6-1-2018

Perceived strength of forensic scientists' reporting statements about source conclusions

William C. Thompson
University of California, Irvine

Rebecca Hofstein Grady
University of California, Irvine

Eric Lai
University of California, Irvine

Hal S. Stern
University of California, Irvine

Perceived strength of forensic scientists' reporting statements about source conclusions

Abstract

Three studies investigated lay people's perceptions of the relative strength of various conclusions that a forensic scientist might present about whether two items (fingerprints, biological samples) have a common source. Lay participants made a series of judgments about which of two conclusions seemed stronger for proving the items had a common source. The data were fitted to Thurstone–Mosteller paired comparison models to obtain rank-ordered lists of the various statements and an indication of the perceived differences among them. The results reveal the perceived strength of several types of statements, relative to one another, including verbal statements regarding strength of support (e.g. 'extremely strong support for same source'), source probability statements (e.g. 'highly probable same source'), random match probabilities (e.g. RMP = 1 in 100 000), likelihood ratios, and categorical statements (e.g. 'identification'). These comparisons in turn provide insight into whether particular statements about the strength of forensic evidence convey the intended meaning and will be interpreted in a manner that is justifiable and appropriate.

Disciplines

Forensic Science and Technology

Comments

This is a manuscript of an article published as Thompson, William C., Rebecca Hofstein Grady, Eric Lai, and Hal S. Stern. "Perceived strength of forensic scientists' reporting statements about source conclusions." *Law, Probability and Risk* 17, no. 2 (2018): 133-155. Posted with permission of CSAFE.

Perceived strength of forensic scientists' reporting statements about source conclusions

William C. Thompson†, Rebecca Hofstein Grady, Eric Lai and Hal S. Stern
University of California, Irvine, CA, USA

Three studies investigated lay people's perceptions of the relative strength of various conclusions that a forensic scientist might present about whether two items (fingerprints, biological samples) have a common source. Lay participants made a series of judgments about which of two conclusions seemed stronger for proving the items had a common source. The data were fitted to Thurstone–Mosteller paired comparison models to obtain rank-ordered lists of the various statements and an indication of the perceived differences among them. The results reveal the perceived strength of several types of statements, relative to one another, including verbal statements regarding strength of support (e.g. 'extremely strong support for same source'), source probability statements (e.g. 'highly probable same source'), random match probabilities (e.g. RMP = 1 in 100 000), likelihood ratios, and categorical statements (e.g. 'identification'). These comparisons in turn provide insight into whether particular statements about the strength of forensic evidence convey the intended meaning and will be interpreted in a manner that is justifiable and appropriate.

Keywords: forensic; report; likelihood ratio; identification; support; source; probability; statistic; conclusion.

Forensic scientists often compare items in order to assess whether they have (or might have) a common source. For example, latent print examiners compare prints lifted from crime scenes with the fingerprints of suspects, footwear examiners compare shoe prints found at crime scenes with test prints made by suspects' shoes, and DNA analysts compare biological specimens associated with crimes to reference samples of known individuals. In this article, we consider how lay people assess the strength of various statements forensic scientists might make when reporting the results of such comparisons.

When choosing how to report their conclusions, forensic scientists should first consider what kinds of statements are warranted—i.e. what statements can be justified logically and empirically given the accuracy of the analytic methods used to reach the conclusion. Among statements that pass this initial test, the forensic scientist should then consider, as a second step, which statements best convey the forensic examiner's conclusions about the strength (probative value) of the forensic evidence. Forensic scientists should avoid making statements that give false or misleading impressions of evidence strength.

The research reported here is designed to assist forensic scientists in performing this second step. It examines the way lay people interpret various statements that might be used in reports and testimony to

explain the probative value of a forensic comparison. Information about the way lay people perceive the strength of a forensic comparison, when reported in various ways, may help forensic scientists decide which reporting statements will best convey their intended meaning.

1. Possible ways to report source conclusions

Forensic scientists can report the strength of source conclusions in a variety of ways. In the studies reported here, we examined reactions to statements of the following types:

1. Likelihood Ratios (LRs) are numerical quantities that represent the relative probability of the observed similarities and discrepancies in the evidence under two alternative propositions (hypotheses) about the source of the items. Typically, the alternative propositions are: (1) that the items have the same source; and (2) that the items have a different source; but more complex propositions are sometimes evaluated.[1] Experts often make the favoured hypothesis the numerator of the LR so that reported values range from one to infinity. A value of one means the results observed when comparing the items are equally likely under the two hypotheses, and hence that the evidence has no value for distinguishing the hypotheses. A value greater than one means the observed results are more likely under one hypothesis than the alternative, and thus that the forensic evidence supports the favoured hypothesis. Larger LRs indicate that the evidence provides stronger support for the favoured hypotheses.

How do forensic examiners come up with LRs? In some disciplines, examiners can rely on databases and statistical modelling. This is most common in fields like forensic DNA analysis (Butler, 2009) and forensic voice comparison (Morrison and Thompson, 2017) where extensive databases exist and methods for statistical modelling have been evaluated in the scientific literature (Evett and Weir, 1998). LRs have been presented in the United States for many years in connection with forensic DNA evidence, particularly in cases involving mixtures of DNA from more than one person. The expert typically says something like: 'The genetic features observed in the evidentiary sample are x times more likely if the sample contains DNA from the victim and defendant than if the sample contains DNA from the victim and a random unknown Caucasian.'

In forensic science disciplines that have not developed databases and statistical models, examiners sometimes report LRs that reflect their beliefs about the relevant likelihoods based on their evaluation of the items in question in light of their training and experience. In some instances, the LR is informed by (and rests partly upon) empirical data, but nevertheless depends on the examiner's subjective evaluation (Biedermann et al., 2012; Morrison and Thompson, 2017). While some commentators have raised concerns about reporting LRs that reflect examiners' subjective beliefs—Risinger (2013) called them 'numbers from nowhere'—those who favor reporting these LRs point out that examiners' subjective beliefs about probability are the basis for the conclusions drawn in many forensic science disciplines (Sjerps and Berger, 2012). If the examiner cannot make an accurate assessment of the relevant probabilities, then the examiner does not know enough to evaluate the strength of the forensic evidence—and hence nothing the examiner says about the value of the evidence should be trusted. The practice of presenting LR that are based on the examiner's subjective beliefs (rather than data-based

[1] Some readers may find a mathematical restatement of the LR approach to be helpful. Let E be the observed results of the comparison; let H_s be the proposition that the item being have the same source and H_d be the proposition that the items have a different source. In order to draw source conclusions under the likelihood ratio approach on the basis of a forensic examination, a forensic examiner must consider both $p(E|H_s)$ and $p(E|H_d)$. The LR is the ratio of these two probabilities.

statistical models) appears to have taken hold in many European countries (Willis et al., 2015; Berger et al., 2011), but is uncommon in the USA.

2. Strength of Support Statements (SoS) are non-numerical statements about the degree to which the results of a forensic comparison support the proposition that the items have the same source (or a different source). In some instances, these statements are intended to be verbal equivalents of likelihood ratios. For example, the UK-based Association of Forensic Science Providers has proposed that forensic scientists use the 'verbal expressions' shown in Table 1 to describe how strongly their evidence supports a particular hypothesis about the evidence (e.g. the hypothesis that two items have a common source; AFSP, 2009). Under this approach, forensic scientists first make a judgment about the LR and then use one of the verbal expressions in the table instead of (or in addition to) the number to describe their conclusions in reports and testimony.

For example, a forensic scientist who concludes (by whatever means) that the results observed in a forensic comparison are 500 times more likely if the items have a common source than if they have a different source would report that the comparison provides 'moderately strong' support for the conclusion that the items have a common source. A forensic scientist who concluded that the results are 100 000 times more likely if the patterns being compared have a common source would say that the evidence provides 'very strong support' for the hypothesis of a common source. Statements of this type are not common in U.S. courts but have been discussed extensively in the academic literature (NIST, 2012; Marquis et al., 2016).

It is interesting to note that the verbal expressions proposed for use in forensic science differ from those that have been suggested in more general discussions of LR and Bayes factors. Kass and Raftery (1995) present a scale in which Bayes factors larger than 10 are viewed as providing 'strong' evidence and values larger than 100 are viewed as 'decisive'. The difference likely stems from the fact that the forensic science community is trying to develop a scale that will be useful across the full range of forensic science disciplines from shoeprints to latent prints to DNA evidence. The large values of LR (Bayes factors) that can be observed in single source DNA analyses require that the verbal equivalent scale cover a wider range than would be common (or needed) in traditional uses of the LR to compare a limited set of statistical models.

3. Match frequencies and random match probabilities. When a comparison reveals matching features in two items, forensic scientists sometimes estimate and report the frequency of the matching features in a reference population. This occurs most commonly in forensic DNA analysis, where genetic databases provide an empirical basis for estimating the frequency of DNA profiles in various human populations. Forensic DNA analysts sometimes present these estimates as *match frequencies*—e.g. 'The blood stain at the crime scene and the reference blood sample from the suspect have the same

TABLE 1 Proposed likelihood ratio terminology (AFSP, 2009)

Numerical expression of probative strength (likelihood ratio)	Verbal expression of probative strength
1–10	Weak or limited
10–100	Moderate
100–1000	Moderately strong
1000–10 000	Strong
10 000–1 000 000	Very strong
>1 000 000	Extremely strong

DNA profile; this profile is estimated to occur in one person in 10 million among Caucasian-Americans.' Alternatively, they may use these estimates to assign *random match probabilities (RMPs)*—e.g. 'The probability that a random Caucasian-American would match this DNA profile is 0.0000001 or 1 in 10 million.'

If the matching features are certain to be observed under the proposition that the items have the same source,[2] then the RMP is the complement of the LR. In these instances, the RMP and LR convey the same information. The RMP and LR are not equivalent, however, if the observed features are not certain to be observed when the items have the same source.[3] In such cases, LRs are generally preferred because RMPs provide an incomplete and potentially misleading account of the strength of the forensic comparison (*see* Curran and Buckleton, 2010; Thompson, 1996; 2009 for examples related to DNA evidence; Morrison and Thompson, 2017 for discussion of this issue in connection with forensic voice comparison).

4. Likelihood of Observed Similarity (LoS)—Some laboratories have opted to make non-numerical statements about the probability of the observed results under the hypothesis that the items being compared have a different source. These are typically statements about match frequencies or random match probabilities. For example, in 2015 the Defense Forensic Science Center (DFSC) of the Department of the Army adopted the following reporting statement for positive latent print comparisons:

> The latent print on Exhibit ## and the record finger/palm prints bearing the name XXX have corresponding ridge detail. The likelihood of observing this amount of correspondence when two impressions are made by different sources is considered extremely low (Department of the Army, 2015).

We understand that the DFSC has since changed their approach to reporting latent print comparisons, but we regard their former reporting statement as a good example of what we call the LoS approach to presenting forensic source conclusions.

5. Source Probability Statements (SP)—Forensic examiners sometimes offer opinions on the probability that two items have a common source. Opinions of this type can be expressed quantitatively, using probabilities or percentages. For example, a forensic scientist might say there is a 99% chance that two items have a common source. It is more common, however, for examiners to express such conclusions with words rather than numbers. For example, the forensic scientist might say it is 'moderately probable'; or 'highly probable'; or 'practically certain' that two items have a common source.

A number of commentators have questioned the appropriateness of source probability statements on epistemological grounds, pointing out that forensic scientists cannot logically draw conclusions about source probabilities based on forensic science alone (Evett, 1998; Buckleton, 2005; Morrison, 2011; Thompson, 2012; Robertson et al., 2016). Consider, for example, whether it is logically possible to draw a conclusion about source probability from the observation that two DNA samples share a genetic profile that would be found in only one person in 1 million. The mere fact that they share a rare profile provides no basis for determining whether the samples in question are likely or unlikely to

[2] I.e., $p(E|H_s)=1.0$
[3] I.e., $p(E|H_s)<1.0$

have a common source unless one makes assumptions or considers evidence about the prior odds that the sample came from the same person.[4]

People sometimes mistakenly assume that a low random match probability means that there is a high probability the matching items have a common source, but this is a logical error that has been called the 'source probability error' (Koehler, 1993; Koehler et al., 1995); the 'fallacy of the transposed conditional' (Evett, 1998); or the 'prosecutor's fallacy' (Thompson and Schumann, 1987). This error is similar to the mistaken (but common) assumption that a *P*-value reflecting the probability of observed data under a null hypothesis reflects the probability that the null hypothesis is true (Wasserstein & Lazar, 2016).

Forensic scientists cannot logically draw conclusion about source probabilities without taking a position on the prior odds that the items in question have the same source. Doing that, however, requires the forensic scientist to delve into matters outside their scientific expertise. A number of commentators have raised questions about when, if ever, forensic scientists should do this (Evett, 1998; Jackson, 2009; Morrison, 2011; Thompson, 2012, 2015; Thompson et al., 2013; National Commission, 2015; Robertson et al., 2016). In light of concerns about whether source probabilities are logically justified and appropriate, it is important to consider whether forensic scientists' source conclusion might be conveyed as effectively to a lay audience by other more justifiable reporting statements.

***6. Categorical conclusions*—**Forensic scientists in some disciplines simply state a bottom line conclusion about whether two items have a common source. In latent print analysis examiners have traditionally reported either that the prints being compared were made by the same finger, or that they were made by different fingers, or that the results of the comparison are inconclusive (Cole, 2014; Eldridge, 2017). The conclusion that prints have the same source has traditionally been described as 'identification' or 'individualization', while the conclusion that they have a different source is called 'exclusion'. Similar reporting terms have also been used in other disciplines when analysts have high confidence that the items being compared do or do not have a common source.

Statements of this type have been criticized on grounds that they imply absolute certainty or, at least, a greater level of certainty than is warranted (NAS, 2009; PCAST, 2016; AAAS, 2017). These statements have been defended, however, on grounds that strong statements are necessary to convey the high probative value of an expert's conclusion that items share a highly discriminating pattern. There is concern that if forensic scientists abandon categorical statements like 'identification' or 'individualization', then people will give their conclusions too little weight. In order to evaluate these claims, it will be helpful to know how lay people perceive the strength of statements about identification and individualization relative to other kinds of reporting statements.

When forensic experts compare items, they typically look for distinguishing features that rule out the possibility of a common source. If they find no distinguishing features, but are uncertain of the probative value of the comparison, they sometimes report that one item 'cannot be excluded' as coming from or having the same source as the other item. For example, they might say that they 'cannot exclude' a shoeprint as having come from a particular shoe, or a tool mark as having been

[4] After comparing two items, a forensic examiner may be able to estimate the likelihood of the observed results under the alternative hypotheses: $p(E|H_s)$ and $p(E|H_d)$. But these likelihoods are not the same as source probabilities; source probabilities are the inverse of these conditionals—i.e., $p(H_s|E)$ and $p(H_d|E)$—mathematically known as the posterior probabilities of the hypotheses. To infer source probabilities (posterior probabilities) from the likelihoods, the examiner must take into account the prior probability that the items have the same source, $p(H_s)$, or different source, $p(H_d)$, because according to Bayes' rule, $p(H_s|E)/p(H_d|E) = p(H_s)/p(H_d) \times p(E|H_s)/p(E|H_d)$. This means that conclusions about source probability cannot rest solely on what the examiner observes when making the comparison but must also depend on assumptions or conclusions about the *a priori* probability the items have the same source. Consequently, examiners must necessarily consider or make assumptions about matters beyond forensic science in order to reach source conclusions.

made by a particular tool (Jackson, 2009). If the patterns observed in two items are indistinguishable, but of unknown probative value, experts might also report that the patterns 'match'.

Unfortunately, it is impossible to evaluate the probative value of such terms without knowing something about the likelihood the expert would report a 'failure to exclude' or a 'match' when comparing items from the same source and from different sources. If the items share features that are common in a given reference population, then a 'match' or 'non-exclusion' might have little probative value for proving the items have a common source; but if the items share rare features, then the comparison might be highly probative. Without information about the relevant likelihoods, it is unclear how much weight *should* be given to a 'match' or 'failure to exclude'. One of the goals of the studies reported here was to learn how much weight people *do* give these ambiguous statements, relative to other reporting statements that might be used.

In sum, there are a variety of ways to characterize the strength of a forensic scientist's source conclusions, but there is not yet consensus on which reporting method is best. Because the question turns, in part, on how the various possible statements are understood by the target audience, we wanted to learn more about how lay people evaluate the strength of possible reporting statements, relative to one another. The studies reported here explore that question.

2. Method

2.1 *Overview*

To measure people's perceptions of the strength of the various statements that experts might make about a forensic comparison, we adopted a method originally used in the field of psychometrics to study perceptions of the strength of physical stimuli (e.g. the brightness of a light, the intensity of a sound). People have difficulty providing meaningful evaluations of the strength of such stimuli on rating scales (e.g. 'How loud is this sound on a scale of 1-10?'). Responses tend to be unreliable, poorly calibrated, and affected by contextual factors, such as the volume of previously heard sounds, and by the nature of the rating scale (Gescheider, 1997; Zeller and Carmines, 1980). People do better, however, when judging the relative strength of stimuli. People are more reliable when reporting which of two sounds is louder than when rating the loudness of various sounds on a scale.

This observation led L.L. Thurstone (1927) to propose paired comparison as a method for ordering the perceived strength or intensity of physical stimuli. Thurstone demonstrated that pair-wise comparison can be used to order multiple items in a scale of strength or magnitude. Subsequently, Mosteller (1951), Bradley and Terry (1952) and Luce (1994) elaborated on the Thurstone approach, developing models for using paired comparison data to scale preferences and perceptions.

Asking people to judge the strength of a statement about forensic evidence struck us as analogous to asking about the strength of a physical stimulus. We doubted that we could obtain meaningful responses through the use of rating scales, so we decided to rely on rankings. We initially attempted a study in which people were asked to rank a large number of statements about the strength of a forensic comparison according to the strength of each statement, but we found that people had difficulty reliably ranking multiple statements. We eventually discovered (or perhaps re-discovered) what psychometricians have known for a century—that people can provide more meaningful responses when asked to evaluate the relative strength of two stimuli, in our case two statements. Accordingly, we decided to follow the psychometric approach and use Thurstone's method of paired comparison as a way of obtaining data that could be used to scale the perceived strength of various reporting statements.

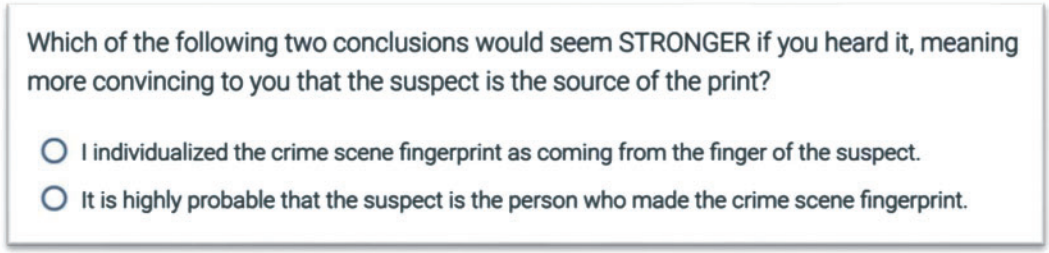

FIG 1. Example of how pairs of statements were presented to participants (Study 2).

We conducted three studies in which jury-eligible adults recruited from an online labour pool evaluated the relative strength of statements used to report a forensic scientist's conclusion following a fingerprint comparison (Study 1 and 2) or DNA comparison (Study 3). We elected to conduct three separate studies, rather than one larger study, in order to expand the number of statements that could be compared and test the generality of findings across different types of forensic evidence (fingerprint or DNA), while still minimizing workload for individual participants.

Participants evaluated the statements in pairs, indicating which of two statements was stronger for proving the items have the same source. Figure 1 is an example of a pair of statement participants were asked to compare.

In each study nine statements were paired randomly, creating a pool of 36 possible comparisons (we did not consider the order in which the statements appeared within the pair to be important). To keep the study to a reasonable length and avoid fatigue, we limited the number of pairs evaluated by each participant to a selection of 16. Some statements were used in all three studies; some in just one or two of the studies. Table 2 presents a complete list of statements used throughout the studies.

2.2 *Recruitment of participants*

We recruited participants for each of the three studies from Amazon's Mechanical Turk (mTurk), an online labour pool frequently used for social science research (Buhrmester et al., 2011) that can offer cost-effective and high quality data (Paolacci and Chandler, 2014), especially when restricted to workers from the USA (Smith et al., 2016).[5] We offered workers \$0.60–\$0.70 to take part in an online study of how people evaluate the strength of forensic science evidence. The invitation was available only to workers with IP addresses in the USA and we used a web utility to screen out respondents whose mTurk ID numbers had previously been used to participate in one of our studies. Those who passed this initial screening were asked to affirm that they were American citizens at least

[5] Thompson and Newman (2015) recruited participants for a study of lay reactions to forensic science evidence using the same recruitment methods used here. In a supplement to the published article (available online at http://supp.apa.org/psycarticles/supplemental/lhb0000134/lhb0000134_supp.html, *see* Table S1), they provided a detailed demographic breakdown, comparing participants recruited from mTurk with a sample of actual jurors who had been recruited from a county jury pool for an earlier study (Thompson, Kaasa & Peterson, 2013). The actual jurors, from an affluent suburban county, tended to be older, more affluent, and better educated than the mTurk recruits; with regard to education and income, however, the mTurk recruits were more representative of American adults (and hence of jurors in general) than the sample from the county jury pool. Thompson and Newman concluded that the mTurk recruits are sufficiently diverse and representative of American jurors to be suitable for a study of lay reactions to forensic evidence.

TABLE 2 Statements used in the studies

Study	Statement	
	Match frequency (RMP)	
2, 3	RMP4:	'one person in 10 million'
1, 2, 3	RMP3:	'one person in 100, 000'
1, 2	RMP2:	'one person in 1, 000'
1	RMP1:	'one person in 10'
	Likelihood ratios (LR)	
3	LR4:	'10, 000, 000 times more likely' if suspect rather than random person is source
3	LR3:	'100, 000 times more likely' if suspect rather than random person is source
	Categorical conclusions (CC)	
3	CC5:	'suspect was the source'
2	CC4:	'individualized ... as coming from the finger of the suspect'
2	CC3:	'identified ... to the finger of the suspect'
2	CC2:	'matches the fingerprint of the suspect'
3	CC1:	'suspect could have been the source'
	Likelihood of observed similarity (LoS)	
2	LoS1:	'likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is considered extremely low'
	Source probability statements (SP)	
1	SP3:	'a practical certainty that suspect was the source'
1, 2, 3	SP2:	'highly probable'
1	SP1:	'moderately probable'
	Strength of support (SoS)	
1, 2, 3	SoS4:	'extremely strong support'
3	SoS3:	'very strong support'
1	SoS2:	'moderate support'
1	SoS1:	'weak support'

18 years of age and to complete an additional series of screening questions designed to detect random or robotic responders and those unable or unwilling to follow instructions. Relatively few prospective participants were eliminated by these additional screening questions (8 in Study 1; 7 in Study 2; 5 in Study 3).

2.3 *Study 1*

Participants ($N = 122$; Mean Age = 34.7 years, $SD = 12.0$; 46.7% male; 47.1% college graduates) were asked to imagine that a fingerprint expert compared a fingerprint found at a crime scene with the fingerprint of a person suspected of committing the crime and had found the two prints indistinguishable. They were told that the strength of this evidence depends on the quality and number of distinguishing features in the prints and that experts need to explain how strong the evidence is in a

particular case. They were then asked to evaluate a series of statements that an expert might make about the strength of fingerprint evidence. These statements were presented in pairs and participants were asked to indicate, by checking a box, which of the two statements in each pair was stronger, where the stronger statement was 'the one that makes the fingerprint evidence sound *more convincing or more conclusive* for proving the suspect made the print at the crime scene'.

In Study 1, we examined perceptions of nine possible statements. There were three statements about match frequency:

> RMP1: 'Given the size and quality of the crime scene print I would expect about one person in 10 to have a fingerprint similar enough to be indistinguishable from it.'
>
> RMP2: 'Given the size and quality of the crime scene print I would expect about one person in 1000 to have a fingerprint similar enough to be indistinguishable from it.'
>
> RMP3: 'Given the size and quality of the crime scene print I would expect about one person in 100, 000 to have a fingerprint similar enough to be indistinguishable from it.'

There were also three statements about source probability (SP):

> SP1: 'Given the size and quality of the crime scene print, it is moderately probable that the suspect is the person who made the crime scene print.'
> SP2: 'Given the size and quality of the crime scene print, it is highly probable that the suspect is the person who made the crime scene print.'
> SP3: 'Given the size and quality of the crime scene print, it is a practical certainty that the suspect is the person who made the crime scene print.'

And there were three statements about strength of support (SoS):

> SoS1: 'Given the size and quality of the crime scene print, these findings provide weak support for the theory that the suspect is the person who made the crime scene print.'
> SoS2: 'Given the size and quality of the crime scene print, these findings provide moderate support for the theory that the suspect is the person who made the crime scene print.'
> SoS4: 'Given the size and quality of the crime scene print, these findings provide extremely strong support for the theory that the suspect is the person who made the crime scene print.'

In the first phase of the study participants were asked to make comparisons of statements in the same category—that is both statements were about match frequency, both about source probability, or both were about strength of support. We wanted to be sure that all participants made multiple within-category comparisons so that we would know whether they were evaluating statements within categories in the expected manner. We also thought that within-category comparisons would be easier than cross-category comparisons and hoped that the easier comparisons would help participants become accustomed to the study before encountering more challenging comparisons. They made seven of these within-category comparisons. In the second phase of the study participants were presented a random selection of nine of the remaining comparisons from the pool of possible comparisons. Comparisons were drawn from the pool without replacement, so that a given comparison could not be presented more than once to the same participant.

During pilot testing we encountered some unexpected responses when participants were asked to compare statements about match frequencies. While the majority of participants indicated that the

statement with the lower match frequency was stronger (e.g. the evidence is stronger if the match frequency is 1 in 100 000 than if it is 1 in 10), a substantial minority indicated the opposite. We determined that these unexpected (and incorrect) judgments arose largely from misinterpretation of the expert's statements. Some participants thought the expert was commenting on the number of possible contributors—that is, the number of people who might have left the fingerprint at the crime scene—rather than the frequency with which an 'indistinguishable' print would be found in the population. In other words, they thought the expert was saying that the suspect was one of either 10 (or 100 000) people who could have left the print, and found the evidence more probative when they mistakenly thought the expert had claimed that it narrowed the possible sources to 10, than when the expert was asserting that the pool of possible sources was 100, 000. Most of those who misinterpreted the expert's statements in this way quickly reversed their judgment, however, when asked to think more carefully about what the expert was saying. For example, when we asked them to think about the number of people besides the suspect who would have 'indistinguishable' fingerprints, under the two frequency estimates, most of them then chose the expected answer and continued to give the expected (correct) response for subsequent comparisons involving two match frequencies. We concluded that the unexpected answers were, in most cases, simply a mistaken response based on careless reading rather than a reflection of participants' considered judgment.

To reduce the chance that such careless errors might distort our findings, we decided to begin the initial phase of the study by asking participants whether statement RMP1 or RMP3 was stronger. To be sure participants understood the expert's statements correctly, those who chose RMP1 were presented with the following prompt:

> Please take a closer look at this question before moving on. Think about the number of people besides the suspect who would have 'indistinguishable' fingerprints. The number of 'indistinguishable' people will be far smaller if one person in 100, 000 is 'indistinguishable' than if one in 10 is 'indistinguishable.'

They were then given the opportunity to answer the question again and we used their second answer as their response to this question. Those who gave the wrong answer a second time were asked to explain (in a text box) why they thought RMP1 was stronger than RMP3, but were not removed from the study.

At the end of the study, participants answered demographic questions about their age, gender and education level.

2.4 *Study 2*

A new group of participants ($N = 127$; mean age $= 36.1$, $SD = 14.6$; 48.8% male; 44.1% college graduates) was screened in the same manner as in Study 1. Participants were asked to imagine that an expert 'finds corresponding ridge detail in the two fingerprints', one from a crime scene and the other from a suspect. They were then asked to evaluate a series of statements that an expert might make about the strength of this fingerprint evidence. As in Study 1, these statements were presented in pairs and participants were asked to indicate, by checking a box, which of the two statements in each pair was stronger, where the stronger statement was 'the one that makes the fingerprint evidence sound *more convincing or more conclusive* for proving the suspect made the print at the crime scene.' Each participant evaluated fifteen pairs of statements.

Nine different statements were evaluated in this study. Some of the statements include the same numbers or phrases as statements in Study1 and are therefore given the same acronym, although the match frequency (RMP) statements in Study 2 have a slightly different wording than those in Study 1. Study 2 included three statements about match frequency:

> RMP2: 'The likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is less than 1 in 1000 (one thousand).'
>
> RMP3: 'The likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is less than 1 in 100, 000 (one hundred thousand).'
>
> RMP4: 'The likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is less than 1 in 10, 000, 000 (ten million).'

It included three categorical conclusions:

> CC2: 'The crime scene fingerprint matches the fingerprint of the suspect.'
> CC3: 'I identified the crime scene print to the finger of the suspect.'
> CC4: 'I individualized the crime scene fingerprint as coming from the finger of the suspect.'

It included a source probability statement:

> SP2: 'It is highly probable that the suspect is the person who made the crime scene fingerprint.'

It included a Strength of Support statement:

> SoS4: 'These findings provide extremely strong support for the theory that the suspect is the person who made the crime scene fingerprint.'

Finally, it included a Likelihood of Similarity (LoS) statement like that adopted in 2015 by the Defense Forensic Science Center of the Department of the Army:

> LoS1: 'The likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is considered extremely low.'

In the first phase of the study, participants were asked to compare RMP statements. To be sure participants were reading these statements in a thoughtful manner, those who gave an unexpected (incorrect) response when making the first comparison (RMP4 versus RMP 2) were given a prompt similar to that in Study 1 and allowed asked to answer the question again. All participants then made two additional comparisons of RMP statements.

In the second phase of the study, participants were presented with randomly paired statements drawn from the entire pool of possible comparisons. Each participant evaluated 12 random pairings in the second phase of the study.

2.5 *Study 3*

In Study 3, a fresh group of participants ($N = 138$; mean age = 33.3, *SD* 9.5; 58.7% male; 48.0% college graduates) was recruited using the same procedure as Study 1 and 2. Participants were asked to imagine that a DNA expert compared blood from a crime scene with blood of a man suspected of committing the crime and that the expert found 'the same genetic characteristics' in both blood

samples. Participants were then asked to evaluate a series of statements that an expert might make about the strength of this DNA evidence. As in the previous studies, these statements were presented in pairs and participants were asked to indicate, by checking a box, '[w]hich statement makes the DNA evidence sound stronger (for proving the suspect is the source of the crime scene blood)?' Each participant evaluated sixteen randomly selected pairs of these statements.

In this study (unlike Study 1 and 2), we did not include a first phase in which participants ranked statements within-category and we did not include any prompts for participants who gave unexpected (incorrect) answers to questions about RMPs. The order and pairing of statements was entirely random, with the sole restriction that no pair could be presented more than once.

Nine different statements were evaluated in this study. Some of these parallel the statements presented in the previous studies and so are given the same acronyms for ease of comparison (although there are minor difference in wording between these statements and comparable statements in the previous studies).

The study included two statements about match frequency (RMPs):

> RMP3: 'The genetic characteristics observed in the crime scene blood and the suspect's blood would be found in approximately one person in 100, 000.'
>
> RMP4: 'The genetic characteristics observed in the crime scene blood and the suspect's blood would be found in approximately one person in 10 million.'

It included two statements about likelihood ratios (LRs) of a similar strength as the RMP statements:

> LR3: 'The genetic characteristics observed in the crime scene blood are approximately 100, 000 times more likely if the blood came from the suspect than if it came from a randomly chosen person.'
>
> LR4: 'The genetic characteristics observed in the crime scene blood are approximately 10 million times more likely if the blood came from the suspect than if it came from a randomly chosen person.'

It included two categorical conclusions (CC):

> CC1: 'The suspect could have been the source of the crime scene blood.'
> CC5: 'The suspect was the source of the crime scene blood.'

It included a source probability statement (SP):

> SP3: 'It is highly probable that the suspect was the source of the crime scene blood.'

And it included two strength of support (SoS) statements:

> SoS3: 'These findings provide very strong support for the theory that the suspect was the source of the crime scene blood.'
> SoS4: 'These findings provide extremely strong support for the theory that the suspect was the source of the crime scene blood.'

2.6 *Statistical analysis*

Data from the paired-comparison experiments were analysed using the Thurstone–Mosteller paired-comparison model (Thurstone, 1927; Mosteller, 1951). The Thurstone–Mosteller model is an example

of a linear paired comparison model. In our context, a linear paired comparison model assumes that there is a linear, underlying dimension of 'perceived strength' for each statement. More precisely, the strength of a statement as perceived by a listener is assumed to be a random variable having a distribution on an underlying linear scale. For the Thurstone–Mosteller model the perception is modeled as following a normal distribution (with constant variance across statements) and the mean of the normal distribution is thus a parameter that characterizes the strength of the statement, often referred to as a strength or merit parameter. The probability that one statement is preferred to another is estimated by evaluating the normal cumulative distribution function at the difference of the two statements' strength parameters. Alternative models apply different distributional assumptions but Stern (1990, 1992) demonstrated that similar results are obtained for a given data set from a wide class of linear paired comparison models.

The Thurstone–Mosteller model can be viewed as a generalized linear model (Critchlow and Fligner, 1991) and standard statistical software can be used to obtain the maximum likelihood estimates of the strength parameters and the standard errors of the estimates. We used the glm function in the R programming environment (R Core Team, 2013) to fit the models to our data. Several points related to the estimates are important to note. It is necessary to identify a reference point for the underlying scale because the probability that a statement is judged stronger than another depends only on the difference between their strength parameters. The statement RMP3 ('one person in 100, 000 would be indistinguishable') was used as the reference point because it was present in all studies; it was assigned strength parameter equal to zero. It is important to note that the strength parameter estimates are not based just on direct comparison of each item to the RMP3 statement, but rather all parameters are estimated simultaneously based on how each statement compared to all of the others. The standard errors of the statements (other than the reference statement) can be used to assess whether the strength parameter for the statement differs significantly from the reference statement. Appropriate standard errors for comparing any two statements (or any other contrast) must account for the correlation of the parameter estimates; these have been calculated using the full variance matrix of the parameter estimates. It is common in the analysis of data from paired-comparison experiments to assume that, given the strength parameters, the results of all comparisons are independent (both within a single participant and across participants). If there were heterogeneity among the rankings of statements within the population, then this would violate the assumption and the standard errors would likely be underestimates. Research is ongoing to address this possibility. A standard generalized linear model goodness-of-fit test can be used to assess the overall fit of the Thurstone–Mosteller model to our data (see, e.g. Dobson, 2001).

3. Results

Table 3 presents Thurstone–Mosteller paired comparison estimates (and standard errors) for each of the three studies, using the statement RMP3 ('one person in 100, 000') as the reference category since it was present in all studies. The estimated strength parameter for RMP3 is reported as zero in each case. The estimate presented for each other statement indicates the strength of that item relative to RMP3, with positive numbers indicating the statement is perceived as stronger than the statement that the match frequency/RMP is 1 in 100 000, and negative numbers indicating it is perceived to be weaker. In Table 3, the items are ordered from strongest to weakest based on their estimated strength parameter in each study. Each pair of items within a study was compared to assess whether the strength parameters were significantly different. Pairs of statements that did not differ significantly are identified by a common letter in Table 3. Thus for Study 1, the statements SP3 and SoS4 did not differ

TABLE 3 *Thurstone–Mosteller paired comparison estimates and standard errors (SE) for the three studies*

Study 1 (Fingerprints)				Study 2 (Fingerprints)				Study 3 (DNA)			
Statement		Coefficient	(SE)	Statement>		Coefficient	(SE)	Statement		Coefficient	(SE)
SP3	a	0.28	(0.10)	RMP4	a	0.60	(0.08)	LR4	a	0.65	(0.08)
SoS4	a, b	0.10	(0.10)	CC2	a	0.57	(0.09)	RMP4	a	0.59	(0.08)
RMP3	b	0.00	–	CC3	b	0.16	(0.09)	CC5	b	0.20	(0.08)
SP2		−0.19	(0.09)	**RMP3**	b, c	0.00	–	LR3	b	0.16	(0.08)
RMP2		−0.58	(0.08)	CC4	c	−0.04	(0.09)	**SoS4**	b, c	0.09	(0.08)
SoS2	c	−0.87	(0.09)	**SoS4**		−0.37	(0.09)	**RMP3**	c	0.00	–
SP1	c	−0.90	(0.10)	LoS1	d	−0.63	(0.09)	SoS3		−0.28	(0.08)
RMP1		−1.40	(0.09)	**SP2**	d, e	−0.81	(0.09)	**SP2**		−0.65	(0.08)
SoS1		−1.64	(0.11)	RMP2	e	−0.89	(0.09)	CC1		−1.44	(0.10)

Notes: Statements that share a letter did *not* differ significantly at $P < 0.05$. Statements in bold appeared in all three studies (with RMP3 used as the reference point in all models, with a strength of 0).

significantly, the statements SoS4 and RMP3 did not differ significantly, and the statements SoS2 and SP1 did not differ significantly. These determinations were made using a 0.05 significance level without adjusting for multiple comparisons. The absolute values of the coefficients should not be compared across studies, as they depend on the particular statements included in each study. For example, if one statement has a value of 0.6 in Study 1, and another has a value of 0.8 in Study 2, that does not mean that the latter statement is stronger than the former, as its weight was derived from comparison with a different set of statements. Values should be used to compare *within* a study, and then relative order can be compared across studies.

In this case, the goodness-of-fit test suggests that the Thurstone–Mosteller model (which incorporates eight parameters—strength parameters for each statement except RMP3) is a substantial improvement over a null model that assumes all statements are equal. At the same time, the Thurstone–Mosteller does not fit the data as well as a fully saturated model that allows a separate parameter for each pair of items. This suggests that the assumed linear scale may benefit from the incorporation of additional predictors.

The study protocol gathered information on characteristics of the respondents including age, gender and educational level. Preliminary analyses were carried out to assess if the rank ordering of statements in Study 2 seemed to depend on any of these predictors. Fitting the model separately to males and female or to young and old respondents indicated that the rank order of the strength of statements did not seem to differ across these groups. There were occasional reversals between neighbouring statements (e.g. RMP4 might be higher than CC2 in one subgroup while CC2 is rated stronger than RMP4 in the other) but these occurred only between statements that were not found to be significantly different in the original analysis. Additional ways of assessing the contribution of individual characteristics to the rankings of the statements is an area for further investigation.

3.1 *Sensitivity to strength of statements within categories*

Statements were generally ranked in the expected order within each category. In other words, statements that we viewed as stronger ranked significantly higher than those that we viewed as weaker.

Study 1 revealed that among Strength of Support (SoS) statements, 'extremely strong support' (SoS4) was viewed as significantly stronger than 'moderate support' (SoS2), which was in turn seen as significantly stronger than 'weak support' (SoS1). Among Source Probability (SP) statements, 'practical certainty' (SP3) ranked significantly higher than 'highly probable' (SP2), which was in turn ranked significantly higher than 'moderately probable' (SP1).

Study 2 included three categorical conclusions (CCs). The statement that the fingerprints 'match' (CC2) was ranked significantly higher than the statement that the expert had 'identified' the print as the suspect's (CC3), which in turn was significantly stronger than the statement that the expert had 'individualized' the print as coming from the suspect (CC4). That the ambiguous term 'match' was ranked highest was not a complete surprise, as other researchers (e.g., McQuiston-Surrett and Saks, 2009) have reported that people give a lot of weight to expert testimony about a 'match'. This finding should raise concerns, however, about experts using the term 'match' in cases where the probative value of the comparison may be weak or unknown. In Study 3, an expert's categorical statement that the suspect 'was the source' of a DNA sample (CC5) was, not surprisingly, ranked significantly higher than the statement that the suspect 'could have been the source' (CC1).

Study 3 also included two likelihood ratios (LRs). As expected, a LR of 10 million was ranked significantly higher than a LR of 100 000.

In all three studies, participants were asked to compare the strength of statements about match frequencies (RMPs). The findings support the conclusion that, in general, people are sensitive to the value of the match frequency (RMP) and give more weight to a forensic comparison when the match frequency is lower than when it is higher, as is appropriate. Our evaluation of participants' reactions to RMP statements was complicated, however, by our discovery during pilot testing that a minority of participants misinterpreted these statements in a manner that caused them to draw the opposite conclusion—i.e. that a forensic comparison is stronger when the match frequency is higher. They apparently misinterpreted the expert's statements about match frequencies as statements about number of possible perpetrators (or the number of suspects).

These findings posed a dilemma regarding how best to conduct the studies and analyse the data. That people can misinterpret statements about RMPs is, of course, important to know, given that one goal of the research is to identify circumstances in which lay people interpret an expert's statements in an unexpected manner. On the other hand, we are interested in people's considered judgment about the strength of experts' statements, not responses based on a superficial and easily corrected misreading of expert's statements. With regard to the design of the studies, we saw two possible ways to resolve the dilemma. One approach (which we adopted in Studies 1 and 2) was to attempt rehabilitation. We asked participants at the beginning of the study to judge the strength of two RMP statements, and then prompted those who gave an incorrect judgment to reconsider the matter, hoping that we would thereafter get more thoughtful, considered judgments concerning RMP statements. A second approach, which we adopted in Study 3, was simply to ignore the problem hoping that our results would not be distorted too much by the careless responses of participants who misread the RMP statements.

With regard to analysis of data, there were also two options. One option was to consider all of the data, including data of participants who gave incorrect answers when comparing the strength of RMP statements. The second option was to exclude from our data participants who made an incorrect judgment on these comparisons so that our results for the overall study would not be affected by participants who had clearly misinterpreted or misunderstood the RMP statements. To test the robustness of our findings, we analysed the data both ways. We report here the data for all participants,

although we also analysed the data for all three studies after excluding participants who gave the incorrect response to the RMP comparison questions. Excluding these participants did not change the rank ordering of any of the statements, although it increased the differences among the coefficients across statements.

We found that 32% of participants in Study 1 and 28% in Study 2 gave the incorrect answer when asked on the initial question to compare the strength of two RMP statements. However, all but four of these participants in Study 1, and all but two in Study 2 changed their answer following our prompt. Sixty-one participants in Study 3 were asked to compare statement RMP4 with RMP3. (The number making this particular comparison was less than the full sample because statements were paired randomly and each participant evaluated only sixteen of the 36 possible pairings of the nine statements examined in the study). Of those who made this comparison, about 82% (50/61) correctly concluded that RMP4 is stronger; about 18% (11/61) incorrectly concluded that RMP3 is stronger.

3.2 *Comparing perceived strength of statements across categories*

The most interesting aspect of these studies is what they reveal about the perceived strength of different types of reporting statements across categories relative to one another. Within each category some statements were perceived as very strong while others were perceived as relatively weak. In no category were all statements perceived as stronger or weaker than the statements in any other category.

In Study 1, for example, the statement that was perceived to be strongest was an SP statement—'practical certainty'—although it did not differ significantly in perceived strength from the SoS statement that the findings provide 'extremely strong support' for a common source. In Study 2, the top statement was a RMP ('1 in 10 million'), which did not differ significantly in perceived strength from the simple categorical claim that the fingerprints 'match'. In Study 3, the top statement was a LR ('10 million times more likely') which, not surprisingly, did not differ in perceived strength from a comparable RMP ('1 in 10 million').

The relative ranking of statements was generally consistent across studies. For example, RMP3 ('1 in 100 000') was consistently, in all three studies, ranked significantly higher than SP2 ('highly probable'). SoS4 ('extremely strong support') was also ranked consistently, in all three studies, as stronger than SP2. There was an inconsistency in the ranking of SoS4 relative to RMP3—the rankings do not significantly differ in Studies 1 and 3, but SoS4 is ranked significantly lower in Study 2. We view this inconsistency as minor and suspect it arose from sampling variability. The overall results suggest that SoS4 ('extremely strong support') is perceived as having about the same weight as saying the RMP is 1 in 100 000 or (as shown in Study 3) that the LR is 100 000.

We were a bit surprised that the categorical statement that the expert has 'identified' (CC3) or 'individualized' (CC4) the items—which many experts believe to be the strongest possible statements about a source determination—were not perceived to be the strongest by our participants. 'Identification' was ranked significantly higher than 'individualization' but both categorical statements were ranked significantly lower than a RMP of '1 in 10 million'. Neither 'identification' nor 'individualization' differs significantly in ranking from a RMP of 1 in 100 000 (RMP3), although both categorical conclusions were ranked significantly higher than 'extremely strong support' (SoS4). We did not directly compare 'identification' or 'individualization' with likelihood ratios, but the results of Study 3, which shows that LRs are viewed as equivalent in strength to comparable RMPs, suggests that 'identification' and 'exclusion' would be treated as roughly equivalent in strength to a LR of 100 000.

The statement once used by the Defense Forensic Science Center (LoS1) was perceived to be significantly weaker than reporting 'identification' or 'individualization'. It was also weaker than saying the results provide 'extremely strong support' for the theory of a common source (SoS4). Relative to the RMP statements, LoS1 fell between '1 in 100, 000' and '1 in 1000', differing significantly from both. It was ranked slightly but not significantly higher than the statement that it is 'highly probable' that there is a common source.

4. Discussion

The three studies reported here provide important insights into people's perceptions of the strength of various statements that forensic scientists might use to report the strength of source conclusions. Using methods that are novel for the study of forensic science evidence, but well established in psychophysics, we found that people evaluate the relative strength of various reporting statements in a manner that is, on the whole, sensible and appropriate. Statements designed to suggest that the strength of evidence was low or moderate were correctly perceived as weaker than statements designed to suggest the strength of evidence was high.

There were, however, two instances in which people appeared to evaluate reporting statements in a manner that is problematic. First, the simple statement that the ridge patterns of two fingerprints 'match' was perceived to be an extremely strong statement about the probative value of a fingerprint comparison. This is problematic because the term 'match' is sometimes used when the expert is uncertain about the probative value of the matching features. It is not clear whether the term match will be equally powerful with other forms of forensic evidence, but our findings suggest that forensic scientists should use this term cautiously, if at all, when reporting their conclusions, particularly when there is uncertainty about the probative value of the matching features for proving the items have a common source.

Secondly, some participants appeared to misunderstand the meaning of statements about match frequencies or random match probabilities (RMPs). While most participants correctly perceived that the value of a forensic comparison (for proving common source) is greater when the match frequency or RMP is lower, some incorrectly believed that a higher RMP means the evidence is stronger. This finding suggests that experts and lawyers should take care to explain the implications of match frequencies or RMPs, and not assume that the import of a low match probability will be apparent to all people.

Our findings also have implications for the broader controversy that surrounds the question of how forensic scientists should report their source conclusions. The debate is partly over what reporting statements are justified. For example, a number of commentators have questioned whether forensic scientists should claim to 'identify' or 'individualize' the source of traces given what is currently known about the sensitivity and specificity of their methods. In 2009, a committee of the National Academy of Sciences declared that 'with the exception of nuclear DNA analysis ... no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source' (NAS, 2009, p. 7). Similar concerns were expressed in a more recent report of the President's Council of Advisors on Science and Technology (PCAST, 2016). Yet, many forensic scientists in the USA continue to report that they have 'identified' or 'individualized' traces such as fingerprints, tool marks and other impressions to a single source. And while there have been efforts within the field to soften reporting language by avoiding claims that examiners can link impressions to a single source 'to the exclusion of

all others' or with 100% certainty (Cole, 2014; Eldridge, 2017), examiners continue to use terms like 'identification', despite criticism of that terminology from distinguished scientific bodies (e.g. NIST, 2012; AAAS, 2017).

As discussed in the introduction, there is also controversy about reporting source probabilities. Because any conclusions about source probability must, of logical necessity, rest in part on evidence or assumptions about the prior odds that the items in question have a common source, and because it is problematic for forensic scientists to base their conclusions on such evidence or assumptions, a strong case can be made that reporting source probabilities is improper. Yet, many forensic scientists continue to report conclusions about the probability or likelihood that two items have a common source. For example, SWGDOC, the Scientific Working Group for Forensic Document Examiners, has published standards on terminology for expressing conclusions of forensic document examiners (SWGDOC, 2013). Among the conclusions that examiner might reach are 'identification (definite conclusion of identity)' and 'strong probability (highly probable, very probable)'. The standard suggests, for example, that examiners issue reporting statements like the following: 'There is *strong probability* that [the suspect] wrote the questioned material, or it is my opinion (or conclusion or determination) that [the suspect] *very probably* wrote the questioned material' (emphasis in original).

Concerns about whether traditional reporting statements can be justified have caused forensic scientists, in recent years, to consider alternative reporting methods. The European Network of Forensic Science Institutes (Willis et al., 2015) has recommended that forensic scientists use likelihood ratios (LRs), perhaps in combination with statements about strength of support (SoS), to explain the strength of their findings. While alternative approaches appear to be taking hold in Europe for a variety of forensic science disciplines, LRs have been used in the USA only in connection with DNA and voice comparison evidence, and have not (yet) been used in pattern matching disciplines like latent print examination, footwear examination, tool mark examination or document examination.

Support for traditional reporting statements may rest in part on concerns about the viability of alternatives. Forensic scientists may worry that likelihood ratios will be difficult to justify and difficult for lay people to understand; they may worry that statements about strength of support (SoS) or likelihood of observed similarity (LoS) are too weak to convey the probative value of forensic source conclusions.

The studies reported here do not address all of those concerns, but should help allay fears that lay people will perceive the alternative reporting statements as weak. Our results suggest that statements involving numbers—RMPs and LRs—are perceived as very powerful. For a fingerprint comparison, the RMP of 1 in 100 000 was as strong as the categorical statement that the examiner had 'identified' or 'individualized' the print; the RMP of 1 in 10 million was even stronger. For DNA evidence, a LR of 100 000 was as strong as the categorical conclusion that the suspect 'was the source'; and a LR of 10 million was stronger still.

Some of the non-numerical statements about SoS were also perceived to be powerful. Saying that a fingerprint comparison provides 'extremely strong support' for the theory that the suspect made the print was seen as roughly equivalent (Study 1) to saying that it was 'a practical certainty that the suspect was the source'; saying the comparison provides 'extremely strong support' for the theory that the suspect made the print was perceived to be stronger than saying that it is 'highly probable' the suspect made the print. Saying a DNA match provides 'extremely strong support' for the theory of a common source was seen as roughly equivalent (in Study 3) to saying that the RMP is 1 in 100 000 or the LR is 100 000. So our results suggest that it is possible to make a strong statement about the

probative value of a forensic comparison by talking about strength of support rather than relying on categorical statements or source probabilities.

Our findings cast new light on the verbal expressions recommended by the Association of Forensic Science Providers (AFSP, 2009; see Table 1). These expressions were not developed through empirical testing; they simply reflect the best judgment of the AFSP about what words will be perceived as equivalent in strength to LRs of various levels. Our findings suggest that two of these expressions—'weak support' and 'moderate support' are indeed perceived in the manner intended. AFSP recommended reporting that a forensic comparison provides 'weak support' for the theory of a common source when the LR is 1–10. Our participants (in Study 1) saw 'weak support' (SoS1) as roughly equivalent in strength to reporting a LR of 10. AFSP recommended reporting that a comparison provides 'moderate support' for a common source when the LR is 10–100. We did not compare 'moderate support' (SoS2) to any LRs, but our participants rated its strength as falling between a RMP of 1 in 1000 and a RMP of 1 in 10, and Study 3 suggested that they treated RMPs and LRs as roughly equivalent.

On the other hand, our participants thought the expressions 'very strong support' and 'extremely strong support' were weaker than their corresponding LRs in the AFSP table of equivalence. AFSP recommended using the term 'very strong support' when the LR is between 10, 000 and 1 million, but our participants (in Study 3) found the term 'very strong support' to be significantly weaker that a LR of 100 000, although significantly stronger than saying a common source was 'highly probable'. AFSP recommended using the term 'extremely strong support' when the LR exceeds 1 million, but our participants found 'extremely strong support' to be equivalent in strength to a RMP of 1 in 100 000 in Study 1 and Study 3, and weaker than that in Study 2. Forensic scientists who are seeking a verbal statement comparable in strength to a LR of 1 million or more may need to find something stronger than 'extremely strong support'.

Of course, it is important to consider whether such strong statements are warranted when describing the strength of forensic sources comparisons in disciplines other than DNA analysis. If it would be an exaggeration to report a LR of 100 000 or higher when explaining the strength of a latent print, tool mark or footwear comparison, then arguably it is also an exaggeration to say that the comparison provides 'extremely strong support' for the theory of a common source (given that 'extremely strong support' is viewed as equivalent in strength to the LR of 100 000). In this regard, it is noteworthy that a number of false identifications have occurred in 'black-box' studies of the accuracy of latent print examiners (Ulery et al., 2011; Pacheco et al., 2014; for reviews see PCAST, 2016; AAAS, 2017). The error rates observed in these studies suggest that the LR describing the strength of a latent print identification may well be closer to 1000 than to 100 000. Hence, reporting that a latent print comparison provides 'extremely strong' or even 'very strong' support for the theory of a common source may be more than adequate to convey an accurate impression of the strength of this evidence. In fact, if one were seeking a verbal expression that is equivalent to reporting a LR of 1000, then the statement proposed by the Defense Forensic Science Center (LoS1) would be a more reasonable choice.

On the other hand, this analysis presumes that lay people give appropriate weight to LRs, which may not be the case (Martire et al., 2013; Thompson et al., 2013; Thompson and Newman, 2015). We urge readers to be cautious about using the perceived strength of LRs or RMPs as a reference point evaluating the appropriateness of lay interpretations of verbal statements. The research reported here provides insight into people's perceptions of the strength of various possible reporting statements *relative to one another*. These studies do not tell us how much weight people will give to any particular statement, nor do they tell us whether the weight people give to a particular statement is appropriate.

These studies indicate, for example, that saying an examiner has 'identified' two fingerprints to a common source is perceived as roughly equivalent to saying 'the likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is less than 1 in 100, 000'. From this finding it seems reasonable to infer that people will give the two statements—a categorical claim of identification and a RMP of 1 in 100 000—roughtly equal weight. But these studies do not tell how much weight that will be, and hence cannot be used to assess whether these statements will cause people to give the fingerprint evidence more or less weight than it deserves.

It is also important to keep in mind that the studies reported here concerned assessments of the various reporting statements in connection with fingerprint and DNA evidence. The same statements may well be viewed differently when presented in connection with other types of forensic evidence. Thompson and Newman (2015) found, for example, that statements about LRs, RMPs and SoS were given more weight by people evaluating DNA evidence than by people evaluating footwear evidence. They also found that people were more sensitive to variations in LRs and SoS statements when evaluating DNA than footwear evidence. Thompson and Newman proposed that people's reactions to forensic science evidence depend only partly on the words the expert uses to characterize the strength of the evidence; their reactions also depend on existing impressions, knowledge and presumptions about the type of evidence in question. This suggests that further research is needed to test whether the findings reported here will generalize to perceptions of forensic comparisons in other forensic disciplines.

More generally, readers should bear in mind that these studies examined lay reactions only to short written statements about the strength of forensic science evidence. While we believe these statements capture the essence of various reporting formats, and are consistent with the kind of statements found in written reports of forensic science findings, it is possible that reactions to such statements will differ when they are delivered by an expert testifying at trial, particularly if the expert has the opportunity to provide further explanatory context and lawyers have the opportunity to question the expert and offer commentary and argument on the implications of the statements. These studies also examine reactions of individuals, rather than judgments reached following group deliberation. Additional research examining lay reactions to more realistic simulated trial testimony would be helpful for determining how well the findings reported here will generalize to jury trials.

Acknowledgements

This research was supported by the Center for Statistical Applications in Forensic Evidence (CSAFE), which in turn is supported by the National Institute of Standards and Technology (NIST) through Cooperative Agreement # 70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

References

AAAS (American Association for the Advancement of Science). (2017). *Forensic Science Assessments: A Quality and Gap Analysis- Latent Fingerprint Examination* (Report prepared by William Thompson, John Black, Anil Jain, and Joseph Kadane). doi: 10.1126/srhrl.aag2874.

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, **49**(3), 161–164.

Berger, C., Buckleton, J., Champod, C., Evett, I., and Jackson, G. (2011). Evidence evaluation: A response to the Appeal Court judgment in *R v T*. *Science & Justice*, **51**, 43–49.
Biedermann, A, Taroni, F., and Champod, C. (2012). How to assign a likelihood ratio in a footwear mark case: An analysis and discussion in the light of *R v T*. *Law Probability and Risk*, **11**, 259–277.
Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**(3/4), 324–345.
Buckleton, J. (2005). A framework for interpreting evidence. In, Buckleton , Triggs , Walsh (eds) (pp. 27–63). Boca Raton, FL: CRC Press.
Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, **6**, 3–5
Butler, J. M. (2009). *Fundamentals of Forensic DNA Typing*. San Diego, CA: Academic Press.
Cole, S. A. (2014). Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States. *Law, Probability & Risk*, **13**, 117–150.
Critchlow, D. E., and Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, **56**(3), 517–533.
Curran, J. M., and Buckleton, J. (2010). Inclusion probabilities and dropout. *Journal of Forensic Sciences*, **55**(5), 1171–1173.
Department of the Army. (2015). Use of the term 'identification' in latent print technical reports. Information paper from the Defense Forensic Science Center. Retrieved from http://onin.com/fp/DFSC_LP_Information_Paper_Nov_2015.pdf
Dobson, A. (2001). *An Introduction to Generalized Linear Models*, 2nd edn. London: CRC Press.
Eldridge, H. (2017). The shifting landscape of latent print testimony: An American perspective. *Journal of Forensic Science and Medicine*, **3**, 72–81.
Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, **38**, 198–202.
Evett, I. W., and Weir, B. S. (1998). *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.
Gescheider, G. (1997). *Psychophysics: The Fundamentals* (3rd edn). Lawrence: Erlbaum Associates.
Jackson, G. (2009). Understanding forensic science opinions. In Fraser and Williams (eds) (pp. 419–445), Cullompton, UK: Willan Publishing.
Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.
Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics Journal*, **34**, 21–39.
Koehler, J. J., Chia, A., and Lindsey, S. (1995). The random match probability (RMP) in DNA evidence: Irrelevant and prejudicial? *JurimetricsJournal*, **35**, 201–219
Luce, R. D. (1994). Thurstone and sensory scaling: Then and now. *Psychological Review*, **101**, 271–277.
Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., . . . and Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, **56**, 364–370.
McQuiston-Surrett, D., and Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what factfinders hear. *Law and Human Behavior*, **33**, 436–453.
Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., and Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, **37**(3), 197–207.
Morrison, G. S. (2011). The likelihood-ratio framework and forensic evidence in court: A response to R v T. *International Journal of Evidence and Proof*, **15**, 1–29.
Morrison, G. S., and Thompson, W. C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science & Technology Law Review*, **18**, 326–433.

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**, 3–9.
NAS (National Academy of Science). (2009). Strengthening forensic science in the United States: A path forward. Retrieved from http://www.nap.edu/catalog.php? record_id=12589
National Commission on Forensic Science (2015). Views of the Commission: Ensuring that Forensic Analysis Is Based on Task-Relevant Information. Retrieved from https://www.justice.gov/archives/ncfs/file/818196/download
NIST (National Institute of Standards and Technology) Expert Working Group on Human Factors in Latent Print Analysis. (2012). *Latent Print Examination* and *Human Factors: Improving the Practice Through a Systems Approach*. NIST Interagency/Internal Report (NISTIR)-7842.
Pacheco, I., Cerchiai, B., and Stoiloff, S. (2014). Miami-Dade research study for the reliability of the ACE-V Process: Accuracy and precision in latent print examinations. Department of Justice Final Report, Award Number: 2010-DN-BX-K268. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf
Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as *a participant pool. Current Directions in Psychological Science*, **23**, 184–188. doi: 10.1177/0963721414531598
PCAST (President's Council of Advisors on Science and Technology). (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Report to the President. Retrieved from https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/
R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Vienna, Austria: Statistical Computing. Retrieved from https://www.R-project.org.
Risinger, D. M. (2013). Some reservations about likelihood ratios and some other aspects of forensic Bayesianism. *Law, Probability & Risk*, **12**, 63–73.
Robertson, B., Vignaux, G. A., and Berger, C. (2016). *Interpreting evidence: Evaluating forensic science in the courtroom*, 2nd edn. New York: John Wiley & Sons.
SWGDOC (Scientific Working Group for Forensic Document Examination). (2013). Standard Terminology for Expressing Conclusions of Forensic Document Examiners. Retrieved http://www.swgdoc.org/images/documents/standards/SWGDOC%20Standard%20Terminology%20for%20Expressing%20Conclusions%20of%20Forensic%20Document%20Examiners%20150114.pdf
Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, **69**, 3139–3148.
Sjerps, M., and Berger, C. (2012). How clear is transparent? Reporting expert reasoning in legal cases. *Law Probability and Risk*, **11**, 317–329.
Stern, H. S. (1990). A continuum of paired comparisons models. *Biometrika*, **77**(2), 265–273.
Stern, H. S. (1992). Are all linear paired comparison models empirically equivalent? *Mathematical Social Science*, **23**(1), 103–117 doi.org/10.1016/0165-4896(92)90040-C
Thompson, W. C. (1996). DNA evidence in the O.J. Simpson trial. *Colorado Law Review*, **67**, 827–857.
Thompson, W. C. (2009). Painting the target around the matching profile: The Texas sharpshooter fallacy in forensic DNA interpretation. *Law, Probability and Risk*, **8**, 257–276.
Thompson, W. C. (2012). Discussion paper: Hard cases make bad law: Reactions to R v. T. *Law, Probability and Risk*, **11**, 347–359.
Thompson, W. C. (2015). Determining the proper evidentiary basis for an expert opinion: What do experts need to know and when do they know too much? In Robertson and Kesselheim (eds) (pp. 133–150). San Diego, CA: Elsevier, Inc.
Thompson, W. C., Kaasa, S. O., and Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies*, **10**(2), 359–397
Thompson, W. C., and Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law & Human Behavior,* **39**(4), 332–349.

Thompson, W. C., and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior.* **11**, 167–187.
Thompson, W. C., Vuille, J., Biedermann, A., & Taroni, F. (2013). The role of prior probability in forensic assessments. *Frontiers in Genetics*, **4**, 220–223.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273–286.
Ulery B. T., Hicklin R. A., Buscaglia J., and Roberts M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences, USA*, **108**(19), 7733–7738.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process and purpose. *The American Statistician*, **70**(2), 129–133.
Willis, S. M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., . . . and Zadora, G. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. Report by the European Network of Forensic Science Institutes. Retrieved from http://enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science/
Zeller, R., and Carmines, E. E. (1980). *Measurement in the Social Sciences*. New York: Cambridge University Press.