were rated as “not found”, and 419 were rated as indications of guilt, innocence or ambiguous. Among these 419 ratings, there were 22 ratings of guilt (Control group = 5, Strong Guilt group = 9, Innocence group = 3, Weak Guilt group = 5), 125 ratings of innocence (Control group = 46, Strong guilt group = 26, Innocence group = 29, Weak Guilt group = 24), and 272 ratings of ambiguous (Control group = 81, Strong Guilt group = 65, Innocence group = 56, Weak Guilt group = 70). The proportion of “not found” among the groups was: Control group = 28 (17.5%), Strong Guilt group = 20 (16.7%), innocence group = 33 (27.3%), Weak Guilt group = 29 (22.7%).

Fig. 2 shows the percentage of ratings of guilt, innocence and ambiguous within those who found the evidence, and the value “not found” is excluded. A Kruskal–Wallis test performed at the total of conclusions showed no statistically significant effect ($p < .05$) for contextual information between groups, Not found: $H(3) = 3.182$, $p = .364$, Guilt: $H(3) = 3.357$, $p = .316$, Innocence: $H(3) = 3.261$, $p = .353$, Ambiguous: $H(3) = 0.914$, $p = .822$.

3.2. Discussion of biasability

The research question related to biasability was: *Are DF examiners biased by contextual information when they make observations, interpretations of observations, or in their conclusions during the analysis of digital traces?*

The results show that the observations were biased by the contextual information, where more traces were observed when Guilt contexts were introduced, and fewer traces were observed when Innocence context was introduced. The examination of biasability at the interpretation and conclusion level did not provide statistically significant results.

Table 3
Proportion of DF examiners per group that interpreted the traces correctly (B).

Interpretations of observed traces	Control (N = 16)	Strong Guilt (N = 12)	Weak Guilt (N = 13)	Innocence (N = 12)	Total (N = 53)	Range
(B1)	.19	.42	.08	.33	.26	.34
(B2)	.19	.33	.23	.0	.19	.33
(B3)	.38	.42	.31	.25	.34	.17
(B4)	.88	.92	1.00	.92	.93	.12
(B5)	.25	.17	.15	.50	.27	.35
(B6)	.19	.08	.31	.08	.17	.23
(B7)	.25	.33	.46	.33	.34	.21
Mean	.33	.38	.36	.34	.36	
SD	.25	.27	.31	.30	.26	
Total interpretations	37	32	33	29		
Interpretations adjusted for group size	23	27	27	22		

The proportion of observations was obtained from the participants' analysis reports. The traces that were not reported may have not been observed or may have been observed and overlooked or explained away as not relevant.

The results suggest that a DF examiner who believes that the suspect is innocent, observes fewer traces, and thus may have less information for developing explanations about what may have happened, including scenarios involving that the suspect may have contributed to or committed the crime. If the traces are overlooked or explained away due to an innocence bias, a possible consequence may be that important traces are not reported, or might be described and framed in a way that fits with the innocence hypothesis. The following is an illustrative example of such an 'innocence framing' of the findings in the conclusion: "It is very likely that M57. biz was exposed to a phishing attack and that the police hypothesis about Jean being innocent, and rather was framed, is correct" (P-31 – translation from Norwegian).

In the same manner, the results suggest that a DF examiner who believes that the suspect may be guilty, observe more traces, and thus obtain more information basis. The following is an example of 'guilt framing' of the findings in the conclusion: "The employees involved are Alison SMITH, manager of M57. biz and Jean JONES, CFO of M57. biz. Additionally, e-mail communication would also suggest that tuckgorge@gmail.com also has an involvement in the leakage of confidential information" (P-62).

The proportion of observed traces according to the survey (Appendix 3) is higher than what was the proportion that was documented in the analysis reports (Table 2). This difference indicates that the traces may have been overlooked or explained away rather than that they actually were not discovered, and sheds light on how the DF examiners contribute to the construction of the criminal case by including and excluding information in their analysis reports.

The DF examiners were allowed to use processing/analysis tools of their own choice. Processing tools may display the information differently. Therefore, we explored whether there was a relationship between the processing tool and the observations, but no relationship was found. The findings of biased observations uncovered in this study are probably not related to what the processing tool displayed or not, but rather about what the DF examiner looked for and perceived as relevant.

A covariate that may have played a role in the observations, interpretations, or conclusions is the data itself (see Dror, 2020). The evidence file contains a lot of "rich" information that may have impacted the DF examiners during the analysis, such as e-mail and instant messaging communication.

Another critical point, often overlooked, is how bias may impact the results by influencing testing strategies (Dror, 2020). An important aspect of this, demonstrated in our study, is how bias may influence whether the examiner ends or continues the further examination. This has been observed in other forensic domains (Dror, 2020), but not in DF. As detailed above, the data showed that when Guilt contexts were received, and particularly the ambiguous Weak Guilt context, the DF examiners continued to look for and observe more traces, in contrast to when Innocence context was received, where the examiners stopped looking for traces much earlier and hence observed fewer traces.

The findings concerning biasability at the observation level shed light on the point at which bias starts affecting the outcome. This knowledge is critical for designing effective bias mitigation measures for DF work, intending to minimize human error.

3.3. Reliability in DF examinations

The second research question was to explore whether DF examiners were consistent when they make observations, interpretations of observations, or conclusions during the analysis of digital traces.

A calculation of Krippendorff's Alpha coefficient (Hayes and Krippendorff, 2007) was performed to measure the consistency within each group, where the participants had analysed the same evidence file based on the same contextual information. According to Hayes and Krippendroff (2007), values of $\alpha = 0.80$ or higher are considered a strong level of consistency, while values lower than $\alpha = 0.667$ are considered an inadequate level of consistency.

For observation and conclusions of traces (A) and interpretations of observed traces (B), the results suggest overall low/inadequate reliability ($\alpha < 0.667$) (see Table 4). The highest reliability score was found at the observation level in observation of traces (A) for examiners receiving Strong Guilt context ($\alpha = 0.51$) and Innocence context ($\alpha = 0.44$).

The most extreme within-group variation at the conclusion level is the variation from rating a trace as an indication of guilt vs an indication of innocence (see Appendix 3). This variation was observed for all groups, and highest for Strong Guilt (40%, A3, A4, A5, and A6) followed by Control (30%, A1, A4, and A6). For Weak Guilt and Innocence, the variation between guilt and innocence ratings was observed in 20% of the evidence (A1 and A6).

3.4. Discussion of reliability

The research question related to reliability was: *Are DF examiners consistent with one another when they make observations, interpretations of observations, or conclusions during the analysis of digital traces?*

The results suggest what is considered a low/inadequate ($\alpha < 0.667$) consistency between DF examiners who examined the same evidence file based on the same contextual information. The low consistency was found on all the examined levels: observations of traces, interpretations of observed traces, and conclusions. The results suggest that if a DF examiner performs an analysis of an evidence file, and another DF examiner would do a re-analysis of the same evidence file, the chances of reaching consistent results are low. This finding is of great importance, since it indicates, quite contrary to the perception of digital evidence as objective and credible, that digital evidence may be affected by the human examiner in all the stages we have examined – during the observation, the interpretation of observation and when forming a conclusion. The low reliability would thus not only have implications for reporting opinion evidence but also for factual/technical reporting.

Although high reliability between DF examiners is anticipated, it is important to be aware that consistency does not imply accuracy or validity. Consistency may arise from a variety of reasons, for example when examiners are consistently biased by similar biasing contextual information. If the irrelevant contextual information biases the DF examiners in the wrong direction, they may perform similar and consistent observations, and all reach the same incorrect conclusions. In such a case, the consensus is reached for the wrong reasons. Consistency is thus no proof of validity.

This entails that quality measures should not only be directed towards the tools and technology, but also the human. It is not possible to calibrate a human the same way as a technical

Table 4
Krippendorff's α coefficient of reliability for observations of traces, interpretations of observed traces, and conclusions.

Reliability	Control N = 16 N = 132*	Strong Guilt N = 12 N = 100*	Weak Guilt N = 13 N = 99*	Innocence N = 12 N = 88*
Observations (A)	.35	.51	.26	.44
Interpretations of observations (B)	.20	.20	.30	.30
Conclusions (A)*	.26	.22	.20	.30

instrument, but measures such as blind proficiency testing through the use of fake test cases may provide knowledge about human performance (Garrett, 2021). Quality control measures such as verification reviews and re-examination of the evidence (Horsman and Sunde, 2020; Sunde and Horsman, 2020) may be important measures to detect possible erroneous or misleading results.

4. General discussion

To the best of our knowledge, this is the first study to explore reliability and biasability in DF decision making. The paper seeks to advance the current knowledge about the potential sources of error in digital forensic investigations. The study aimed at examining whether contextual information biases the DF examiners' observations, interpretations of observations, and conclusions during the examination of a dataset, and to further explore whether DF examiners were consistent in their observations, interpretations of observations, and conclusions when they analysed the same dataset with the same contextual information.

The results showing low reliability are probably the most important findings in this study. The study shows low reliability between experts examining the same evidence file with the same contextual information, in their observations, interpretation of observations, and their conclusions. The results thus indicate a low probability for a second DF examiner observing and interpreting the same traces, and concluding consistently compared to the initial DF examiner (given the same evidence file and contextual information). The study contributes to a growing body of knowledge about the lack of reliability in forensic decision making (e.g. Ulery et al., 2012; Lidén and Dror, 2020), and highlights the need for systematic implementation of quality control measures such as blind peer review (Horsman and Sunde, 2020; Sunde and Horsman, 2020) in DF examinations.

In terms of biasability, the pre-study suggests that when DF work is commissioned, it is common to provide contextual information to the DF examiner and that some of this information may be task-irrelevant. The current study showed that DF examiner observations were affected by the biasing contextual information.

Our results are consistent with basic research on confirmation bias (Nickerson, 1998). The current study contributes to the growing volume of studies indicating that forensic science decision making can be affected by contextual information (e.g., Elaad et al., 1994; Cooper and Meterko, 2019) and that confessions may affect the collection and evaluation of evidence (Hasel and Kassin, 2009; Kassin et al., 2012; Kukucka and Kassin, 2014).

The findings regarding biasability and reliability suggest that human factors have a significant impact on the outcome of a DF examination and that the results of DF examinations should not be presumed to be objective and credible. It is thus necessary to gain more knowledge on how to minimize and/or detect bias in DF decision making.

To minimize bias during forensic work, excluding or blinding of irrelevant contextual information has been identified as a possible solution for several forensic science domains. This would also be an important measure in DF examination, however, most likely it would not totally solve the contextual bias challenge. This is due to the nature of digital evidence, which normally contains lots of both task-relevant and irrelevant information within the dataset under investigation which cannot be easily excluded. A bias countermeasure could be that the starting point of every DF examination is a balanced set of alternative hypotheses, where both innocence and guilt hypotheses are represented, that the DF examiner systematically considers all the hypotheses during the analysis, and report the results accordingly. Transparency about what they knew (what task-relevant and task-irrelevant information they received), and which hypotheses guided the examination are of great importance to facilitate a review of the process and result, and enable detection of biased decision making.

In terms of ecological validity, the experimental setting was designed to be as close as possible to the real work situation of the DF examiner. The pre-study of submission forms indicated that it is common to provide contextual information when DF work is commissioned and that some of this information may be task-irrelevant. The contextual information provided to the DF examiners in the current study would thus be within what would be a normal situation encountered in actual casework. The DF examiners were allowed to use their normal workstation, and the analysis software of their preference, and could report the results as they would do in their casework.

There were some limitations in the experimental setting that should be commented on. A possible limitation of the study was the sample size of 53 DF examiners, which may limit the statistical power. The time frame may also have been a limitation, and a few commented that 4–5 h may have been less time than they normally would have spent on the analysis of an evidence file. However, the evidence that was observed (A) or interpreted (B) by at least 31% of a group was included in the further analysis of biasability and reliability. Since the study -as most studies- used volunteer participation, there is also a risk of self-selection bias. Since they knew they were taking part in research, a Hawthorne-effect (e.g. Leedy and Omerod, 2014) may have occurred, however – we suggest this would rather impact the results in a "positive direction" in terms of evidence observation and interpretation of the observed evidence, for example, that they examined the evidence file more thoroughly and accurately than they would normally do, and thus observed more traces.

In terms of future directions for research, it is important to explore how bias could be minimized during DF work. To minimize potentially flawed results from DF examinations, there seems to be a need for quality assurance and control, which also calls for more research on effective measures (see e.g. Mattijssen et al., 2020; Sunde and Horsman, 2020). The results related to biasability

underpins a need for research on contextual bias and context management. There is a need to gain more knowledge about what information has the most biasing effects, and which contextual information is task-relevant vs task-irrelevant. With this knowledge, it is possible to design context management procedures for DF that ensure access to the task-relevant contextual information, while excluding or limiting access to task-irrelevant contextual information.

5. Conclusion

The study explored the reliability and biasability in DF decision making and involved 53 DF examiners who analysed the same evidence file. The aim was to examine whether contextual information biases the DF examiners' observations, interpretations, or conclusions and whether DF examiners make consistent observations, interpretations, or conclusions when they examine the same evidence file with the same contextual information.

The results showed that the DF examiners' observations were affected by the biasing contextual information. To minimize bias, it is important to ensure that DF examinations are based on task-relevant contextual information, and minimize the exposure to task-irrelevant contextual information. The results also showed low reliability between DF examiners in theirobservations, interpretations of observations, and conclusions. The low reliability indicates that there is a serious and urgent need for quality assurance in DF examinations and control of DF results to prevent erroneous results from cascading into the investigation process.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the DF examiners who dedicated some of their valuable time and participated in this study, and also Torstein Schjerven and Jørn Helge Jahren for their contribution to pilot testing of the quasi-experiment. We are also very grateful for the support from Håvard Aanes on the statistical measurements used in this study and for the contribution to the control of inter-coder reliability from Christine S. Nordsletten and Torstein Eidet. We thank Helene O. I. Gundhus from the University of Oslo, and Fergus T. Toolan, Johanne Yttri Dahl, and Gunnar Thomassen from the Norwegian Police University College for insights and constructive criticism of former versions of the paper. We also thank the anonymous reviewers for their constructive critique and helpful comments which have improved the paper.

This paper was funded by a grant from the Norwegian Police University College awarded to the first author for pursuing a PhD.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fsidi.2021.301175.

APPENDIX 1. Case vignettes

Confidential information leakage

M57. biz is a small US based company, with office in your country. The company, which develops and sells body art equipment (tattoo, piercing etc.), is in the start-up phase. The manager for the M57. biz office in your country is Alison Smith, and the CFO is Jean Jones. The company has 4 programmers, 2 in marketing, and 1 in business development. Only Alison and Jean have a permanent office space, while the other employees work from home office. All employees participate in a daily online meeting. There are in-person meetings for all employees in the M57. biz office once every two weeks. Most documents are exchanged by e-mail.

A spreadsheet (m57plan.xls) containing confidential information was recently posted as an attachment in a forum of a competitor's website. When this was discovered, Alison reported the incident to the police as information theft. Alison told police that Jean, the CFO, was responsible for updating the spreadsheet, and that it was probably sent from Jean's computer.

The attachment posted on the competitor's website looked like this:

M57.biz company				
Name		Position	Salary	SSN (for background check)
Alison	Smith	President	$140,000	103-44-3134
Jean	Jones	CFO	$120,000	432-34-6432
Programmers:				
Bob	Blackman	Apps 1	90,000	493-46-3329
Carol	Canfred	Apps 2	110,000	894-33-4560
Dave	Daubert	Q&A	67,000	331-95-1020
Emmy	Arlington	Entry Level	57,000	404-98-4079
Marketing				
Gina	Tangers	Creative 1	80,000	980-97-3311
Harris	Jenkins	G & C	105,000	887-33-5532
BizDev				
Indy	Counterching	Outreach	240,000	123-45-6789
Annual Salaries			$1,009,000	
Benefits		30%	$302,700	

You are tasked with analysing a copy of the hard drive from Jeans computer, and find out:

What has happened, and what was Jeans involvement in the reported incident?

Participants other than the Control Group received one of the following additional narratives:

(Strong Guilt):

Jean was arrested for information theft, and in a police interview earlier today, she confessed that she had committed the criminal offense.

(Weak Guilt):

Alison told police that there has been a long-lasting wage dispute with the programmers in the firm, who claim to be underpaid. One of the programmers - Emmy Tuckford Arlington – has fronted the dispute on behalf of the programmers in M57. biz. Jean has supported the programmers in this conflict, and has told Alison that the company can afford to pay them better salaries. Jean is about to be interviewed by the police about the reported incident. However, the chief investigating officer wants an analysis of Jean's computer before the police interview, to look for traces indicating that she was involved in the reported incident.

(Innocence):

As a result, Jean was arrested for information theft and questioned about the incident in a police interview. However, after the police interview, the police believe she is innocent, and that she was framed during a phishing attack.

APPENDIX 2. Observed traces and interpretations of traces

Observed traces (A) (coding: found/not found)

1. Spreadsheet with the filename m57biz.xls stored in the Desktop folder. It contained information similar to what was leaked.
2. Spreadsheet with the filename m57biz.xls located in the outlook.pst file. It contained information similar to what was leaked.
3. AIM chat between alison57 and Jean User. There are indications of that someone might be impersonating Alison – the boss is referred to as a male.
4. Mail correspondence between jean@m57.biz and Alison@m57.biz. The address is used by the boss – Alison, which asks Jean to create a spreadsheet with information about the employees.
5. Mail correspondence between Jean@m57.biz and Alex@m57.biz. The address is also used by the boss – Alison, with information relevant to the information leakage.
6. Mail correspondence between Jean@m57.biz and Tuckgorge@gmail.com, outside the firm M57. biz e-mail domain. Tuckgorge@gmail.com is used together with the display name Alison@m57.biz, and is recipient of the mail attachment m57biz.xls.
7. Mail correspondence between Jean@m57.biz and Bob@m57.biz, who asks about the leaked information.
8. Mail correspondence between Jean@m57.biz and Carol@m57.biz, who comments that she have noticed that something is wrong.
9. Trace of mounting of a USB (type, model), which may contain relevant traces or information about the incident.
10. Several user accounts had been activated, including the Administrator account.
11. Several user accounts had been activated, including the Devon user account.

Valid interpretations of observed traces (B) (coding: corresponding interpretation/non-corresponding interpretation)

1. Excel spreadsheet m57plan.xls was not on the evidence file.
2. That excel spreadsheets m57plan.xls and m57biz.xls were different/did not contain the exact same information.
3. That excel spreadsheets m57biz.xls located in Desktop folder and in outlook. pst had the same hash value.
4. That the mail with the attachment m57biz.xls was sent to tuckgorge@gmail.com and not alison@m57.biz (which was only the display name in e-mail that was replied to).
5. That simsong@xy.dreamhostps.com was defined as return path in the header of the e-mail message where the spreadsheet with confidential information was requested. This e-mail message was replied to and the spreadsheet m57biz.xls was enclosed.
6. That the USB was mounted shortly before the m57biz.xls was sent out of M57. biz domain.
7. That other user accounts (Devon and/or Administrator) than the Jean user account were active around the time when the m57biz.xls was sent.

APPENDIX 3. Conclusions of whether the trace indicated guilt/innocence/ambiguous

(Coding: not found, guilt, innocence, ambiguous)

Trace	N = Control (16) Strong guilt (12) Innocence (12) Weak guilt (13)	% of N stated that they identified the trace	Rating: Guilt*	Rating: Innocence*	Rating: Ambiguous*
A1	Control	100%	2 (13%)	1 (6%)	13 (81%)
	Strong guilt	100%	5 (42%)	0 (0%)	7 (58%)
	Innocence	92%	1 (9%)	2 (18%)	8 (73%)
	Weak guilt	100%	3 (23%)	1 (8%)	9 (69%)
	Mean	98%			
A3	Control	81%	0 (0%)	3 (23%)	10 (77%)
	Strong guilt	67%	1 (13%)	1 (13%)	6 (75%)
	Innocence	67%	0 (0%)	1 (13%)	7 (88%)
	Weak guilt	54%	0 (0%)	0 (0%)	7 (100%)
	Mean	67%			
A4	Control	94%	1 (7%)	7 (47%)	7 (47%)
	Strong guilt	100%	1 (8%)	7 (58%)	4 (33%)
	Innocence	83%	0 (0%)	6 (60%)	4 (40%)
	Weak guilt	100%	0 (0%)	6 (46%)	7 (54%)
	Mean	94%			
A5	Control	87%	0 (0%)	6 (43%)	8 (57%)
	Strong guilt	100%	1 (8%)	5 (42%)	6 (50%)
	Innocence	83%	0 (0%)	4 (40%)	6 (60%)
	Weak guilt	100%	0 (0%)	3 (23%)	10 (77%)
	Mean	93%			
A6	Control	100%	2 (13%)	13 (81%)	1 (6%)
	Strong guilt	100%	1 (8%)	8 (67%)	3 (25%)
	Innocence	92%	1 (9%)	10 (91%)	0 (0%)
	Weak guilt	100%	2 (15%)	8 (62%)	3 (23%)
	Mean	98%			
A7	Control	94%	0 (0%)	9 (60%)	6 (40%)
	Strong guilt	100%	0 (0%)	3 (25%)	9 (75%)
	Innocence	83%	1 (10%)	4 (40%)	5 (50%)
	Weak guilt	92%	0 (0%)	2 (17%)	10 (83%)
	Mean	92%			
A8	Control	87%	0 (0%)	6 (43%)	8 (57%)
	Strong guilt	83%	0 (0%)	2 (20%)	8 (80%)
	Innocence	67%	0 (0%)	2 (25%)	6 (75%)
	Weak guilt	85%	0 (0%)	4 (36%)	7 (64%)
	Mean	81%			
A9	Control	62%	0 (0%)	1 (10%)	9 (90%)
	Strong guilt	50%	0 (0%)	0 (0%)	6 (100%)
	Innocence	50%	0 (0%)	0 (0%)	6 (100%)
	Weak guilt	46%	0 (0%)	0 (0%)	6 (100%)
	Mean	52%			
A10	Control	56%	0 (0%)	0 (0%)	9 (100%)
	Strong guilt	67%	0 (0%)	0 (0%)	8 (100%)
	Innocence	54%	0 (0%)	0 (0%)	6 (100%)
	Weak guilt	38%	0 (0%)	0 (0%)	5 (100%)
	Mean	54%			
A11	Control	62%	0 (0%)	0 (0%)	10 (100%)
	Strong guilt	67%	0 (0%)	0 (0%)	8 (100%)
	Innocence	50%	0 (0%)	0 (0%)	6 (100%)
	Weak guilt	61%	0 (0%)	0 (0%)	8 (100%)
	Mean	60%			

*% of those who found the trace.

References

Casey, E., 2002. Error, uncertainty and loss in digital evidence. International Journal of Digital Evidence 1 (2).

Christensen, A.M., Crowder, C.M., Ousley, S.D., Houck, M.M., 2014. Error and its meaning in forensic science. J. Forensic Sci. 59 (1), 123–126. https://doi.org/10.1111/1556-4029.12275.

Cooper, G.S., Meterko, V., 2019. Cognitive bias research in forensic science: a systematic review. Forensic Sci. Int. 297, 35–46. doi:10.1016/j.forsciint.2019.01.016.

Dror, I.E., 2016. A hierarchy of expert performance. Journal of Applied Research in Memory and Cognition 5 (2), 121–127. https://doi.org/10.1016/j.jarmac.2016.03.001.

Dror, I.E., 2020. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. Anal. Chem. 92 (12), 7998–8004. https://doi.org/10.1021/acs.analchem.0c00704.

Dror, I.E., Hampikian, G., 2011. Subjectivity and bias in forensic DNA mixture interpretation. Sci. Justice 51 (4), 204–208. https://doi.org/10.1016/j.scijus.2011.08.004.

Dror, I.E., Murrie, D., 2018. A Hierarchy of Expert Performance (HEP) applied to forensic psychological assessments. Psychol. Publ. Pol. Law 24 (1), 11–23. https://doi.org/10.1037/law0000140.

Dror, I.E., Charlton, D., Péron, A.E., 2006. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci. Int. 156 (1), 74–78. https://doi.org/10.1016/j.forsciint.2005.10.017.

Elaad, E., Ginton, A., Ben-Shakhar, G., 1994. The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. J. Behav. Decis. Making 7 (4), 279–292. https://doi.org/10.1002/bdm.3960070405.

Gardner, B.O., Kelley, S., Murrie, D.C., Blaisdell, K.N., 2019. Do evidence submission forms expose latent print examiners to task-irrelevant information? Forensic Sci. Int. 297, 236–242. https://doi.org/10.1016/j.forsciint.2019.01.048.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing Science to Digital Forensics with Standardized Forensic Corpora, DFRWS 2009, Montreal, Canada. Available at: https://simson.net/clips/academic/2009.DFRWS.Corpora.pdf. (Accessed 20 April 2020).

Garrett, B.L., 2021. Autopsy of a Crime Lab. University of California Press.

Garrett, B.L., Neufeld, P.J., 2009. Invalid forensic science testimony and wrongful convictions. Va. Law Rev. 95 (1), 1–97.

Hartley, S., Winburn, A.P., 2021. A Hierarchy of Expert Performance (HEP) applied to forensic anthropology. J. Forensic Sci. (in press).

Hasel, L.E., Kassin, S.M., 2009. On the presumption of evidentiary independence: can confessions corrupt eyewitness identifications? Psychol. Sci. 20 (1), 122–126. https://doi.org/10.1111/j.1467-9280.2008.02262.x.

Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. 1 (1), 77–89. https://doi.org/10.1080/19312450709336664.

Horsman, G., Sunde, N., 2020. Part 1: the Need for peer review in digital forensics. Forensic Sci. Int.: Digit. Invest. 35, 301062. https://doi.org/10.1016/j.fsidi.2020.301062.

Huang, C., Bull, R., 2020. Applying Hierarchy of Expert Performance (HEP) to investigative interview evaluation: strengths, challenges and future directions. Psychiatr. Psychol. Law. https://doi.org/10.1080/13218719.2020.1770634.

James, J.I., Gladyshev, P., 2013. A survey of digital forensic investigator decision processes and measurement of decisions based on enhanced preview. Digit. Invest. 10 (2), 148–157. doi:10.1016/j.diin.2013.04.005.

Kassin, S.M., Bogart, D., Kerner, J., 2012. Confessions that corrupt: evidence from the DNA exoneration case files. Psychol. Sci. 23 (1), 41–45. https://doi.org/10.1177/0956797611422918.

Krippendorff, K., 2011. Computing Krippendorff's Alpha-Reliability. http://repository.upenn.edu/asc_papers/43.

Kukucka, J., Kassin, S.M., 2014. Do confessions taint perceptions of handwriting evidence? An empirical test of the forensic confirmation bias. Law Hum. Behav. 38 (3), 256–270. https://doi.org/10.1037/lhb0000066.

Leedy, P.D., Ormrod, J.E., 2014. Practical Research: Planning and Design. Pearson Education.

Lidén, M., Dror, I.E., 2020. Expert reliability in legal proceedings: "Eeny, meeny, miny, moe, with which expert should we go?". Sci. Justice. https://doi.org/10.1016/j.scijus.2020.09.006.

Mattijssen, E.J., Witteman, C.L., Berger, C.E., Brand, N.W., Stoel, R.D., 2020. Validity and reliability of forensic firearm examiners. Forensic Sci. Int. 307, 110112. https://doi.org/10.1016/j.forsciint.2019.110112.

Nickerson, R.S., 1998. Confirmation bias: a ubiquitous phenomenon in many guises. Rev. Gen. Psychol. 2 (2), 175–220.

Nili, A., Tate, M., Barros, A., 2017. A critical analysis of inter-coder reliability methods in information systems research. In: Australasian Conference on Information Systems. Hobart, Australia.

Page, H., Horsman, G., Sarna, A., Foster, J., 2019. A review of quality procedures in the UK forensic sciences: what can the field of digital forensics learn? Sci. Justice 59 (1), 83–92. https://doi.org/10.1016/j.scijus.2018.09.006.

Pollitt, M., Casey, E., Jaquet-Chiffelle, D.O., Gladyshev, P., 2018. A Framework for Harmonizing Forensic Science Practices and Digital/multimedia Evidence. The Organization of Scientific Area Committees for Forensic Science (OSAC), USA.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Wadsworth Cengage Learning.

Skovlund, E., Fenstad, G.U., 2001. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J. Clin. Epidemiol. 54 (1), 86–92. https://doi.org/10.1016/s0895-4356(00)00264-x.

Sunde, N., Dror, I.E., 2019. Cognitive and human factors in digital forensics: problems, challenges, and the way forward. Digit. Invest. 29, 101–108. https://doi.org/10.1016/j.diin.2019.03.011.

Sunde, N., Horsman, G., 2020. Part 2: the phase-oriented advice and review structure (PARS) for digital forensic investigations. Forensic Sci. Int.: Digit. Invest. 36, 301074. https://doi.org/10.1016/j.fsidi.2020.301074.

Tart, M., Pope, S., Baldwin, D., Bird, R., 2019. Cell site analysis: roles and interpretation. Sci. Justice 59 (5), 558–564. https://doi.org/10.1016/j.scijus.2019.06.005.

Ulery, B.T., Hicklin, R.A., Buscaglia, J., Roberts, M.A., 2012. Repeatability and reproducibility of decisions by latent fingerprint examiners. PLoS 7, e32800. https://doi.org/10.1371/journal.pone.0032800.

van Baar, R.B., van Beek, H.M.A., van Eijk, E.J., 2014. Digital forensics as a service: a game changer. Digit. Invest. 11, 54–62. https://doi.org/10.1016/j.diin.2014.03.007.

van Beek, H.M.A., van Eijk, E.J., van Baar, R.B., Ugen, M., Bodde, J.N.C., Siemelink, A.J., 2015. Digital forensics as a service: game on. Digit. Invest. 15, 20–38. https://doi.org/10.1016/j.diin.2015.07.004.

van Beek, H.M.A., van den Bos, J., Boztas, A., van Eijk, E.J., Schramp, R., Ugen, M., 2020. Digital forensics as a service: stepping up the game. Forensic Sci. Int.: Digit. Invest. 35, 301021. https://doi.org/10.1016/j.fsidi.2020.301021.

Watkins, K., McWhorte, M., Long, J., Hill, B., 2009. Teleporter: an analytically and forensically sound duplicate transfer system. Digit. Invest. 6, 43–47. https://doi.org/10.1016/j.diin.2009.06.012.

Wilson-Kovacs, D., 2019. Effective resource management in digital forensics. Policing: Int. J. 43 (1), 77–90. doi:10.1108/pijpsm-07-2019-0126.