INVESTIGATION · FOUNDATIONS

Evidence Weighting Frameworks for Conflicting Information

Investigations rarely produce evidence that all points the same way. This article explains the main frameworks for weighing conflicting evidence, including source and credibility grading, likelihood ratios, analysis of competing hypotheses, argument charting, story models and evidence hierarchies, and how to choose and combine them.

ModQu research article · 2,339 words · 27 references (APA 7th edition)

Almost every investigation produces conflicting information. Two witnesses describe the same meeting differently. A document says one thing and a database another. A forensic finding points towards a suspect while an alibi points away. Investigators must decide how much weight each item deserves and what conclusion the evidence as a whole supports. Doing this by intuition alone invites well-documented errors, including giving too much weight to vivid or early information and too little to evidence that contradicts a favoured explanation (Nickerson, 1998). Over the past century, lawyers, forensic scientists, intelligence analysts and medical researchers have developed frameworks that make weighing more explicit, consistent and open to review. This article describes the main frameworks, what each does well, and how investigators can use them together.

Why conflicting evidence is hard

Weighing evidence involves several distinct judgements. The investigator must assess whether a source is reliable, whether a specific item of information is credible, how relevant it is to each question, how strongly it supports one explanation over another, and how it combines with other items. Schum (1994) showed that even a single item of testimony rests on a chain of inferences: whether the witness was in a position to observe, whether they observed accurately, whether they remember accurately, and whether they are reporting truthfully. A weakness at any link reduces the weight of the whole. When items conflict, the investigator must also decide whether the conflict reflects error, deception, different perspectives or a genuine ambiguity in events.

People handle these tasks poorly without help. Heuer (1999) described how analysts tend to seize on a single hypothesis, interpret ambiguous evidence as consistent with it, and fail to notice that much of their evidence is equally consistent with other explanations. Dror and Charlton (2006) showed that even experienced fingerprint examiners changed their conclusions about the same prints when given contextual information suggesting a different answer. Frameworks for weighing evidence are, in part, defences against these tendencies.

Grading sources and information

The simplest frameworks separate the reliability of a source from the credibility of a particular item of information. Intelligence services have long used two-part grading systems in which a source is rated on its track record and the information on whether it is confirmed by other sources, consistent with what is known, and logical. The separation matters because reliable sources sometimes pass on wrong information and unreliable sources are sometimes right. Irwin and Mandel (2019) reviewed these systems and found that their categories are often poorly defined, that analysts apply them inconsistently, and that the two ratings tend to be merged in practice. They recommended clearer definitions and numerical expressions of uncertainty.

For investigators, a source-and-credibility grading is a useful first step. Each item in the evidence file can be tagged with its source, how that source obtained the information (direct observation, documents, hearsay), whether it has been corroborated, and any reason to doubt it. Corroboration must be independent: two reports that trace back to the same original source count as one.

Relevance and probative value

Legal systems provide a basic structure for weighing. Evidence is relevant if it makes a fact of consequence more or less probable than it would be without the evidence (Fed. R. Evid. 401). Its probative value is how much it shifts that probability. Courts may exclude relevant evidence if its probative value is substantially outweighed by the risk of unfair prejudice or confusion. These concepts translate directly to investigations. An investigator should ask of each item: which question does it bear on, which way does it point, and how far does it move the answer?

Likelihood ratios

The most precise framework for the weight of a single item of evidence is the likelihood ratio. It compares the probability of observing the evidence if one proposition is true with the probability of observing it if an alternative proposition is true. A ratio of 100 means the evidence is 100 times more probable under the first proposition than the second. The likelihood ratio measures the strength of the evidence itself, separately from how probable the propositions were to begin with (Aitken et al., 2010; Evett & Weir, 1998).

The approach forces investigators to state the alternative explicitly. A fibre match is strong evidence that the suspect's jumper was the source only if such fibres are rare; evidence that is equally likely under both propositions has a ratio of 1 and no weight, however striking it seems. Cook et al. (1998) showed that propositions can be framed at different levels, from the source of a trace to the activity that left it to the offence itself, and that the weight of the same evidence can differ greatly between levels. A DNA profile may strongly support the proposition that a person was the source of a stain, but say much less about how or when the stain was deposited.

Forensic science bodies in Europe recommend expressing evaluative opinions as likelihood ratios, with verbal equivalents to help readers (European Network of Forensic Science Institutes [ENFSI], 2015).

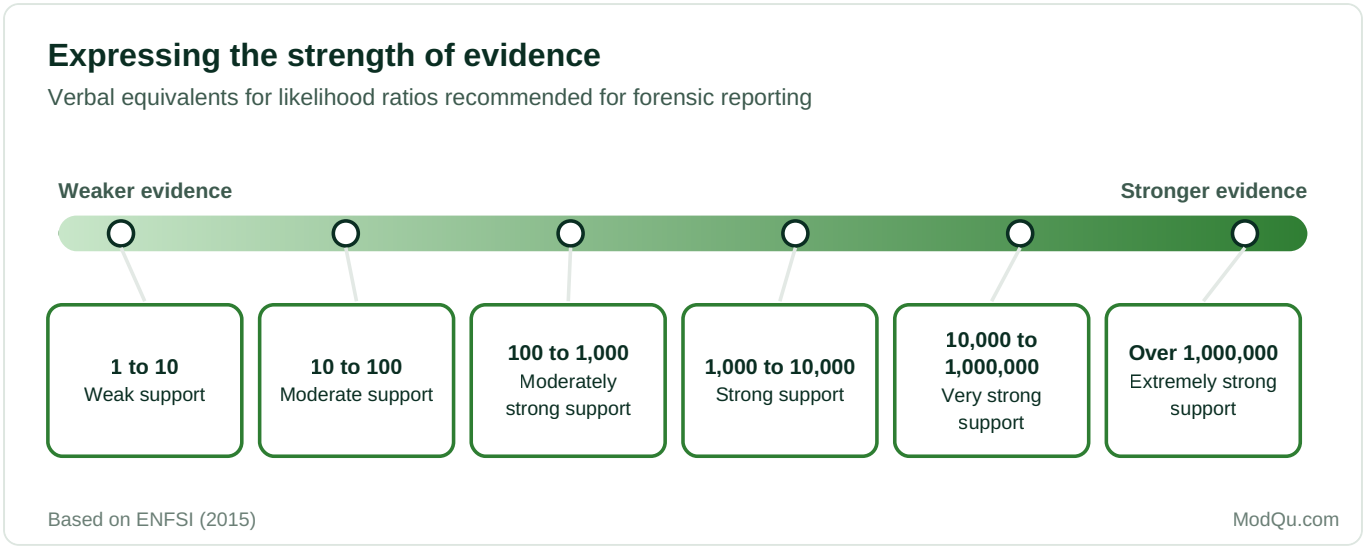


Figure 1. A likelihood ratio says how many times more probable the evidence is under one proposition than another. Verbal scales help readers, but the words should always be tied to the numbers they represent, because people interpret probability words very differently. Diagram by ModQu.

The verbal labels need care. Kent (1964) found that readers interpreted probability words such as "probable" across very wide ranges, and the same problem applies to "strong support". Labels should always be tied to the numbers they stand for. Likelihood ratios also require data on how common the evidence is in relevant populations. In *R v T* (2010), the Court of Appeal of England and Wales criticised an expert who had used likelihood ratios for footwear marks based on databases the court considered inadequate, illustrating that the framework is only as good as its inputs. Thompson and Schumann (1987) showed that people readily confuse the probability of the evidence given innocence with the probability of innocence given the evidence, so even a correct ratio can be misread.

Combining evidence: Bayesian networks

Likelihood ratios can in principle be combined, but only if the items of evidence are independent. Often they are not: two witnesses may have discussed what they saw, or a forensic finding and a confession may both depend on the same disputed event. Bayesian networks represent the dependencies between hypotheses and items of evidence as a graph and calculate how belief in each hypothesis should change as evidence is added. Fenton et al. (2013) proposed a general structure for building such networks for legal arguments, and Lagnado et al. (2013) described reusable "idioms" for common patterns, such as evidence whose reliability is itself uncertain, or alibi evidence.

Bayesian networks make reasoning explicit and can expose errors, such as double-counting dependent evidence. Their weakness is that they require numerical estimates that may be difficult to justify, and their outputs can appear more precise than the inputs warrant. They are most useful for checking the structure of reasoning in complex cases, even when the numbers are approximate.

Analysis of competing hypotheses

Analysis of competing hypotheses (ACH), developed by Heuer (1999) for intelligence analysis, is a structured method for weighing evidence against several explanations at once. The analyst lists all reasonable hypotheses, lists the significant evidence and arguments, and builds a matrix showing whether each item is consistent or inconsistent with each hypothesis. The key insight is that evidence consistent with all hypotheses has no diagnostic value, however much of it there is. The analyst then focuses on refuting hypotheses rather than confirming them, tentatively accepting the one with the least evidence against it, and identifies which items, if wrong, would change the conclusion (Heuer & Pherson, 2015).

ACH suits investigations because it brings alternative explanations into view and counters the tendency to build a case around one suspect. Its limitations include that results depend on which hypotheses and evidence are listed, and that it treats evidence as consistent or inconsistent without fully capturing degrees of weight. Dhami et al. (2015) note that structured techniques such as ACH have rarely been tested for whether they improve accuracy, and call for more research. Even so, the discipline of listing alternatives and asking which evidence discriminates between them is valuable in itself.

Argument charting

Wigmore (1913) devised a method of charting the chain of inferences from each item of evidence to the ultimate facts in issue, including the inferences that support and undermine each link. Anderson et al. (2005) modernised and taught the method as a way to organise complex cases. Toulmin's (1958) model of argument, in which a claim is supported by data through a warrant and qualified by possible rebuttals, offers a simpler version. Tecuci et al. (2016) extended these ideas into computer-supported methods that combine evidence, hypotheses and arguments.

Charting makes visible exactly where conflicts arise. If two witnesses disagree, the chart shows which inferential links differ: perhaps one was not in a position to see, or one had reason to misreport. The investigator can then seek evidence that bears on that specific link rather than on the overall conclusion.

Stories and relative plausibility

Decision-makers often reason not item by item but by building stories. Pennington and Hastie (1992) found that jurors organise evidence into narratives and choose the story that best accounts for the evidence, is coherent and is complete. Allen and Pardo (2019) argue that legal fact-finding is best understood as comparing the relative plausibility of competing explanations, rather than calculating probabilities for each item.

Story-based reasoning has strengths: it captures how evidence fits together, and investigators naturally think this way. It also has dangers. A coherent story can be persuasive even when key parts lack evidence, and once a story is adopted, conflicting evidence tends to be explained away. A good practice is to construct the best possible story for each competing explanation, identify the evidence each story must explain away, and ask which gaps in each story could be tested.

Evidence hierarchies

In medicine and public policy, evidence frameworks rank types of study by their susceptibility to bias. The GRADE approach rates the quality of a body of evidence, starting from the study design and then adjusting for risk of bias, inconsistency, indirectness, imprecision and publication bias (Guyatt et al., 2008). Hill (1965) proposed considerations for judging whether an association is causal, including strength, consistency, specificity, temporal order and a dose–response relationship.

Investigators can adapt these ideas. Some kinds of evidence are generally more resistant to error than others: contemporaneous records are usually more reliable than later recollections, and physical or digital traces are less subject to memory distortion than testimony. But hierarchies are only starting points. Wixted and Wells (2017) showed, for example, that an eyewitness's initial identification made with high confidence under fair conditions can be highly accurate, while later identifications are not. The weight of any type of evidence depends on how it was obtained.

Comparing the frameworks

Frameworks for weighing evidence

Each answers a different question; most investigations benefit from combining several

	Best for	Main limitation
Source and credibility grading	Screening items and tracking corroboration	Categories often vague and merged
Likelihood ratios	Weight of a single item between two propositions	Needs data on how common the evidence is
Bayesian networks	Combining dependent items of evidence	Numbers can look more precise than they are
Competing hypotheses	Keeping alternatives in view	Depends on the hypotheses listed
Argument charts	Locating exactly where evidence conflicts	Time-consuming for large cases
Story comparison	Seeing how evidence fits together	Coherent stories can hide gaps
Evidence hierarchies	Judging types of evidence by bias risk	Ignores how an item was obtained

ModQu analysis of sources cited in this article ModQu.com

Figure 2. The main frameworks for weighing evidence, the task each does best, and its main weakness. Diagram by ModQu.

Using the frameworks together

No single framework handles every task. A practical approach for investigators combines them in stages:

1. **Grade each item.** Record its source, how the source knew, corroboration and any reasons for doubt.
2. **List the hypotheses.** Include every reasonable explanation, not only the leading one.
3. **Identify diagnostic evidence.** Use an ACH-style matrix to find which items discriminate between hypotheses and which are consistent with all of them.
4. **Weigh key items.** For the most diagnostic items, ask how much more probable they are under one hypothesis than another, using likelihood ratios where data exist and reasoned estimates where they do not.
5. **Check dependencies.** Make sure corroborating items are genuinely independent and that the same underlying fact is not being counted twice.
6. **Test the stories.** Build the strongest account for each surviving hypothesis and identify what each must explain away.
7. **Record the reasoning.** State the conclusion, the standard of proof applied, the evidence relied on, the evidence against, and how conflicts were resolved.

Intelligence analytic standards call for exactly this kind of transparency: distinguishing information from judgement, describing the quality of sources, expressing uncertainty and considering alternatives (Office of the Director of National Intelligence [ODNI], 2015).

A short example

The following hypothetical case shows the approach. An internal investigation examines whether a manager leaked confidential pricing data to a competitor. Three items conflict: an IT log shows the file was downloaded from the manager's account, the manager says they were on leave and their password may have been used by someone else, and a colleague says the manager mentioned the competitor's offer shortly before the leak.

Grading shows that the IT log is reliable as to the account but not as to the person; the manager's account is self-interested but specific; and the colleague's recollection is second-hand and was given after rumours had spread. The investigator lists three hypotheses: the manager leaked the file; someone else used the manager's credentials; the file reached the competitor by another route. The download log is consistent with the first two, so it does not discriminate between them. The investigator then seeks evidence that does: the device and location of the download. The records show that the download came from a device registered to another employee, from the office, on a day the manager's leave is confirmed by travel records. The weight shifts decisively to the second hypothesis, and the investigation turns to the other employee. The report records each step, including why the colleague's recollection was given little weight.

Conclusion

Conflicting evidence is the normal condition of investigation. Frameworks for weighing it, from simple source grading to likelihood ratios, Bayesian networks, analysis of competing hypotheses, argument charts, story comparison and evidence hierarchies, each make part of the reasoning explicit. None removes the need for judgement, and each has

limitations. Used together, they help investigators separate reliable from unreliable information, keep alternative explanations in view, identify which evidence actually discriminates, avoid double-counting and explain their conclusions in a way others can check.

References

Aitken, C., Roberts, P., & Jackson, G. (2010). *Fundamentals of probability and statistical evidence in criminal proceedings: Guidance for judges, lawyers, forensic scientists and expert witnesses* (Practitioner Guide No. 1). Royal Statistical Society.

Allen, R. J., & Pardo, M. S. (2019). Relative plausibility and its critics. *The International Journal of Evidence & Proof, 23*(1–2), 5–59. https://doi.org/10.1177/1365712718813781

Anderson, T., Schum, D., & Twining, W. (2005). *Analysis of evidence* (2nd ed.). Cambridge University Press.

Cook, R., Evett, I. W., Jackson, G., Jones, P. J., & Lambert, J. A. (1998). A hierarchy of propositions: Deciding which level to address in casework. *Science & Justice*, *38*(4), 231–239. https://doi.org/10.1016/S1355-0306(98)72117-3

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 10*(6), 753–757. https://doi.org/10.1177/1745691615598511

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification, 56*(4), 600–616.

European Network of Forensic Science Institutes. (2015). *ENFSI guideline for evaluative reporting in forensic science*.

Evett, I. W., & Weir, B. S. (1998). *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sinauer.

Fed. R. Evid. 401.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science, 37*(1), 61–102. https://doi.org/10.1111/cogs.12004

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Heuer, R. J., Jr., & Pherson, R. H. (2015). *Structured analytic techniques for intelligence analysis* (2nd ed.). CQ Press.

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Irwin, D., & Mandel, D. R. (2019). Improving information evaluation for intelligence production. *Intelligence and National Security, 34*(4), 503–525. https://doi.org/10.1080/02684527.2019.1569343

Kent, S. (1964). Words of estimative probability. *Studies in Intelligence, 8*(4), 49–65.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: A framework for evidential reasoning. *Argument & Computation, 4*(1), 46–63. https://doi.org/10.1080/19462166.2012.682656

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Office of the Director of National Intelligence. (2015). *Intelligence Community Directive 203: Analytic standards*.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology, 62*(2), 189–206. https://doi.org/10.1037/0022-3514.62.2.189

R v T [2010] EWCA Crim 2439.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley.

Tecuci, G., Schum, D. A., Marcu, D., & Boicu, M. (2016). *Intelligence analysis as discovery of evidence, hypotheses, and arguments: Connecting the dots*. Cambridge University Press.

Thompson, W. C., & 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*(3), 167–187. https://doi.org/10.1007/BF01044641

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Wigmore, J. H. (1913). *The principles of judicial proof as given by logic, psychology, and general experience, and illustrated in judicial trials*. Little, Brown.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10–65. https://doi.org/10.1177/1529100616686966