Robust Trust in Expert Testimony

Christian Dahlman †
christian.dahlman@jur.lu.se

Lena Wahlberg †
lena.wahlberg@jur.lu.se

Farhan Sarwar †
farhan.sarwar@psychology.lu.se

ABSTRACT

The standard of proof in criminal trials should require that the evidence presented by the prosecution is robust. This requirement of robustness says that it must be unlikely that additional information would change the probability that the defendant is guilty. Robustness is difficult for a judge to estimate, as it requires the judge to assess the possible effect of information that the he or she does not have. This article is concerned with expert witnesses and proposes a method for reviewing the robustness of expert testimony. According to the proposed method, the robustness of expert testimony is estimated with regard to competence, motivation, external strength, internal strength and relevance. The danger of trusting non-robust expert testimony is illustrated with an analysis of the Thomas Quick Case, a Swedish legal scandal where a patient at a mental institution was wrongfully convicted for eight murders.

Keywords: expert testimony, robustness, Thomas Quick case.

1. Robust Evidence

It is obvious that some evidence is better than other evidence, but it is not obvious to everyone that 'better evidence' can mean two different things. A&B can be better evidence than A for a certain hypothesis H, in the sense that the

† Lund University, Sweden.

probability that H is true is increased more by the conjunction of A and B than by A on itself, $P(H|A,B) > P(H|A)$, but A&B can also be better evidence than A for H in the sense that it takes more relevant information into account. These qualities must be distinguished from each other. A&B is always better evidence than A in the latter sense, but it is by no means necessary that A&B is better evidence than A in the former sense. It could just as well be the case that taking account of B decreases the probability of H, $P(H|A,B) < P(H|A)$. The distinction was first observed by Charles Sanders Peirce.

> [...] to express the proper state of belief, not one number but two are requisite, the first depending on the inferred probability, the second on the amount of knowledge on which that probability is based (Peirce 1932, 421).

The notion that evidence is better if it takes more information into account has been explored in different ways, with different terminology. John Maynard Keynes says that new information can improve the 'weight' of the evidence (Keynes 1921, 71), while Peter Gärdenfors and Nils-Eric Sahlin explain the improvement in terms of 'epistemic reliability' (Gärdenfors & Sahlin 1982, 362). Neil Cohen says that more knowledge increases the ´confidence´ in the evidence (Cohen 1985, 405), and Alex Stein says that it makes the evidence more ´resilient' (Stein 2005, 48). We will use the term 'robustness'. In the tradition of Scandinavian legal theory, we will say that the evidence for a hypothesis is more *robust* if it takes more relevant information into account (Ekelöf 1992, 128-129; Strandberg 2012, 515-607).

The robustness of evidence is a question of sensitivity to new information. Robust evidence is not sensitive to new information. It is not likely that additional information will alter the assessment. When the evidence is not robust, we have a situation where the assessment is sensitive to new information. The probability that more information will change the probability of the hypothesis is relatively high.

In criminal trials, the standard of proof requires that it has been proved beyond reasonable doubt that the defendant is guilty. This requirement should entail two conditions. Firstly, the probability of the hypothesis that the defendant is guilty (H_g), given the evidence that has been introduced to the court (E_i), must reach a certain standard (p^*), e.g. 99%.

$$P(H_g|E_i) \geq p^*$$

Secondly, this assessment must be robust. It must be unlikely that the assessment that the probability of the hypothesis reaches the p*-standard would change, in the hypothetical case that further evidence is introduced to the point where all attainable evidence (E_{max}) is given to the court. This condition introduces a second order probability over the first order probability measured by the p*-standard, and sets a standard of proof (p**) for the second order probability.

$$P(P(H_g \mid E_{max}) \geq p^*) \geq p^{**}$$

An increase in the second order probability means that it is more likely that the assessment of the first order probability will hold in the light of additional information. It should be noted that the p*-standard can differ from the p**-standard. It could, for example, be the case that p* = 0.99 and p** = 0.90. The p**-standard should be high enough to prevent conviction in cases where the evidence introduced to the court is too scarce. At the same time, it is important not to set the p**-standard too high. If the p**-standard is extremely high, no defendant can ever be convicted. There is always some risk that further information could change the assessment.

The setting of the p**-standard must also consider the cost of additional evidence. Gathering evidence comes with a cost, and it would be very costly to require that all attainable evidence is given to the court. The cost of additional evidence must, therefore, be balanced against the social cost of wrongful convictions (Keynes 1921, 77). In cases that deal with major offenses, like murder or rape, the social cost of a wrongful conviction is very high. The p**-standard should, therefore, be high in these cases. In cases that concern minor offenses, e.g. traffic violations, the social cost of a wrongful conviction is considerably lower. This is a reason for setting the p**-standard lower in such cases.

A problem with the p**-standard is that it requires us to assess what would happen if we had information that we do not have. This seems almost paradoxical. How can we assess the effect of evidence that we do not have? How can we know something about what we do not know? At a closer look, this is not as strange as it seems. Imagine, for example, that we know that Mrs. Brown was at the scene of the crime when it took place, but we do not know what Mrs. Brown knows about the incident. We have not heard testimony from Mrs. Brown, since she has not been called as a witness. In this situation, it is

possible, that our assessment of the probability that the defendant is guilty, $P(H_g \mid E_i)$, would change dramatically if we get to hear what Mrs. Brown has to tell. Maybe she made some very important observations that exonerate the defendant. It is, of course, also possible that Mrs. Brown's testimony would make no difference at all. Her observations could turn out to be completely irrelevant. We do not know how Mrs. Brown's testimony will affect our assessment of the hypothesis that the defendant is guilty, but our knowledge that she was at the scene of the crime makes it more probable that a testimony from her will change that assessment than a testimony from a randomly picked person in the general population. As a general rule, the probability that a person has some important information to offer is greater if we know that the person was present at the scene of the crime when the crime took place. It is generalizations like these that make it possible to assess the second order probability against the p**-standard. We know from experience that some inquiries are more likely to produce information that will change the first order probability than other inquiries. The probability that further investigations will change the assessment of the first order probability is higher, if there is a possible line of inquiry that has not been explored, but which, according to general experience, has the ability to change the assessment. Hearing the testimony of a person who was at the scene of the crime when it was committed is an example of such an inquiry. Another example of a kind of inquiry that typically has the ability to change the assessment of the first order probability is DNA profiling. In cases where DNA profiling is possible but has not been conducted, the probability that the defendant is guilty is sensitive to the additional information that a DNA investigation would produce.

The assessment of the second order probability can be summarized as follows. If a certain piece of evidence has not been introduced to the court, in spite of the fact that this would have been possible and the evidence is of a kind that, generally, is likely to change the probability that the defendant is guilty, the evidence assessment based on the existing evidence is not robust. If, on the other hand, there is no potential piece of evidence that is likely to make such a difference, the existing evidence is robust. The dimension of robustness applies to all types of evidence. In the remainder of this article, we will discuss the usefulness that this notion can have to court assessments of expert testimony.

2. Robustness and Expert Testimony

Expert testimony plays a crucial role in many legal cases, but is by definition difficult for non-experts to assess. It goes without saying that judges and juries to a large extent must trust the opinions of an expert. At the same time, it would be naïve to think that an expert's opinion is always correct - there are both empirical and theoretical reasons to take seriously the risk that an expert witness goes wrong (see e.g. Huber 1993, Angell 1997, Meester et al. 2006, Dwyer 2008, Wahlberg 2010 a and b, and Råstam 2012). Just like non-experts, some alleged experts are dishonest or incompetent, and even an honest and competent expert can commit a reasoning error, disregard relevant studies or misunderstand the factual question raised by a legal norm. If experts were trusted indiscriminately, many verdicts would hence rest on inadequate facts.

Ideally, therefore, the trust that judges and juries put in experts should be critical, not blind. Yet, the idea that experts should be trusted "critically" has a paradoxical flavour. Judge Learned Hand made this paradox explicit in an article published in *Harvard Law Review* in 1901. According to Hand, the jury is placed in an impossible position when the prosecution and the defence call expert witnesses that make contradictory statements and the jury has to assess which expert to trust.

> [...] how can the jury judge between two statements each founded upon an experience confessedly foreign in kind to their own? It is just because they are incompetent for such a task that the expert is necessary at all. [...] Knowledge of such general laws can be acquired only from a specialized experience such as the ordinary man does not possess [...] The jury by hypothesis have no such experience directly, it being of a kind not possessed by ordinary men [...] Therefore, when any conflict between really contradictory propositions arises, or any reconciliation between seemingly contradictory propositions is necessary, the jury is not a competent tribunal. [...] [the jury] will do no better with the so-called testimony of experts than without, except where it is unanimous. (Hand 1901, 54-56)

As a response to the paradox, Hand proposed that juries should be composed of experts: In a case of poisoning, the jury would be composed of people with expert knowledge in toxicology, in a case of murder by arson the jury would be composed of people with special knowledge on fires, and so on. In such a system, expert witnesses would no longer be necessary. Hand's proposition was never adopted by the American legal system. On the contrary, the use of

expert witnesses in court has increased tremendously (Graham 1977, 35), and experts' testimony is still assessed by juries and courts and not by peers. Fortunately, the paradox that Hand puts forward can be solved - at least in part. Jurists and philosophers who have engaged with the problem of when to trust an expert witness have put forward a number of tools that can be used by non-experts to assess the reliability of putative experts. Below, we will present and systematize these tools. More importantly, we will show that the notion of robustness makes a valuable contribution to this toolbox. On our definition, the robustness of scientific evidence *qua* evidence for a certain hypothesis is a measure of how likely it is that additional attainable evidence would alter the probability of the hypothesis. Hence, and as explained in more detail below, the robustness of scientific evidence is in part a measure of the extent to which the available tools for evaluating expert testimony have been put into use. In other words, the notion of robustness allows courts to assess scientific evidence *without* actually putting the suggested tools into use. This is important, considering that applicable procedural rules often constrain courts' mandate to initiate further investigations.

The crucial question of when to trust an expert has engaged both jurists and philosophers. Within common law, judges and legislators have developed criteria for the admissibility of expert testimony. A well-known example is the so-called *general acceptance test* which was first laid down in *Frye v. United States 293 F. 1013, D.C. Circ., 1923*. The Court in *Frye* held that in order to be admissible, expert testimony must be based on scientific principles and discoveries that are "sufficiently established to have gained general acceptance in the particular field" (at 1024). Another, more recent example is *Daubert v. Merrell Dow Pharmaceuticals 509 U.S. 579 (1993)*, where the court referred to the works of Karl Popper and Carl G Hempel and identified testability, peer review, error rate and general acceptance as criteria for determining the reliability of expert testimony. In *General Electric Co. v. Joiner, 522 U.S. 136 (1997)*, the Supreme Court later stated that nothing in the Daubert guidelines requires a court "to admit opinion evidence which is connected to existing data only by the *ipse dixit* of the expert" (at 137), and thereby implicitly encouraged courts to scrutinize the inferences underlying expert testimony. In the subsequent *Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999)*, the Court explained that the Daubert criteria might apply to non-scientific expert testimony too, depending on "the particular circumstances of the particular case at issue" (at 150). (Similarly, the discussion in this article focuses on

robustness assessment of scientific expert testimony, but is applicable to relevantly similar non-scientific evidence too.)

In philosophy, Douglas Walton, Alvin Goldman and others have contributed to the development of criteria by which non-experts can evaluate an expert's statement. Walton has devised a list of critical questions that non-experts can use to challenge an argument from expert opinion. The list includes questions regarding the alleged expert's education, experience and personal reliability (Walton 1997, 223):

> *Expertise* question: How credible is E as an expert source?
> *Field* question: Is E an expert in the field that A is in?
> *Opinion* question: What did E assert that implies A?
> *Trustworthiness* question: Is E personally reliable as a source?
> *Consistency* question: Is A consistent with what other experts assert?
> *Backup evidence* question: Is A's assertion based on evidence?

Similarly, Goldman has identified and discussed five sources of evidence that a non-expert can use in determining the reliability of expert testimony: "arguments presented by contending experts", "agreement from additional putative experts", "appraisal by 'meta-experts' of the expert's expertise", "evidence of the expert's interests and biases" and "past track records" (Goldman 2001, 93).

The referred discussions identify measures that a non-expert can take to review an expert's testimony. Roughly put, the notion of robustness adds to this picture that it estimates the relevance of the inquiries made, as compared to those omitted. Hence, a robustness estimate requires consideration not only of the degree to which possible inquiries have been performed, but also of the omitted inquiries' capacity to alter the probability of the hypothesis.

The various inquiries so far discussed are to a large extent over-lapping. In this section, we will propose a tentative taxonomy (summarized in Figure 1 below) arranged according to the different aspects of expert testimony that these inquiries address. As will be elaborated below, this taxonomy can be used as a basis for courts' robustness assessments. First, we note that some of the measures for reviewing expert testimony that have been put forward in the literature relate to the reliability of the expert's *person* whereas others relate to the reliability of the expert's *reasoning*. These two different objects provide the first partition in our taxonomy. We will refer to reviews of the expert's person

as *ad hominem reviews* and to reviews of the expert's reasoning as *de re reviews*.

Ad hominem reviews challenge the reliability of an expert's opinion by drawing attention to attributes of the expert's person that put her reliability into question. This kind of review is likely to be the most obvious way for a non-expert to confront an expert's opinion. An argument that attacks an arguer's person rather than her reasoning is often treated as a fallacy (Copi and Cohen 2002, 143). However, in contexts like the present, where the arguer's reliability as a source is a relevant factor for trusting her conclusion in the first place, drawing attention to personal attributes that affect her reliability is both relevant and warranted (Walton 1997, Hahn et al. 2009, Dahlman et al. 2011, Dahlman and Wahlberg 2015). Roughly, attributes of relevance to *ad hominem* reviews can be divided into two categories: those that relate to competence, and those that relate to motivation (Dahlman and Wahlberg 2015). The expert's *competence* is of obvious relevance to her reliability and moreover a factor that is relatively easy to assess. Not surprisingly, then, this is a factor that is frequently highlighted in discussions on the reliability of expert testimony. Thus, we have seen that Douglas Walton points out that we can ask critical questions about the expert's experience, education, and field of expertise (Walton 1997), and that Alvin Goldman advises us to make use of meta-experts and past track records to assess the expert's reliability in this respect (Goldman 2001). Similarly, professional organisations have carved out standards that their members must meet when testifying as expert witnesses. For example, the American Psychological Association's (2013) and the British Psychological Associations' (2010) demand that expert witnesses possess the psychological and legal knowledge, experience, training, and required skills to perform the requested expert role. The competence should be established either by professional certification or by providing proof of active practice and up-to-date knowledge in the area in which the expertise are requested. An expert's *motivation* is perhaps more difficult to control, but likewise a factor that is regularly stressed as relevant. Walton (1997) and Goldman (2001) both point at the importance of taking into account evidence of the expert's interests and biases, and many legal rules and policies go as far as to treat secondary interests as reasons for disqualification.

In contrast, *de re* reviews address not the expert's person, but her reasoning. By definition, an expert's reasoning is in part a result of knowledge and skills that the non-expert lacks. *Prima facie*, it is more difficult for a non-

expert to call in question an expert's reasoning than to call in question the expert's competence or motivation. At closer look, however, several ways in which non-experts can contest an expert's reasoning can be discerned. To begin with, *de re* reviews can address the *external strength* of the expert's assumptions and conclusions and examine how her opinion relates to external factors, such as available evidence and the views of other experts. Thus, Walton (1997) suggests assessors to ask questions about the evidence that backs up the expert's assertion as well as about how well the assertion accords with the views of other experts. Goldman (2001) recommends consideration of arguments presented by contending experts and the level of agreement from additional putative experts. Another example of the relevance of external strength is provided by the so called *general acceptance test* formulated in *Frye v. United States 293 F. 1013, D.C. Circ., 1923*, in which the Court ruled that to be admissible, expert testimony must be based on scientific principles and discoveries that are "sufficiently established to have gained general acceptance in the particular field". In the superseding case *Daubert v. Merrell Dow Pharmaceuticals 509 U.S. 579 (1993)*, the Court mentioned peer review and general acceptance, which both relate to external strength, as criteria relevant for determining the reliability of expert testimony.

A *de re* review can also address the *internal strength* of an expert's reasoning. For example, the review can assess the consistency of the expert's own reasoning and examine to what extent the expert's conclusion follows from her premises. It should be noted that this kind of assessment addresses formal properties of the expert's reasoning and therefore does not necessarily require additional expert knowledge. In this vein, the Court in *General Electric Co. v. Joiner*, 522 U.S. 136 (1997) held that "a court may conclude that there is simply too great an analytical gap between the data and the opinion proffered" (at 146). As a parallel, many professional guidelines, such as the British Psychological Society's guidelines for psychologists as expert witnesses (2010) require experts to provide the court with criteria that allow the court to evaluate the basis of the expert's opinion (Standard 1.5). Internal strength can potentially also be assessed by generic quality criteria for scientific evidence. For example, in the spirit of Hempel and Popper, the *Daubert* court mentioned *testability* as a relevant criterion for assessing the reliability of scientific evidence. The idea seemed to be that testability is an intrinsic quality, which can be assessed *a priori*, without considering further evidence.

Finally, *de re* reviews can address the *relevance* of the expert's reasoning by assessing its relation to the questions at stake in the particular case (Walton 1997). An assessment of relevance requires that the expert's statement is sufficiently transparent to allow for inter-disciplinary comparisons. Many expert statements, such as "the accident didn't cause A's disability" may *appear* transparent but in fact contain implicit assumptions and values. There is hence a risk that these statements conceal significant epistemological differences between legal and scientific standards of proof, as well as ontological differences between legal and scientific notions (such as *cause* and *disability*), which can hinder effective cross-discsiplinary communication. (See for example Cranor 1993, and Shrader-Frechette and McCoy 1993. See also Wahlberg 2010 a and b for a comprehensive discussion on epistemological and ontological differences between law and science).

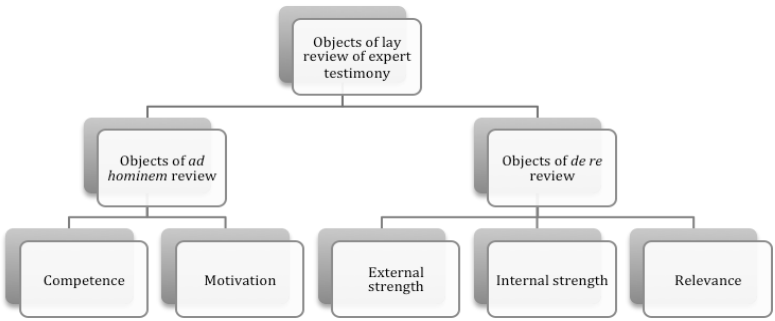

Figure 1. Lay review of expert testimony

Figure 1 above summarizes the objects of the above-discussed inquiries for reviewing expert testimony. A taxonomy of this sort can certainly be of help for those who might want to make further inquiries into the reliability of the expert's opinion. However, this taxonomy is also a potentially useful tool for those assessing the robustness of inquiries already performed. A robustness assessment is an estimate of how sensitive the current probability of the hypothesis is to additional inquiries, and the now proposed taxonomy can hence serve as a check-list for considering what inquiries a non-expert *could*

make. The assessor should then ask herself 1) to what extent such inquiries *have* been performed and 2) how likely it is that they (given her knowledge of their typical relevance) *would* alter the current probability of the hypothesis, if performed. For example, an assessor should ask to what extent measures such as further control of the expert's secondary interests (*ad hominem* review pertaining to motivation), and consultation of additional experts on the same topic (*de re* review pertaining to external strength) are likely to alter the probability of the statement to which the expert testifies. A typical case of low robustness with respect to external strength is at hand when the expert has stated her opinion but not disclosed the assumptions and premises that the opinion is based on, or explained what degree of support these premises and assumptions have in the scientific community. Insufficient robustness means that the evidence should be deemed not to meet the standard of proof required. If, on the other hand, it is likely that the current probability of the hypothesis will sustain in the light of additional evidence, the present evidence is robust. In the remainder of this article, we will show how a robustness evaluation along these lines could have been put into use by the courts in the infamous Swedish Thomas Quick cases, by many considered as the biggest scandal in Swedish legal history.

3. The Thomas Quick Case

The Säter Clinic is a psychiatric care facility in Mid Sweden, located in the Dalarna forest 200 km north west of Stockholm. It is a high security facility that treats convicted criminals who have been sentenced to forensic psychiatric care. In 1992 one of the patients at the Säter Clinic was a 42-year-old man called Thomas Quick. The name on his birth certificate was Sture Bergwall, but he had legally changed his name to Thomas Quick to disassociate himself from his father. Quick had been convicted for armed robbery, assault with a deadly weapon, and several sexual offences against young boys. He had been diagnosed with personality disorder and *pedofilia cum sadismus*. In the spring of 1992 Thomas Quick read a newspaper article about an unsolved police case from 1980, the disappearance of an eleven-year-old boy, Johan Asplund, in Sundsvall. The police suspected that Johan Asplund had been abducted and possibly killed, but in spite of extensive investigations his body had not been found. Thomas Quick told his therapist at Säter that the newspaper article

about Johan Asplund's disappearance gave him very uneasy feelings. He was not sure, but he had a feeling that he was responsible for what had happened to Asplund. Over the course of the following months Quick reached a point in his therapy sessions where he confessed that he had killed Johan Asplund, chopped up the body, and buried the pieces (Råstam 2012, 118-122). Quick said to his therapist that he wanted to take responsibility for his actions, and wished to contact the police.

The first police interview was conducted at Säter in March 1993. The police were impressed by Quick's vivid story of the killing, and he was escorted to Sundsvall to show the police to the location where he had buried the remains of Asplund. The police made several excavations on locations indicated by Quick, but no body parts or other evidence was found. After several months of interviews with Quick the police were stuck with a confession that was not backed up by any forensic evidence. In the meantime, Quick had confessed to several other killings. One of them was the murder of Charles Zelmanovits, a fifteen-year-old boy from Piteå who had disappeared in 1976. Some parts of Zelmanovits body, his skull and some bones dressed in decomposed clothes, had recently been found in the woods north of Piteå, and in September 1993 several Swedish newspapers had published articles about the unsolved case and the findings in the woods. The police interviewed Quick about Zelmanovits, and Quick explained that he had killed Zelmanovits and buried parts of the body in different places. A problem with Quick's confession was that he was not able to remember any details that would confirm that his confession was genuine. He was asked to describe the clothes that Zelmanovits was wearing but was not able to recall them correctly. In April 1994 the police called in Sven Å Christianson to help out with the investigation (van der Kwast 2015, 41-42). Christianson was a professor in psychology at Stockholm University, and an expert in issues related to memory. He suggested to the police that they should use a method known as the 'cognitive interview' (Fisher & Geiselman 1992) to help Quick recall details that he had difficulties to remember. In a cognitive interview the interviewer uses various techniques to have the subject mentally recreate and reenact an event. Sven Å Christianson describes in one of his scientific publications how the cognitive interviews with Quick were conducted.

> [...] memories of smells, body positions, various sounds and emotions were triggered. After the reinstatement of his internal context, he [Quick] showed strong emotions and could describe vivid memories of the killings. He was

> able to give specific details, which he had not had access to in previous interrogations (Christianson & Engelberg 1997, 241).

In November 1994 Thomas Quick stood trial for the murder of Charles Zelmanovits. The case presented by the prosecutor, Christer van der Kwast, consisted of Quick's confession, testimony from Detective Sergeant Seppo Penttinen, and an expert testimony from Professor Christianson. There was no forensic evidence. Penttinen testified about the interviews that he had conducted with Quick, and said that Quick had described several details about the vegetation on the location in the forest where the skull and bones had been found, and some details on how the remains of Charles Zelmanovits had been buried. Penttinen testified that Quick had been able to describe these details without information or help from him or other police officers. According to the prosecution, this proved that Quick had knowledge about the crime that only the killer could have, and, thereby, proved that Quick must be the killer.

Christianson had written an expert opinion about Thomas Quick that was submitted as evidence by the prosecution. It addressed the issue of false confessions and listed three circumstances that have been established by science to indicate the possibility of a false confession: 1) situations where the confessor is seeking attention, 2) situations where the confessor has something to gain by confessing, and 3) situations where the confessor is unsure about his own memory and is convinced by others that he is guilty.[1] Christianson was called by the prosecution as an expert witness and testified that Quick's confession was genuine (Josefsson 2013, 367). According to Christianson, there were no circumstances in the Quick case that indicated a false confession. Thomas Quick's defense attorney, Claes Borgström, did not question the prosecutor's case and did not bring in any evidence against it. Quick had instructed Borgström that he wanted to be convicted, and Borgström assisted him in accordance with this instruction. On 16 November 1994 Thomas Quick was found guilty of the murder of Charles Zelmanovits, and was sentenced to continued psychiatric care. The court says in its verdict that the testimony by Penttinen strongly supports that the murder was committed by Quick, and the testimony by Christianson strongly supports that Quick's confession was genuine.[2]

[1] Piteå Tingsrätt, B 179/94, Christianson, S.Å., Sakkunnigyttrande angående betingelser för Thomas Quicks (500426-7190) utsaga i psykologiskt avseende, p. 2-3.

[2] Piteå Tingsrätt, B 179/94, Dom 1994-11-16, p. 11-12.

Quick continued his treatment at the Säter Clinic and continued confessing murders in unsolved cases. In late 1994 he confessed to the murder of two Dutch hikers, Marinus Stegehuis and Janni Stegehuis, who had been stabbed to death in a tent at Lake Appojaure in Lapland, in the summer of 1984. Quick stood trial for double murder in January 1996, and was found guilty. Just like the Zelmanovits case, there was no forensic evidence. Quick was convicted on the testimony of Detective Sergeant Penttinen and Professor Christianson. For this trial, Christianson handed in an expert opinion that ended with the following statement.

> In this report I have discussed false confessions of various types ... Each and every one of these types fit badly with the circumstances of Thomas Quick's confession.[3]

Over the following years, Quick was convicted for five more murders that he had confessed. In May 1997, he was convicted for the murder of Yenon Levi, an Israeli tourist found dead in Hedemora, in June 1998 he was convicted for the murder of Therese Johannesen, a nine year old Norwegian girl who had disappeared in Drammen, in June 2000 he was convicted for the murders of Trine Jensen and Gry Storvik, two young Norwegian women who had been found dead on a parking lot in Oslo, and, finally, in June 2001, he was convicted for the murder of Johan Asplund, the very first murder he had confessed. None of the cases relied on forensic evidence. In each case, the court found that Quick's confession and the testimonies of Detective Sergeant Penttinen and Professor Christianson were enough a guilty verdict.

Parallel to his work as a consultant to the police Sven Å Christianson also took interest in Thomas Quick as a scientific research subject. Christianson was interested in the psychology of serial killers, and interviewed Quick in detail about his childhood and his emotions when he killed his victims. The result of this research was published by Christianson in a book (Christianson 2010, 401-421) entitled *Inside the Head of a Serial Killer* (*I huvudet på en seriemördare*).

However, not everyone was convinced that Thomas Quick was guilty. Some sceptics said that serial killers normally follow some sort of pattern, and pointed out that no such pattern could be seen in the killings that Quick had been convicted for. Some victims were men, others were women. Some victims

[3] Gällivare Tingsrätt, B 26/95, Christianson, S.Å., *Sakkunnigyttrande angående betingelser för Thomas Quicks (500426-7190) utsaga i psykologiskt avseende*, p. 9. See, also, Råstam 2012, p. 256.

where children, others were adults. The modus operandi was different for each crime. Some victims were strangled, other victims were stabbed, and some were clubbed to death with a heavy object. The crimes had been committed at various geographical locations, spread all over Sweden and Norway. Another circumstance that raised doubt about Quick's guilt was the sheer number of confessions. By 2001 he had confessed to 39 killings, and not even Sven Å Christianson believed that all of them were genuine (Christianson 2010, 86). In some cases, it was obvious that Quick's confession did not correspond to the truth. For example, he confessed that he had killed two Somali boys that were reported missing in 1996, not knowing that the boys had later been found and were alive and well. So, if some of Quick's confessions were false, could it not be the case that they were all false? In 2008, Hannes Råstam, an investigative reporter working for Swedish Television (SVT), started to take interest in the Quick case. Råstam went to the Säter Clinic to interview Quick, who had now changed his name back to his birth name Sture Bergwall. During the course of these interviews, Quick/Bergwall confided in Råstam that all of his confessions were false. He explained to Råstam that the psychologists at Säter and the police rewarded him for his confessions by giving him their undivided attention, and granting him extra doses of the medicine that he asked for (Råstam 2012, 94; Josefsson 2013, 461). At the time, Quick was heavily addicted to benzodiazepines. They even rewarded him for his confessions by granting him a leave of absence to go to Stockholm for a couple of days (a rather imprudent decision, considering that they were dealing with a confessed serial killer). Råstam digged deeper into the Quick confessions than anyone had done before, and unearthed a number of things that undermined the prosecution's case. Råstam went through transcripts and videos from the interviews that Seppo Penttinen had conducted, and showed that they were full of leading questions that had helped Quick "remember" the right details (Råstam 2012, 198-202, 206, 222-224, 268-269, 283-284, 297-299, 312-316, 322-325, 370-374).

In 2009 Sture Bergwall requested the court of appeal to order a retrial in the case of Yenon Levi. The request had been worked out by defense attorney Thomas Olsson, and was based on the withdrawal of the confession, in combination with the weaknesses in the police investigation that Råstam had exposed. The request for a retrial was granted, and in September 2010 Quick/Bergwall was acquitted for the murder of Yenon Levi. Subsequent requests for retrials were handed in for all of the other convictions. In June

2011 Quick/Bergwall was acquitted for the murder of Therese Johannesen, in August 2012 he was acquitted for the murder of Johan Asplund, in November 2012 he was acquitted for the murders of Trine Jensen and Gry Storvik, in July 2013 he was acquitted for the murders of Marinus and Janni Stegehuis, and in November 2013 he was acquitted for the murder of Charles Zelmanovits.

The Thomas Quick case is by many considered to be the greatest scandal in the history of Swedish criminal law, and the people responsible have been massively criticized. Christer van der Kwast, who prosecuted all of the cases, has been criticized for leading the investigation in a way that was strongly biased towards the theory that Quick was guilty, Seppo Penttinen has been accused of committing perjury, when he testified that he had not asked leading questions in his interviews, Christianson has been blamed for architecting the fatal interview technique practiced by Penttinen, and Borgström has been criticized for his passive performance as Quick's defense attorney. In contrast, the judges who convicted Quick have not been criticized. The general view seems to be that the judges cannot be blamed, since Quick appeared to be guilty, given the evidence that was presented to them in court. The evidence that exonerated Quick surfaced afterwards, and, as it were, you cannot blame the judges for not taking account of information that they did not have at the time. This line of reasoning overlooks that the standard of proof should not only require that the probability that the defendant is guilty, given the evidence that has been presented, meets the p*-standard. The standard of proof should also require that the probability that this assessment would not be changed by additional information meeta the p**-standard. The evidence must be robust. The evidence presented in the trials against Thomas Quick did not meet this requirement. The judges who convicted Quick should not have trusted Penttinen and Christianson blindly. The judges should have reviewed Penttinen's and Christianson's reliability with regard to *competence* and *motivation*, and they should have reviewed the value of their testimonies critically, with regard to *external strength*, *internal strength* and *relevance* (figure 1 above).

Penttinen testified that he had not helped Quick with leading questions, and the judges trusted him. If the judges had reviewed his testimony more critically with regard to *motivation*, they would have realized that Penttinen was in fact evaluating himself when he made this statement, and was therefore motivated to cover up any mistake that he might have made during the interviews. The transcripts and videos from the interviews should have been

studied by an independent expert, to check if Penttinen had helped Quick with leading questions. The absence of such an inquiry made the case weak with regard to robustness, and the judges should have realized this, instead of trusting Penttinen uncritically. If the transcripts and videos from the interviews had been evaluated by an independent interrogation expert, that expert would have found, like Råstam, that the interviews were full of leading questions.

Christianson testified that none of the circumstances that indicate false confessions were present when Thomas Quick made his confessions. As we have seen, Christianson made this assessment on the assumptions that Quick was not seeking attention and had nothing to gain from confessing. The evidence that was presented to the court included no information that supported these assumptions, and no information that supported the opposite assumptions. In fact, the evidence that the verdicts were based on did not include inquiries into these matters. If such evidence had been brought in, the court would have seen evidence to the effect that Quick was seeking attention and evidence that he had plenty to gain from confessing. If the newspaper reporters who covered the Quick case had been called to testify about their experiences when interviewing Quick, they would have testified that he was seeking attention. Quick was very keen to be interviewed and always made very theatrical statements that made good headlines ("I am an Evil Man"[4], "I Must Carry my Guilt"[5] etcetera). Quick even wrote newspaper articles himself about his case.[6] If the psychiatrists that treated Quick at the Säter Clinic had been called to testify, they would have informed the court that Quick was rewarded for his confessions with extra medication. If the medical records had been introduced as evidence, they would have shown that Quick received extra doses of benzodiazepines as a payment for his confessions. There is an entry in Quick's medical records from 1994 that reveals that the prosecutor van der Kwast put pressure on the staff at Säter to give Quick the pills he asked for, with the argument that Quick "must receive something in return" (Råstam 2012, 157). The judges never received the information that such inquiries would have produced, and they should have realized that the evidence that they were given was not robust, since these lines of inquiry had not been pursued. They should have realized that the missing lines of inquiry typically have the

[4] *Expressen*, 2 September 1994 ("Jag är en ond man").

[5] *Dagens Nyheter*, 10 April 1995 ("Jag måste bära min skuld").

[6] E.g. *Dagens Nyheter*, 12 July 1994 ("Jag flydde för att dö") and 1 January 1995 ("Jag kan bli frisk").

ability to change the picture completely. The probability that the case against Thomas Quick would not hold for additional information was so high, that he should not have been convicted on the existing evidence. As we have seen, an expert testimony based on assumptions that are not supported by evidence does not have the *external strength* to pass as robust evidence. It should be noticed that a critical review of the assumptions in Christianson's expert testimony can be conducted without expertise in psychology. As we observed above, it is problematic for judges to review expert testimony critically, as judges lack expert knowledge on the issue of the testimony (Learned Hand's Paradox). With regard to Christianson's assumptions, however, this does not pose a problem. You do not need any special expertise in psychology to see that a confessor who is rewarded with extra doses of a drug that he is addicted to has something to gain by confessing.

Furthermore, if the reliability of Christianson as an objective scientific expert had been critically reviewed with regard to *motivation*, the judges would have learned that Christianson also had another relationship with Quick. He was interviewing Quick for his study on the psychology of serial killers. This clearly put Christianson in a conflict of interests. The study of Quick as a serial killer relied on the assumption that Quick was guilty. If his confessions were false the entire study would be worthless. Christianson therefore had a strong personal interest in Quick being guilty. This information obviously undermines Christianson's reliability as an objective scientific expert. The court should not have trusted Christianson blindly. They should have reviewed his reliability critically, and they should have realized that the lack of inquires with regard to his motivation weakened the robustness of the evidence provided by his testimony.

In conclusion, the case against Thomas Quick was not robust enough for a conviction. It did not meet the p**-standard in any of the murders that he confessed. The judges who found him 'guilty beyond reasonable doubt' did not apply the standard of proof correctly. They should have acquitted him. Someone might say that this harsh criticism is unfair, since it is passed in hindsight, with all the information uncovered by Råstam and others, that the judges who convicted Quick did not have. We disagree with this defense of the incorrect convictions. It is true that the judges who convicted Quick did not know all that we know today, but they knew that they lacked information on many things, where additional evidence typically makes a difference, and they should have realized that this made the case against Quick insufficiently robust.

ACKNOWLEDGMENTS

Research funded by the Swedish Research Council and the Ragnar Söderberg Foundation.

REFERENCES

Angell, M. (1996). *Science on Trial : the Clash of Medical Evidence and the Law in the Breast Implant Case*, New York: W.W. Norton.

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *Am Psychol*, 68(1), 7-19. doi: 10.1037/a0029889

British Psychological Society: Expert Witness Advisory Group (2010). *Expert Witnesses: Guidelines and Procedure for England and Wales*, Leicester.

Christianson, S.Å., Engelberg, E. (1997). Remembering and Forgetting Traumatic Experiences – A Matter of Survival, in Conway, M. (ed.) *Recovered Memories and False Memories,* Oxford University Press. Oxford. 230-250.

Christianson, S.Å., Sakkunnigyttrande angående betingelser för Thomas Quicks (500426-7190) utsaga i psykologiskt avseende

Christianson, S.Å. (2010). *I huvudet på en seriemördare*. Norstedts. Stockholm.

Cohen, N. (1985). *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*. New York University Law Review. 385-422.

Copi, I. and Cohen, C. *Introduction to Logic*. 11th ed. Upper Saddle River: Prentice Hall, 2002.

Cranor, C. (1993). *Regulating toxic Substances: a Philosophy of Science and the Law*, New York: Oxford University Press.

Dagens Nyheter, 12 July 1994, "Jag flydde för att dö".

Dagens Nyheter 1 January 1995 "Jag kan bli frisk".

Dagens Nyheter, 10 April 1995, "Jag måste bära min skuld".

Dahlman, C., Reidhav, D. & Wahlberg, L. (2011). Fallacies in Ad Hominem Arguments. *Cogency*, 3(2), 105-124.

Dahlman, C., & Wahlberg, L. (2015). Appeal to Expert Testimony – a Bayesian Approach in Bustamante, T. et al (Ed.) *Argument Types and Fallacies in Legal Argumentation*, Springer, Dordrecht.

Dwyer, D. (2008). *Judicial Assessment of Expert Evidence*. Cambridge: Cambridge University Press.

Ekelöf, P.O. (1992). Rättegång IV. 6th ed. Norstedts. Stockholm.

Expressen, 2 September 1994, "Jag är en ond man".

Fisher, R. & Geiselman, R. (1992). *Memory Enhancing Techniques for Investigative Interviewing – The Cognitive Interview*. Thomas. Springfield (Il.).

Goldman, A. (2001). Experts: Which Ones Should You Trust?. *Philosophy and Phenomenological Research*, 63(1), 85-110.

Graham, M. H. (1977). Impeaching the Professional Expert Witness by Showing Financial Interest. *Indiana Law Review*, 53, 35-53.

Gärdenfors, P. & Sahlin, N. (1982). Unreliable Probabilities, Risk Taking and Decision Making, *Synthese*, 53, 361-386.

Hahn, U., Oaksford, M., & Harris, A. (2013). Testimony and Argument: A Bayesian Perspective. In F. Zenker (Ed.) B*ayesian Argumentation: The practical side of probability*, Dordrecht: Springer, 15-38.

Hand, L. (1901). *Historical and Practical Considerations Regarding Expert Testimony*, Harvard Law Review, 15(1), 40-58.

Huber, P. (1993). *Galileo's Revenge: Junk Science in the Courtroom*, Basic Books.

Josefsson, D. (2013) *Mannen som slutade ljuga*. Lind & Co. Stockholm.

Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan. London.

van der Kwast, C. (2015) *Bortom rimligt tvivel*. Bladh by Bladh. Stockholm.

Meester, R., Collings, M., Gill, R., van Lambalgen, M. (2006). On the (ab)Use of Statistics in the Legal Case Against Nurse Lucia de B. *Law, Probability and Risk*, 5(3-4), 233-250.

Peirce, C.S. (1932) *Collected Papers. vol 2*. Harvard University Press. Cambridge Ma.).

Råstam, H. (2012). *Fallet Thomas Quick – Att skapa en seriemördare*, Stockholm: Ordfront. Published in English as *Thomas Quick: The Making of a Serial Killer*. Canongate. Edinburgh, 2013

Shrader-Frechette, K.S. & McCoy, E.D. (1993) *Method in Ecology: Strategies for Conservation*, Cambridge: Cambridge University Press.

Stein, A. (2005). *Foundations of Evidence Law*. Oxford University Press. Oxford.

Strandberg, M. (2012). *Beviskrav i sivile saker*. Fabbokforlaget. Bergen.

Wahlberg, L. (2010 a). *Legal Questions and Scientific Answers: Ontological Differences and Epistemic Gaps in the Assessment of Causal Relations*, Lund: Lund University.

Wahlberg, L. (2010 b). Rätt svar på fel fråga: Typ III fel vid användningen av expertkunskap. *Juridisk Tidskrift* 4, 889-900.

Walton, D. (1997). *Appeal to Expert Opinion*, University Park: Pennsylvania State University Press.

TABLE OF CASES

US

Frye v. United States 293 F. 1013, D.C. Circ., 1923
Daubert v. Merrell Dow Pharmaceuticals 509 U.S. 579 (1993)
General Electric Co. v. Joiner, 522 U.S. 136 (1997)
Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999)

Sweden

Piteå Tingsrätt, B 179/94
Gällivare Tingsrätt, B 26/95