Cognitive biases in the peer review of bullet and cartridge case comparison casework: A field study

Erwin J.A.T. Mattijssen[a,b], Cilia L.M. Witteman[b],
Charles E.H. Berger[a,c], Reinoud D. Stoel[a]

[a] Netherlands Forensic Institute, PO Box 24044, 2490 AA The Hague, The Netherlands
[b] Radboud University Nijmegen, Behavioural Science Institute, PO Box 9104, 6500 HE Nijmegen, The Netherlands
[c] Leiden University, Institute for Criminal Law and Criminology, PO Box 9520, 2300 RA Leiden, The Netherlands

Abstract
Objective: Forensic judgments and their peer review are often the result of human assessment and are thus subjective and prone to bias. This study examined whether bias affects forensic peer review.
Hypotheses: We hypothesized that the probability of disagreement between two forensic examiners about the proposed conclusion would be higher with "blind" peer review (reviewer saw only the first examiner's comparison photos) than with "non-blind" peer review (reviewer also saw the first examiner's interpretation and proposed conclusion). We also hypothesized that examiners with a higher perceived professional status would have a larger effect on the reported conclusion than examiners with a lower status.
Method: We acquired data during a non-blind and a blind peer review procedure in a naturalistic, covert study with eight examiners (3–26 years of experience). We acquired 97 conclusions of bullet and cartridge case comparisons in the blind and 471 in the non-blind peer review procedure.
Results: The odds of disagreement between examiners about the evidential strength of a comparison were approximately five times larger (95%-CI [3.06, 8.50]) in the blind than in the non-blind procedure, with disagreement about 42.3% and 12.5% of the proposed conclusions, respectively. Also, the odds that their proposed conclusion was reported as the final conclusion were approximately 2.5 higher for the higher-status examiners than for lower-status examiners.
Conclusions: Our results support both the hypothesis that bias occurs during non-blind forensic peer review and the hypothesis that higher-status examiners determine the outcome of a discussion more than lower-status examiners. We conclude that blind peer review may reduce the probability of bias and that status effects have an impact on the peer reviewing process.

Keywords: Firearm examination; Decision making; Cognitive bias; Peer review; Verification; Forensic comparison.

1. Introduction

1.1. Forensic sciences

The forensic science disciplines provide scientific evidence to be used in a court of law. The courts tend to view the testimonies of forensic expert witnesses as objective [1,2] and generally do not doubt them. In many forensic science disciplines, however, a human decision maker, the forensic examiner, is the main instrument of analysis and interpretation [3]. The forensic examiner compares visual patterns and provides a judgment about the sources of these patterns. These judgments are subjective [4,5] and their validity is questioned by other scientists [6–8].

1.2. Cognitive biases

The critical attitude towards the validity of examiner judgments is understandable when we consider that when humans make judgments under uncertainty they tend to resort to heuristics to simplify complex tasks [9]. The resulting efficiency gains can come at the cost of errors resulting from cognitive bias. In the forensic sciences, most attention goes to the effects of confirmation bias and contextual bias (e.g., [10,11]). Confirmation bias can be explained as the tendency to seek out and interpret information in accordance with one's pre-existing beliefs or as the tendency to retain, or an averseness to abandon, a favored hypothesis (for a review see [12]). Both research studies (for a review see [13]) and real-world casework [14] have shown that confirmation bias influences forensic decision making. Contextual bias is the tendency to develop expectations about the outcome of an examination and to draw conclusions guided by task-irrelevant context information (e.g., [15]). The implementation of context information management is proposed in order to minimize the risk of these biases (e.g., [2,16]). This could include for example the management of case information [17,18] and (linear) sequential unmasking, where the evidence is not examined simultaneously with the reference material but before examination of, and comparison with, the reference material [19,20].

Anchoring could be another bias in forensic work. It is related to confirmation bias but with a more restricted focus on numerical judgments (e.g., probability or likelihood judgments). The anchoring-and-adjustment mechanism explains the phenomenon that adjustment from a pre-existing anchor may be insufficient [9,21]. This anchoring effect has been shown to be remarkably robust (for a review see [22]), manifesting itself independently of the relevance of the anchor, the expertise of the decision maker, the decision maker's motivation, and whether the judgment is made in a laboratory or real-world situation, including judicial decision making (e.g., [23,24]). Consequently, it is likely that it also occurs in forensic casework.

1.3. Forensic peer review

Forensic peer review (also referred to as verification [25]) is a procedure during which a second examiner reviews the outcomes of the examination of another examiner. This includes the comparison results, interpretation, proposed conclusion and draft report. The aim of peer review in forensic casework is to improve validity of conclusions and to prevent errors [26–

29]. In this paper, when we talk about 'valid conclusions', we refer to conclusions that are warranted by the available evidence, the procedures used, and the forensic examiners' knowledge.

Peer review is considered to be optimal if all conclusions are reviewed (both same-source and different-source conclusions) and if the second examiner does not know the conclusion of the first examiner and can formulate an independent conclusion [29]. This is called 'blind' peer review or 'blind' verification (e.g., [2,30]). This form of peer review results in an independent interpretation of the same comparison results, where the focus of peer review lies on the interpretation phase. This is different from a completely independent 'doubling' of examinations, where two examiners perform all phases of the examination independently (including the comparison phase) and not only the peer review phase.

When forensic peer review is applied in a non-blind manner, the second examiner gets all intermediate outcomes (results of the comparison, interpretation, and proposed conclusion) from the first examiner. In this procedure confirmation or anchoring bias could occur. We consider the interpretation and resulting conclusion proposed by the first examiner as an anchor, which the second examiner may adjust insufficiently while formulating their proposed conclusion. This would result in sub-optimal peer review, with too small a contribution from the second examiner, and in fewer disagreements about the conclusion between the two examiners.

1.4. Examiner status

Both the non-blind and the blind peer review procedures (discussed in detail in Section 2.2.2) assume that all examiners are equally competent in carrying out comparisons and that their considerations carry equal weight in a discussion following a disagreement. Yet examiners may have different roles. At the Netherlands Forensic Institute a distinction is made between two types of firearm examiners: reporting and non-reporting examiners. The reporting examiners are qualified to perform comparisons, to perform shooting scene reconstructions and other ballistic examinations, to sign reports, testify in court, and they can be held accountable for their reports and testimony. The non-reporting examiners are equally qualified to perform comparisons, but they are not allowed to perform any of the other tasks.

Such differences in roles and accountability can result in differences in perceived professional status even for tasks where equal competence is assumed. In discussions, higher-status group members tend to generally have more influence on the group's conclusions than lower-status group members [31,32]. This effect could be explained by for instance the findings that in a discussion the considerations of higher-status members are more easily accepted by lower-status group members than the other way around [33,34], and that group members seem to underestimate (and overestimate) the performance of lower-status (and higher-status) group members [35]. When peer review is affected by differences in perceived professional status this could result in a suboptimal discussion, with unequal contributions of the examiners.

1.5. Overview of current study and hypotheses

Although it is commonly accepted that peer review has added value, this is corroborated by only a few studies [25]. To our knowledge there are no studies that compared different forensic peer review procedures. The current study examined the effect of different peer review procedures and of differences in perceived professional status on proposed conclusions. We did this in a real-world setting, during forensic peer review in casework.

1.5.1. Peer review procedure effect

We hypothesized that bias occurs in non-blind forensic peer review, when the second examiner knows the interpretation and proposed conclusion of the first examiner before giving their own judgment. We expected that in non-blind peer review the second examiner's proposed conclusion would be biased towards that of the first examiner. We expected that in blind peer review disagreement about the proposed conclusion (when two examiners proposed different conclusions) would occur more often than in the non-blind procedure. This difference would also be discernable for the judged evidential strength (see Section 2.2.1 for more information). These differences between the two procedures in the proportion and degree of disagreement between examiners might be attributed to confirmation and anchoring bias.

1.5.2. Status effect

We hypothesized that when there is a discussion between a reporting and a non-reporting examiner about the evidential strength of the result of a comparison, the examiner with the higher perceived status (the reporting examiner) would have a larger effect on the outcome. We consequently expected that the conclusion proposed by the reporting examiner would more often be reported as the final conclusion than that of the non-reporting examiner.

2. Method

2.1. Participants

We collected the data using real-world forensic bullet and cartridge case comparison casework with eight forensic examiners from the firearms section of the Netherlands Forensic Institute. Seven of these examiners only worked on firearms related cases and one also worked on tool mark cases. There were three women and five men, with their experience as a forensic firearm examiner ranging from approximately three to 26 years at the start of data acquisition (M = 16, SD = 9, Mdn = 19). In the group of eight participants there were five reporting and three non-reporting examiners. Whether an examiner was a reporting or non-reporting examiner was determined by the description, and corresponding salary, of the job for which they had applied. Both types of examiners had varying degrees of education, ranging from secondary vocational education to higher professional education for the reporting examiners and from pre-university education to higher professional education for the non-reporting examiners. All examiners had followed the same internal training to learn to compare bullets and cartridge cases, including working on approximately 90 comparison cases under supervision during a period of three years or longer and they were thus equally

qualified to perform these comparisons. The non-reporting examiners had on average less years of experience, ranging from three to eight years, than the reporting examiners, whose experience ranged from 16 to 26 years. Although the non-reporting examiners had less experience, they had overall performed more comparisons per year than the reporting examiners, as that was their sole job. The two groups may therefore be considered to have comparable experience and competence carrying out comparisons, as is indeed assumed by our standard operating procedures.

2.2. Materials and measures

2.2.1. Comparison conclusions

To test our hypotheses we considered the conclusions proposed by the first and second examiner during peer review, and the reported conclusion. Such a conclusion represents a judgment of the evidential strength of the outcomes of for example a bullet comparison. To reach such a conclusion, firearm examiners will compare the striation patterns (features that are acquired when a bullet is fired through a barrel and are caused by imperfections in the barrel's interior) in a questioned bullet with those in (an)other fired bullet(s) found at the crime scene or in reference bullets fired with a firearm submitted for examination. Firearm examiners compare these striation patterns using a comparison microscope, which enables them to view both bullets at the same time. The result of this phase is a judgment of the degree of similarity of striation patterns of the bullets. The examiners combine the judged degree of similarity with an assessment of how distinctive these striation patterns are, to judge how likely it is to find this degree of simi larity when the bullets are fired with either the same or with different firearms. The ratio of these two judged likelihoods is the evidential strength, expressed as a likelihood ratio [36]. More specifically, this likelihood ratio (LR) is determined by dividing the judged probability of the findings of the examination (E) given one hypothesis (H1: the bullet was fired with the seized firearm) by the judged probability of the findings given an alternative hypothesis (H2: the bullet was fired with another firearm):

$$LR = \frac{p(E|H1)}{p(E|H2)} \quad (1)$$

The forensic firearm examiners at the Netherlands Forensic Institute provide a numerical judgment of the likelihood ratio of the results of a comparison. Because these likelihood ratios are not calculated but judged, the examiners report a range of likelihood ratios to avoid an impression of precision, given the subjective character of these judgments. They report this range as a verbal expression chosen from a scale, where each verbal expression is directly defined by ranges of numerical likelihood ratios (Table 1). An example of such a conclusion is: The results of the comparison are 'more probable' (between 10 and a 100 times more) if the bullet found at the crime scene was fired with the seized firearm than if it was fired with some other firearm. The same scale is used when the results of a comparison provide support for the same-source hypothesis (H1) as when they provide support for the different-source hypothesis (H2). When necessary, the examiners can also combine verbal expressions to indicate a broader range of evidential strength. An 'approximately equally

probable' conclusion (judged likelihood ratio of 1–2) means that the results provide no support for either the same-source or the different-source proposition. When the first and the second examiner agree on the interpretation of the range of numerical evidential strength and would thus report the same verbal expression(s), they report the conclusion and do not discuss it any further. When the second examiner does not agree with the conclusion proposed by the first examiner, they will discuss their respective considerations.

[Table 1]

2.2.2. Peer review procedures

When this study started, the firearms section of our institute had a peer review procedure in place in which all casework was reviewed. Following this procedure, a second examiner received the complete case file including the results of the comparison, the written interpretation and the proposed conclusion of the first examiner (non-blind peer review). The non-blind procedure consisted of 3 phases.

The blind peer review procedure (without the interpretation and resulting proposed conclusion) that was implemented later consisted of the same phases but with an additional phase (Phase 2), an independent interpretation and proposed conclusion by the second examiner (see Fig. 1 for a schematic representation of all phases). During Phase 2 of the blind peer review the second examiner only received the comparison photos and was asked to give their own interpretation and to provide a judgment on evidential strength, recording their interpretation and conclusion. Only after this additional Phase 2 was performed did the second examiner receive the complete case file from the first examiner. The second examiner then completed the remaining steps of Phase 3 and 4.

The 4 phases of the blind peer review procedure are:

- *Phase 1.* The first examiner performs the comparison of the markings present in the evidential material using a comparison microscope (e.g., comparison of the striation patterns in a bullet from a crime scene with the striations in bullets fired with a seized firearm). Digital photos are taken of the striation patterns (the results), and their comparison leads to a judged degree of similarity. The first examiner interprets those judged degrees of similarity and provides a judgment about their evidential strength, the proposed conclusion.
- *Phase 2.* In the blind peer review procedure, the second examiner interprets the comparison photos and provides a judgment about their evidential strength, the proposed conclusion. The non-blind peer review procedure skips Phase 2.
- *Phase 3.* The first examiner provides the complete case file, including the comparison photos, the interpretation and the proposed conclusion, to a second examiner. The second examiner performs a peer review.
- *Phase 4.* The second examiner provides feedback to the first examiner after peer review. When the second examiner does not agree with the interpretation and conclusion proposed by the first examiner this will lead to a discussion. During such a discussion, the reported final conclusion is decided upon by the two examiners.

2.3. Procedure

We acquired the data for this study in a real-world setting, during forensic casework, by means of a naturalistic, covert study. It was not possible to randomize the cases amongst the participants and procedures because we wanted to keep the participants unaware of the real goal of the study. To acquire the data we introduced a feedback form for each comparison conclusion in the case file, purportedly as a quality control measure shortly after the Netherlands Forensic Institute had defined the already used verbal expressions of evidential strength by numerical likelihood ratio ranges (Table 1). Our cover story was that the data would be used to monitor the consequences of these definitions for the reported conclusions. Both the first and second examiner filled out the same feedback form during Phase 1 and 3, respectively, and if disagreement occurred in Phase 4. For each of the three comparison conclusions (from the first examiner, the second examiner, and the reported conclusion) one feedback box such as shown in Fig. 2 was given. The examiners would first circle whether they had found support for a same-source (Hypothesis 1) or different-source (Hypothesis 2) conclusion or neither and they would subsequently circle the chosen numerical and corresponding verbal expression of evidential strength. The information provided on the feedback form had to be the same as their comparison conclusion(s) in the case notes. Most questions on the feedback form asked for information about the identity of the first and of the second examiner and the proposed preliminary and final conclusions. One question asked specifically about disagreement between the two examiners on the conclusion proposed by the first examiner. The question was (translated from Dutch): 'Does the second forensic examiner immediately agree with the proposed conclusion of the first forensic examiner?'. When the answer to this question was affirmative, the parts of the feedback form about the proposed conclusion of the second examiner and about the reported conclusion became redundant and did not have to be filled out.

[Figure 1]
[Figure 2]

We collected most of the data (55%) from the non-blind peer review procedure in a period of approximately eight months. After that period the first author formed a pilot group with three of the forensic examiners to construct and implement the blind peer review procedure. Those three examiners were not informed about the goals of this study; they participated with the sole goal of implementing a new procedure intended as a quality improvement based on the literature (e.g., [2,30]). After approximately three months the pilot group agreed that the blind peer review procedure could be implemented in the complete firearms section. During the following eleven months we collected data using the blind peer review procedure. To facilitate the pilot, the blind peer review procedure was only implemented for smaller cases (with a maximum of eight items of evidence). The pilot group made this choice because the additional time needed for Phase 2 of the blind peer review procedure depends on the number of items of evidence. By limiting our study to smaller cases the procedure could be tested efficiently. We expected no differences between smaller and

larger cases as the difference only concerned the number of shots fired and not the comparison procedures or the difficulty of the comparisons[1]. As a consequence of the three month implementation period during which only the pilot group tested the blind procedure and because the blind procedure was only implemented for smaller cases, we collected the remaining 45% of the data for the non-blind peer review procedure during the same fourteen month period as the blind peer review procedure data. The same feedback form was used to acquire data in both procedures. See Table 2 for an overview of the data collection and the number of examiners involved per period.

[Table 2]

3. Results

3.1. Data analysis

We performed descriptive statistics to provide information about the number of conclusions we acquired, what type of conclusions were reported, and how often and to what degree the first and second examiner disagreed about the proposed conclusion.

To address our hypothesis that bias would occur during non-blind forensic peer review we estimated three regression models. The first estimated regression model was for the probability of disagreement, the second for the degree of disagreement, where we considered the number of steps between the proposed conclusions of the two examiners on the conclusion scale (Table 1), and the third for the probability of changing the conclusion of the first examiner to the reported conclusion. We used generalized linear mixed models, specifying a binary logistic regression for both the occurrence of disagreement between examiners and changing the conclusion, and a linear regression model for the degree of disagreement. The three regression models contained the fixed effect of the peer review procedure, and a random effect of examiner pairs (thus taking into account that each examiner pair occurred multiple times).

To address the hypothesis that the examiner with the higher perceived status would have a larger effect on the outcome of a discussion we selected those comparisons of both the non-blind and blind procedure which fulfilled all of the following criteria ($N = 77$):

- there was disagreement about the proposed conclusion;
- the disagreement was between a reporting and a non-reporting examiner; and
- either the proposed conclusion of the reporting or that of the nonreporting examiner was reported as the final conclusion and not an alternative conclusion.

[1] After data acquisition we tested the assumption that comparison difficulty did not depend on case size. To do so, we considered the lower bound of the evidential strength of the reported conclusion as a proxy for comparison difficulty. The results of a Mann-Whitney *U* test showed that there were no significant differences in the evidential strengths of smaller ($N = 317$, $Mdn = 10000$, $M = 403241$, $SD = 899126$) and larger cases ($N = 251$, $Mdn = 10000$, $M = 344885$, $SD = 473921$; *Mann-Whitney-U* = 38455.5, $p = .481$). This indicates that indeed smaller and larger cases were equally difficult.

We first ranked the forensic examiners by the proportion of discussions following a disagreement in which the conclusion they had proposed became the finally reported conclusion. We then performed a one-sided, one-sample non-parametric binomial test to examine whether the conclusions proposed by reporting examiners were more often reported as the final conclusions than those of non-reporting examiners. Because there were multiple discussions per pair of examiners and because the number of discussions were not equally distributed over the possible examiner pairs we first aggregated the data, resulting in thirteen examiner pairs for the analysis. To explore whether possible status effects could also be explained by order effects (i.e., whether the reporting examiner was the first or the second examiner) we performed additional analyses. We included the same thirteen examiner pairs, but not all pairs showed both orders of examiners. This resulted in seven aggregated data points where the reporting examiner was the first examiner and thirteen data points where the non-reporting examiner was the first examiner. We performed a one-sample non-parametric binomial test on both groups separately (first examiner is a reporting examiner and first examiner is a non-reporting examiner).

3.2. Descriptive statistics

We acquired a total of 568 conclusions of comparisons for the two peer review procedures together. These conclusions refer to the conclusion for a single one-on-one comparison of for example two bullets or to the conclusion for a group of comparisons between for example five bullets. When a conclusion refers to a group of comparisons this means that the separate one-on-one comparisons within that group each provided the same degree of support for the same hypothesis. When the separate one-on-one comparisons in a group resulted in varying degrees of support or in support for different hypotheses, multiple comparison conclusions were reported. This strategy was chosen to compensate for dependencies between one-on-one comparisons of bullets or cartridge cases fired with the same firearm and similar availability of features. In 483 (85.0%) of these conclusions, support for the same-source hypothesis was reported, in 24 (4.2%) support for a different-source hypothesis, and in 61 (10.7%) no support for either the same-source or different-source hypothesis was reported. Of all these conclusions, there were 100 disagreements (17.6%) between a first and second examiner. There were no disagreements about contradicting source judgments (support for the same-source vs support for the different-source proposition), only about the evidential strength. In six out of the 100 disagreements however, one of the examiners did not find support for either a same-source or different-source conclusion (no source judgment) while the other examiner found support for a same-source conclusion. Of these six disagreements four resulted from the non-blind procedure and two from the blind procedure.

Disagreements on the evidential strength could be on the lower or upper bound of the numerical evidential strength or on both. For the analyses of the differences in judged evidential strength we considered the number of steps between the proposed conclusions of the two examiners on the conclusion scale used (Table 1). We used this number of steps instead of the actually judged numerical evidential strength because the upper bound of the strongest conclusion (“extremely more probable”) was not given and could be infinite, which is not usable in statistical analyses. When there was a disagreement, the number of steps

between the proposed conclusions of the two examiners ranged from 0 to 3 (M = 0.88, SD = 0.61) for the lower bound of the judged evidential strength and from 0 to 4 (M = 1.02, SD = 0.93) for the upper bound. Of the reported conclusions, 425 (74.8%) covered one evidential strength scale level (as shown in Table 1), 113 (19.9%) two levels, 26 (4.6%) three levels and 4 (0.7%) four levels.

When examiners discussed the disagreement between their proposed conclusions they could choose one of the two proposed evidential strengths as the final reported conclusion or to decide upon an alternative evidential strength to be reported. The proposed evidential strength of one of the examiners was reported in 83 out of the 100 disagreements following discussion. For the other 17 disagreements the lower and/or upper bound of the evidential strength of the reported conclusion differed from both proposed evidential strengths. Table 3 provides an overview of the decisions made to report either the lowest or highest proposed evidential strength following discussion, or an alternative, which can be a lower, intermediate or higher evidential strength than those proposed. We provide this information for both the lower and upper bound of the evidential strength as the disagreement could be about either one or both. Based on a visual exploration of the results the likelihood of reporting the lowest or highest proposed evidential strength seems approximately equal. When an alternative evidential strength is reported this is most often an evidential strength between the two proposed evidential strengths.

[Table 3]

3.3. Peer review procedure effect

3.3.1. Probability of disagreement

The frequency of disagreement between a first and second examiner on the proposed conclusion differed between the non-blind and blind peer review procedures. There was disagreement on 12.5% (59 out of 471) and 42.3% (41 out of 97) of the proposed conclusions in the non-blind and blind peer review procedures, respectively. The results of the generalized linear mixed model showed that the probability of disagreement was significantly larger for blind peer review than for non-blind peer review ($Exp(B)$ = 5.10; $p < .001$). The odds ratio ($Exp(B)$) indicates that the odds of disagreement were approximately 5 times larger in the blind peer review procedure than in the non-blind procedure (with a 95% confidence interval of [3.06, 8.50]).

3.3.2. Degree of disagreement

When considering the lower bounds of the proposed conclusions, the results of the generalized linear mixed model showed that the number of levels on the conclusion scale between the two examiners was 0.28 (with a 95% confidence interval of [0.19, 0.37]; d = 0.69) larger for blind peer review than for non-blind peer review (M_{blind} = 0.39, S_{blind} = 0.59; $M_{non\text{-}blind}$ = 0.11, $SD_{non\text{-}blind}$ = 0.36; t = 6.13, $p < .001$). The results also showed that the number of levels between the two examiners was 0.28 (with a 95% confidence interval of [0.16, 0.39]; d = 0.51) larger for blind peer review than for non-blind peer review (M_{blind} =

0.41, SD_{blind} = 0.77; $M_{non\text{-}blind}$ = 0.13, $SD_{non\text{-}blind}$ = 0.48; t = 4.53, p < .001) when considering the upper bounds of the proposed conclusions.

3.3.3. Probability of changing the conclusion

The frequency with which the conclusion changed from the proposed conclusion of the first examiner to the reported conclusion also differed between the non-blind and blind peer review procedures. Change occurred in 12.1% (57 out of 471) and 25.8% (25 out of 97) of the proposed conclusions in the non-blind and blind peer review procedures, respectively. The results of the generalized linear mixed model showed that the probability of changing the conclusion from the proposed conclusion of the first examiner to the reported conclusion was significantly larger for blind peer review than for non-blind peer review ($Exp(B)$ = 2.60; p = .001). The odds ratio ($Exp(B)$) indicates that the odds of changing the conclusion were approximately 2.5 times larger for the blind peer review procedure than for the non-blind procedure (with a 95% confidence interval of [1.50, 4.51]).

3.4. Status effect

3.4.1. Effect on the outcome of a discussion

Table 4 shows the ranking of the forensic examiners by the proportion of discussions, following a disagreement, in which their proposed conclusion became the finally reported conclusion. With the exception of one reporting examiner who was involved in only one discussion, there seemed to be a clear separation between the reporting and non-reporting examiners when looking at the percentages, where the reporting examiners ranked higher.

The results of the one-sided, one-sample non-parametric binomial test showed that the proportion of proposed conclusions by reporting examiners reported as the final conclusions was significantly higher (Mdn = 0.80, M = 0.71, SD = 0.35) than the equal test proportion of 0.5 ($t(13)$ = 2.22, p = .011). The observed proportions were higher for the reporting examiners (0.71, N = 55) than for the non-reporting examiners (0.29, N = 22). In other words, the odds that the proposed conclusion of an examiner was reported as the final conclusion following a discussion were approximately 2.5 higher for reporting than for non-reporting examiners.

[Table 4]

3.4.2. Order effect

The results of the one-sample non-parametric binomial tests showed that the proportion of proposed conclusions by reporting examiners reported as the final conclusions was significantly higher (Mdn = 0.83, M = 0.80, SD = 0.28) than the equal test proportion of 0.5 ($t(13)$ = 2.77, p = .003) when a non-reporting examiner was the first examiner. When a reporting examiner was the first examiner no significant difference from the equal test proportion of 0.5 was observed (Mdn = 0.50, M = 0.43, SD = 0.45).

4. Discussion

The result that there was no disagreement about contradicting source judgments (same-source vs different-source judgments) corresponds with findings in earlier studies (e.g., [37,38]). In those studies too, a high reliability of source judgments was found. The finding that there was disagreement about the degree of support in only 17.6% of the comparison conclusions is hard to relate to earlier findings in firearm comparison studies. Such studies typically focus on categorical conclusions (same-source, different-source or inconclusive judgments) and not on the judged evidential strength. Looking at this result solely in the context of this study, we believe that this is a fairly high degree of reliability of judgments. When there was disagreement about the proposed conclusion there did not seem to be a tendency to report either the lowest or highest proposed evidential strength.

4.1. Peer review procedure effect

The odds of disagreement between examiners were approximately five times larger in the blind than in the non-blind peer review procedure. The participant group or individual roles did not change during data acquisition and no other changes occurred, such as in applied procedures, worksheets or reporting templates. Therefore we may conclude that we found support for the hypothesis that bias occurs when the second examiner is first shown the interpretation and proposed conclusion of the first examiner (non-blind peer review). This result is in line with the results of earlier studies showing that confirmation bias and contextual bias occur in forensic decision making (for reviews see [2,13]). These results attest to the need to implement context information management in forensic decision making (e.g., [2,16]), which some forensic institutes have already shown to be feasible (e.g., [17,18,39]).

Furthermore we found that the degree of disagreement, measured as the number of steps between the proposed conclusions of the first and second examiner on the conclusion scale (Table 1), was larger for the blind than for the non-blind peer review procedure. This result corresponds with the results of earlier studies on anchoring (e.g., [9,40]). The main difference between the blind and non-blind peer review procedure was that in the blind procedure the second examiner first only received the comparison photos of the first examiner and could thus not be influenced by that examiner's interpretation and proposed conclusion. The finding that in that situation the second examiner more often disagreed with the conclusion proposed by the first examiner (42.3% vs 12.5%) showed that there was quite some latent disagreement that was not addressed when peer review was performed in a non-blind manner.

We emphasize that these disagreements only concerned the judged evidential strength and not the source judgments themselves. The examiners did not disagree on whether the evidence supported a same-source or different-source conclusion, for which examiners did not provide contradicting judgments. Providing judgments on the degree of support in the form of a likelihood ratio is fairly new. The results of this study show that there was between-subject variability in these judgments. On average the proposed conclusions of the first and second examiner differed by one step on the conclusion scale (Table 1). Future work would do well to also address the reliability and validity of evidential strength judgments.

The different role of the second examiner in the non-blind and blind procedure could also have affected the odds of disagreement. In the non-blind procedure the second examiner was expected to implicitly interpret the comparison photos and to formulate a proposed conclusion, and then to compare this to the first examiner's interpretation and proposed conclusion. In the blind procedure the second examiner explicitly wrote down their interpretation and conclusion, without first seeing the interpretation and proposed conclusion of the first examiner. This shift from implicit to explicit interpretation and conclusion may have affected the second examiners' perception of their role: from a check of soundness of another examiner's interpretation to performing a complete interpretation themselves. The current study design did not enable us to look into this effect in more depth. Future work may address the extent to which this change of role influences the likelihood of a disagreement.

We did not see the need to actively prevent discussion between examiners prior to Phase 4 of the procedure, since the standard operating procedures applied in the firearms section prescribe a serial execution of the phases. Therefore, discussion between examiners was not expected prior to Phase 4. Neither did we introduce a system that prevented the second examiner from changing their proposed conclusion prior to Phase 4 after learning of the first examiner's interpretation. We counted on the examiners' professional ethics to prevent them from doing this.

Apart from the observed difference in the odds of disagreement between the blind and the non-blind procedure we also saw an effect of the procedure on the reported conclusions. The results showed that the odds of changing the first examiner's proposed conclusion were approximately 2.5 times larger in the blind than in the non-blind peer review procedure. We recognize that this effect could have been a direct result of the larger odds of disagreement between examiners during blind peer review than during non-blind peer review. Even so, these results indicate that this blind peer review procedure not only affected the odds of a disagreement, but also had an impact on the reported conclusion.

4.2. Status effect

There was an overall difference between the reporting and non-reporting examiners. The reporting examiners tended to rank higher in the list of examiners whose proposed conclusion was also the finally reported one. At the group level, the odds that the proposed conclusion of a reporting examiner was also the final conclusion, following a discussion, were approximately 2.5 times larger than for a non-reporting examiner. Additionally, we explored the influence of examiner order on the outcome of a discussion. The results suggest that apart from differences in status the order of examiners also had an influence on whose proposed conclusion was reported. When a reporting examiner was the first examiner, the outcome of the discussion was approximately just as often their proposed conclusion as the second examiner's, and when a reporting examiner was the second examiner their proposed conclusion was reported more often. Future studies could look into this effect in more detail.

These results indicate an influence of the type of forensic examiner on the outcome of a discussion. Because of the design, where examiners were included in the study in their natural role, we could not randomize examiners to the role of reporting or non-reporting examiner. As a result, we cannot exclude that other factors besides type of examiner caused

the effect. However, our results are in line with previous studies and could well be explained by a difference in perceived professional status [31,32]. Because the reporting examiners had additional roles and accountability (they are allowed to sign reports, perform shooting scene reconstructions, and testify in court) and were compensated accordingly, they had a higher perceived professional status. This per- ceived difference in professional status also became apparent in a discussion in the firearms section of our institute following the presentation of the results.

The status effect in this study is in line with earlier findings. For example, during a discussion the considerations of higher-status members are more easily accepted by lower-status group members than the other way around [33,34]. In addition, group members seem to underestimate and overestimate the performance of lower and higher-status group members, respectively [35]. Also, lower-status group members seem reluctant to repeat themselves, while higher-status group members do not show this self-censorship [41]. Based on these findings, it can be expected that higher-status group members will put forward more of their considerations during discussion [42] and that their influence on the discussion will be higher than that of lower-status members [31,32,42].

Apart from differences in perceived professional status the outcome of a discussion might also in part be explained by potential differences in the quality of judgments between the two groups of examiners. Because the correct source and evidential strength judgments of a comparison are unknown in casework, it was not possible to test this possibility. The difference in quality of judgments was however expected to be small, because both examiner groups had followed the same internal training and were both similarly qualified to perform these comparisons.

The differences in influence on the outcome of a discussion could be detrimental for the aim of peer review to improve validity of conclusions and to prevent errors (e.g., [26]). If the considerations of non-reporting examiners are often overruled, even though the standard operating procedures assume that the examiners are equally competent in carrying out comparisons, this diminishes the added value of peer review. This effect might be minimized by implementing procedures in which the perceived lower-status group members are always the first to give their views. That would ensure that these views are at least verbalized and considered. Another possibility to minimize a status effect could be to implement a peer review procedure in which examiners are unaware of each other's identities. It would also be possible to opt for a "majority wins" approach, where an additional examiner interprets the evidence. This alternative seems most suitable for categorical conclusions as there is a limited number of possible proposed conclusions. However, when judging (lower and upper bounds of) evidential strengths there are far more possible outcomes. This increase in possible conclusions increases the likelihood of an additional examiner providing a new alternative as a proposed conclusion. Furthermore, such a "majority wins" approach could result in a situation where examiners no longer discuss the considerations underlying their proposed conclusions, which would decrease the potential to learn from each other.

4.3. Limitations and Future directions

We acquired the data for this study in a naturalistic, covert study to ensure high ecological validity. We kept the examiners unaware of what we were actually studying. This means that we could not randomly assign examiners to the reporting and non-reporting examiner groups. Furthermore, we could not randomize examiner pairs, reporting and non-reporting examiner order, or peer review procedure. Also, we could not randomly assign cases to examiners because current practice expects them to start working on the case with the nearest reporting deadline when they are available for casework. As a result, we could not ensure an equal contribution of each examiner to the acquired data. We also could not rule out individual influences, such as that some examiners might be more prone to disagree about proposed conclusions, and that some might have a larger influence on the outcome of a discussion as a result of personality traits. The small participant group, consisting of five reporting and three non-reporting examiners, adds to these potential internal validity threats. We have taken these threats into account in the analyses by including the random effect of examiner pair to study the peer review procedure effect, by aggregating the data on examiner pairs and by investigating the effect of examiner order to study the status effects. We must take care in generalizing the results of this study to a broader population of forensic (firearm) examiners because all participants were members of one existing firearms section.

The fact that the first author is an active member of the firearms section creates the threat that the examiners could have become aware of the goal of the study during the almost two years of data acquisition. We recognized this potential threat during the study design and told the examiners that the goal of the feedback forms was to monitor consequences of the newly introduced link between the numerical and verbal expression of evidential strength. We did not mention that we wished to study bias during peer review. We waited for eight months with the implementation of the blind peer review procedure after the feedback form was introduced. We did this to decrease the likelihood that the examiners would become aware of the link between the feed- back form and the goal of this study. Also, we asked each examiner the following two questions during the manuscript review process: 1) What did you think was the aim of the feedback forms?, and 2) Were you aware that one goal of the feedback forms was to study whether bias occurs during peer review? All eight examiners (including the three who were part of the pilot group to implement the blind peer review procedure) gave the same answer to the questions. They said they thought that the feedback form was implemented to acquire data about the interpretations and reported conclusions as a general quality control measure. They thus accepted our cover story. All eight examiners also said that they were not aware that one of the goals of the feedback forms was to study bias during peer review.

Because in forensic comparisons the judgments of evidential strength are subjective without a known ground truth, it is at this moment not possible to say whether the final conclusions resulting from blind forensic peer review were better than those resulting from non-blind peer review or not. Even so, the results provide evidence that when peer review is performed in a blind manner at the very least the considerations of two forensic examiners are more often discussed, following a disagreement. This is beneficial when the aim of peer review is to improve validity of conclusions and to prevent errors (e.g., [26]). With this aim

in mind, it is highly recommended to perform forensic peer review in a blind manner, where the second examiner does not get to see the interpretation and proposed conclusion of the first examiner. In addition, the higher probability of discussion will lead to a more frequent exchange of considerations, providing more opportunities to learn from each other's knowledge and experience.

Although our blind peer review procedure already reduced the potential for bias to occur during the process, there are still possibilities for further improvement. In our procedure, the second examiner was aware of the identity of the first examiner. As we showed, knowing the identity of the other examiner could influence the outcome of the procedure. Furthermore, in our procedure the second examiner only interpreted the first examiner's comparison photos, and did not perform an independent comparison of the features. This could in part have led to the observed high degree of reliability of source judgments, as both examiners interpreted the same results. Additionally, it could then happen that corresponding striations between for example two bullets, providing support for a same-source conclusion, are not found by the first examiner and only differences in striation patterns are observed and photographed. That would typically result in support for a different-source conclusion although the bullets could have been fired with the same firearm. As the second examiner only sees the comparison photos of the first examiner's comparison they will also not see the corresponding striation patterns. The second examiner would then endorse the different-source conclusion of the first examiner. Such effects could be avoided by implementing a procedure where two examiners both perform the comparison (Phase 1) and interpretation (Phase 2) independently, doubling of the examination. Such a procedure would theoretically be superior to our blind peer review of the comparison photos, but it would also take far more time. Our blind peer review procedure on average takes about 30 min more per case than non-blind review, and implementing an independent comparison phase could add hours. This means that in practice it is not feasible for our institute to work this way. A procedure in which the examiners are unaware of each other's identity and in which both the comparison and interpretation phase are performed independently would be ideal, and should result in even less bias in forensic casework. However, practical and time constraints may need to be overcome.

The studied non-blind and blind peer review procedures and the suggested doubling of the examination and interpretation can be compared with the ACE-V procedure. This procedure, which is predominantly applied in fingerprint examination but can also be applied in other disciplines, consists of an Analysis, Comparison, Evaluation and Verification phase [43]. Our non-blind peer review procedure is similar to the situation where the first examiner performs the analysis, comparison and evaluation phase (ACE) and a second examiner performs the verification (V) based on the outcomes of all ACE phases. In the blind peer review procedure the second examiner receives the outcomes of the analysis and comparison phase (AC), then performs an independent evaluation (E) and verifies (V) whether this second interpretation is in agreement with that of the first examiner. When doubling the examination two examiners would perform the analysis, comparison and evaluation phases independently (ACE + ACE) and would then verify whether their outcomes of all phases are in agreement (V). In our study we have deliberately chosen to use the term 'peer review' rather than 'verification' as used in the ACE-V procedure, because we consider 'peer review'

to be the more neutral term, which does not imply that the second examiner should agree with (verify) the first examiner's outcomes.

Another source of bias in forensic casework, not related to the peer review procedure, is the unequal number of same-source and different-source conclusions. Assuming that these judgments were valid this would mean that at least 85% of the comparison cases were same-source comparisons. Such base-rate information, where there are far more same-source than different-source comparisons, could create an undesirable expectation about the outcome of an examination (e.g., [15,44]). Examiners will more often expect a same-source comparison. A possible way to minimize bias resulting from this base-rate in- formation is to include fake cases in the normal flow of comparison cases to shift the base-rate towards 50–50. At the same time such fake cases could be used to gather information about the validity of forensic source judgments in casework circumstances. This can be done by comparing the reported support for a same-source or different-source conclusion to the known source of the samples in a fake case (e.g., [45,46]).

5. Conclusions

We found that examiners are more likely to disagree in the blind peer review procedure than when they see the other's interpretation and proposed conclusion. We also found that examiners with a higher perceived status (reporting examiners) have a larger effect on the outcome of a discussion than non-reporting examiners. To minimize the occurrence of both of these effects and to enhance the validity of examiner judgments we propose the implementation of a peer review procedure in which all conclusions are reviewed, where examiners do not see the other's judgment before they give their own, and where examiners do not know who the other examiner is.

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.

References

[1] I.E. Dror, S.M. Kassin, J. Kukucka, New application of psychology to law: Improving forensic evidence and expert witness contributions, J. Appl. Res. Memory Cognit. 2 (2013) 78–81.

[2] S.M. Kassin, I.E. Dror, J. Kukucka, The forensic confirmation bias: problems, perspectives, and proposed solutions, J. Appl. Res. Memory Cognit. 2 (2013) 42–52.

[3] R.D. Stoel, W. Kerkhoff, E.J.A.T. Mattijssen, C.E.H. Berger, Building the research culture in the forensic sciences: announcement of a double blind testing program, Sci. Justice 56 (2016) 155–156.

[4] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice 51 (2011) 204–208.

[5] M.J. Saks, R.F. Kidd, Human information processing and adjudication: trial by heuristics, Law Soc. Rev. 15 (1980) 123–160.

[6] Committee on Identifying the Needs of the Forensic Sciences Community: National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, USA, 2009.

[7] Executive Office of the President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, 2016.

[8] Forensic Science Regulator, Cognitive Bias Effects Relevant to Forensic Science Examinations, 2015.

[9] AA. Tversky, D. Kahneman, Judgment under uncertainty: heuristics and biases, Science (New York, N.Y.), 185 (1974) 1124–1131.

[10] I.E. Dror, D. Charlton, A.E. Peron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Sci. Int. 156 (2006) 74–78.

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

[12] J. Klayman, Varieties of confirmation bias, in: J. Busemeyer, R. Hastie, D.L. Medin (Eds.), Psychology of Learning and Motivation, Academic Press, 1995, pp. 385–418.

[13] G.S. Cooper, V. Meterko, Cognitive bias research in forensic science: a systematic review, Forensic Sci. Int. 297 (2019) 35–46.

[14] U.S. Department of Justice: Office of the Inspector General, A Review of the FBI's Handling of the Brandon Mayfield Case, 2006.

[15] R.D. Stoel, C.E.H. Berger, W. Kerkhoff, E.J.A.T. Mattijssen, I.E. Dror, Minimizing contextual bias in forensic casework, in: K.J. Strom, M.J. Hickman (Eds.), Forensic Science and the Administration of Justice: Critical Issues and Directions, SAGE Publications Inc, Thousand Oaks, California, 2014.

[16] W.C. Thompson, Painting the target around the matching profile: the Texassharpshooter fallacy in forensic DNA interpretation, Law, Probability Risk 8 (2009) 257–276.

[17] B. Found, J. Ganas, The management of domain irrelevant context information in forensic handwriting examination casework, Sci. Justice 53 (2013) 154–158.

[18] E.J.A.T. Mattijssen, W. Kerkhoff, C.E.H. Berger, I.E. Dror, R.D. Stoel, Implementing context information management in forensic casework: Minimizing contextual bias in firearms examination, Sci. Justice 56 (2016) 113–122.

[19] I.E. Dror, W.C. Thompson, C.A. Meissner, I.L. Kornfield, D.E. Krane, M.J. Saks, D.M. Risinger, Letter to the editor – context management toolbox: a linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making, J. Forensic Sci. 60 (2015) 1111–1112.

[20] D.E. Krane, S. Ford, J.R. Gilder, K. Inman, A. Jamieson, R. Koppl, I.L. Kornfield, D.M. Risinger, N. Rudin, M.S. Taylor, W.C. Thompson, Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, J. Forensic Sci. 53 (2008) 1006–1007.

[21] T. Mussweiler, F. Strack, Numeric judgments under uncertainty: the role of knowledge in anchoring, J. Exp. Soc. Psychol. 36 (2000) 495–518.

[22] A. Furnham, H.C. Boo, A literature review of the anchoring effect, J. Socio-Econ. 40 (2011) 35–42.

[23] B. Englich, T. Mussweiler, F. Strack, Playing dice with criminal sentences: the influence of irrelevant anchors on experts' judicial decision making, Pers. Soc. Psychol. Bull. 32 (2006) 188–200.

[24] T.D. Wilson, C.E. Houston, K.M. Etling, N. Brekke, A new look at anchoring effects: basic anchoring and its antecedents, J. Exp. Psychol. Gen. 125 (1996) 387–402.

[25] K.N. Ballantyne, G. Edmond, B. Found, Peer review in forensic science, Forensic Sci. Int. 277 (2017) 66–76.

[26] M.J. Saks, D.M. Risinger, R. Rosenthal, W.C. Thompson, Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States, Sci. Justice 43 (2003) 77–90.

[27] S.V. Stevenage, A. Bennett, A biased opinion: Demonstration of cognitive bias on a fingerprint matching task through knowledge of DNA test results, Forensic Sci. Int. 276 (2017) 93–106.

[28] G. Whitman, R. Koppl, Rational bias in forensic science, Law, Probability Risk 9 (2010) 69–90.

[29] I.E. Dror, Practical solutions to cognitive and human factor challenges in forensic science, Forensic Sci. Policy Manage.: Int. J. 4 (2013) 105–113.

[30] G. Edmond, A. Towler, B. Growns, G. Ribeiro, B. Found, D. White, K. Ballantyne, R.A. Searston, M.B. Thompson, J.M. Tangen, R.I. Kemp, K. Martire, Thinking forensics: cognitive science for forensic practitioners, Sci. Justice 57 (2017) 144–154.

[31] C.R. Sunstein, The law of group polarization, J. Political Philosophy 10 (2002) 175–195.

[32] F.L. Strodtbeck, R.M. James, C. Hawkins, Social status in jury deliberations, Am. Sociol. Rev. 22 (1957) 713–719.

[33] P.R. Costanzo, H.T. Reitan, M.E. Shaw, Conformity as a function of experimentally induced minority and majority competence, Psychonomic Sci. 10 (1968) 329–330.

[34] M. Deutsch, H.B. Gerard, A study of normative and informational social influences upon individual judgment, J. Abnormal Soc. Psychol. 51 (1955) 629–636.

[35] C.L. Ridgeway, Social Status and Group Structure, in: M.A. Hogg, R.S. Tindale (Eds.) Blackwell Handbook of Social Psychology: Group Processes, 2001, pp. 353–354.

[36] C.C.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, second ed., John Wiley and Sons, Chichester, UK, 2004.

[37] T.G. Fadul, G.A. Hernandez, S. Stoiloff, S. Gulati, An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides, in, Miami-Dade Police Department Crime Laboratory, 2011.

[38] T.P. Smith, A.G. Smith, J.B. Snipes, A validation study of bullet and cartridge case comparisons using samples representative of actual casework, J. Forensic Sci. 61 (2016) 939–946.

[39] N.K.P. Osborne, M.C. Taylor, Contextual information management: An example of independent-checking in the review of laboratory-based bloodstain pattern analysis, Sci. Justice 58 (2018) 226–231.

[40] B. Englich, T. Mussweiler, Sentencing under uncertainty: anchoring effects in the courtroom, J. Appl. Soc. Psychol. 31 (2001) 1535–1551.

[41] G. Stasser, W. Titus, Hidden profiles: a brief history, Psychol. Inq. 14 (2003) 304–313.

[42] C. Christensen, A.S. Abbott, Team medical decision making, in: G.B. Chapman, F.A. Sonnenberg (Eds.), Decision Making in Health Care – Theory, Psychology, and Applications, Cambridge University Press, 2000, p. 273.

[43] J.R. Vanderkolk, Chapter 9 – Examination Process, in: E.H. Holder, L.O. Robinson, J.H. Laub (Eds.) The Fingerprint Sourcebook, US Department of Justice, National Institute Of Justice, 2011, pp. 9.1–9.26.

[44] I.E. Dror, Human expert performance in forensic decision making: Seven different sources of bias, Aust. J. Forensic Sci. 49 (2017) 541–547.

[45] W. Kerkhoff, R.D. Stoel, C.E.H. Berger, E.J.A.T. Mattijssen, R. Hermsen, N. Smits, H.J. Hardy, Design and results of an exploratory double blind testing program in firearms examination, Sci. Justice 55 (2015) 514–519.

[46] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, C.E.H. Berger, F.W. Didden, J.H. Kerstholt, A part-declared blind testing program in firearms examination, Sci. Justice 58 (2018) 258–263.

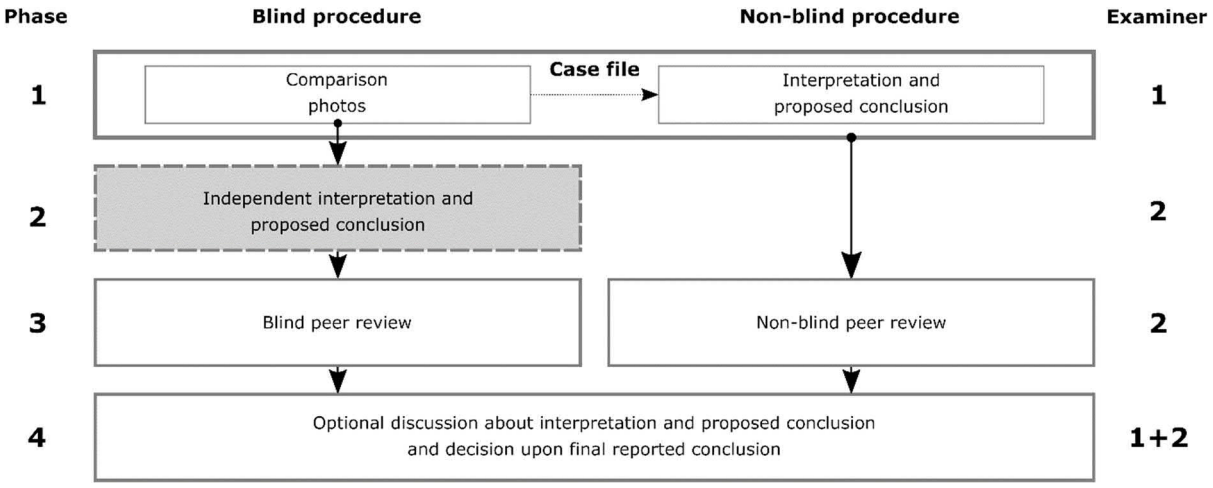

Figure 1. Schematic representation of the non-blind (with an anchor) and blind (without an anchor) peer review procedures.

Hypothesis 1	Hypothesis 2	Likelihood ratio	1 - 2	2 - 10	10 - 100	100 - 10,000	10,000 - 1,000,000	> 1,000,000
		Verbal conclusion	Approximately equally probable	Slightly more probable	More probable	Appreciably more probable	Far more probable	Extremely more probable

Figure 2. Representation of the feedback box for the proposed and final conclusions. The numerical likelihood ratio ranges and verbal expressions are similar to those in Table 1.

Table 1. Relation between the judged likelihood ratios ranges and the corresponding verbal expressions.

Judged likelihood ratio range	Verbal expression
1-2	Approximately equally probable
2-10	Slightly more probable
10-100	More probable
100-10,000	Appreciably more probable
10,000-1,000,000	Far more probable
> 1,000,000	Extremely more probable

Table 2. Percentages of data acquired and the number of examiners involved per study period for the non-blind and blind peer review procedures.

Period	Non-blind Procedure		Blind Procedure	
	% of data acquired	Number of examiners involved	% of data acquired	Number of examiners involved
First 8 months	55	8	0	NA
Pilot of 3 months	10	8	15	3
Last 11 months	35	8	85	8

Table 3. Overview of decisions and changes of evidential strengths between proposed and reported conclusions.

Reported evidential strength	Same evidential strength as one of the proposed conclusions (N = 83)		Alternative evidential strength to those of the proposed conclusions (N = 17)	
	Disagreement about lower bound	Disagreement about upper bound	Disagreement about lower bound	Disagreement about upper bound
Same as both proposed evidential strengths	18	30	3	1
Lower than both proposed evidential strengths	0	0	1	0
Lowest proposed evidential strength chosen	29	30	4	0
Intermediate of both proposed evidential strengths	0	0	4	9
Highest proposed evidential strength chosen	36	25	3	3
Higher than both proposed evidential strengths	0	0	2	2

Table 4. The number of discussions for which that examiner's proposed conclusion is reported as the final conclusion out of the total number of discussions per forensic examiner and the resulting percentages of these ratios. The forensic examiners are ranked by the percentage of discussions where their proposed conclusion became the finally reported conclusion.

Type of examiner	Frequency	Percentage (%)
Reporting	5/5	100
Reporting	15/20	75.0
Reporting	12/16	75.0
Reporting	23/33	69.7
Non-Reporting	13/40	32.5
Non-Reporting	4/14	28.6
Non-Reporting	5/25	20.0
Reporting	0/1	0.0