Law and statistics, a perfect combination?

A study of the level of interpretation of statistical forensic evidence in criminal trials by Dutch courts

Abstract

This paper examines the impact of the publication of a report of the Dutch council of jurisdiction in 2009 on the level of interpretation of forensic evidence by Dutch courts by using an ordered probit model. This paper also investigates if there is a difference in the interpretation level between the low and high courts in the Netherlands as well as if the seriousness of the crime and the penalty given to the defendant have any impact on the latent variable. The results suggest that the level of interpretation of statistical forensic evidence by Dutch courts did not improve significantly after the publication of the report. Likewise, judges at the high courts do not interpret forensic evidence in a better way than their colleagues at the low courts and there is no evidence that both the seriousness of the crime and the penalty have a significant impact.

Author: Koen Bastiaansen

Student number: 384791

Supervisor: C.M. Oosterveen MSc

Date: 12-07-2017

Erasmus School of Economics

Erasmus University Rotterdam

Introduction

Nobody wants to be imprisoned for several years for a crime he did not commit. Unfortunately, sometimes this happens. A recent example in the Netherlands is the case of Lucia the Berk. “Lucia de Berk acquitted” was the header of the NRC on the 14th of April 2010 (NRC, 2010). Her story is one of the biggest judicial errors in the history of Dutch law. Lucia de Berk was a nurse in a hospital where in a short period of time seven children died under suspicious circumstances. The only common factor of those deaths was that they happened while she was working in the hospital. Statisticians calculated that the chance all those children died during here working hours while she had nothing to do with it was very small. She was proven guilty and sentenced to life imprisonment in 2003. Also the Higher and Supreme Court proven her guilty and did not change the imprisonment. In 2008 there was evidence that the children died as a result of bad healthcare and not because of intoxication by Lucia de Berk. She was released from prison in 2008, after an imprisonment of 6.5 years, and acquitted by the high court in 2010.

One of the most important pieces of evidence in this case was statistical forensic evidence. Lucia de Berk did not confess the murders and moreover nobody caught her doing it or found her DNA on the bodies. Hence, the court asked the Dutch Forensic Institute (NFI) to calculate the chance to observe 7 death children during her working hours given she had nothing to do with it. The NFI reported that this chance was less than 1 in 342 million, so the null hypothesis she had nothing to do with those deaths was rejected (Meester, Collins, Gill, & Lambalgen, 2006). It is said that without this statistical evidence she never was convicted in the first place.

Statistics is becoming more important in the courtroom nowadays (Raad voor de rechtspraak, 2009). Hence, judges should know where statistics is about, otherwise errors like the case of Lucia the Berk might happen more in the future. Due to new technologies more evidence can be found to convict a perpetrator. One of those technologies is matching the DNA found at the crime scene to a person. Having a DNA databank can make this process less time-consuming. However, because of the increasing significance of statistical evidence in the justice system it is very important for judges to interpret the statistical results correctly. Unfortunately, practice has shown that many judges make errors with the interpretation of statistical results. An often seen mistake is the so-called prosecutor’s fallacy or inverse fallacy. This means that instead of making a statement about the probability of the evidence

given the hypothesis judges make a statement about the probability of the hypothesis given the evidence (Sonnemans & Dijk, 2011).

For all judges it is the norm to make as few mistakes as possible, but for judges who are responsible to investigate the guilt of defendants in criminal cases this norm is even more important, because of the possible consequences. Absolute certainty does not exist in the justice system. Society does not see the difference between the situation where someone is wrongfully convicted and the case a perpetrator is not convicted because of a lack of evidence. However, there is a big difference between those two situations. It can be argued that it is worse to convict an innocent person (type I error) than to acquit a perpetrator (type II error). The first situation is definitely an error of the court, but the second situation can be justified. When there is not enough evidence to proof the guilt of the defendant, it is better to acquit the person than convict him based on a feeling or unconvincing evidence. Legal standards as beyond reasonable doubt and 'convincingly proven' provide guidance to judges by making their decision. These standards reflect a kind of trade-off between the two errors. However, these standards are unavoidably vague and can be interpreted differently by different judges (Sonnemans & Van Dijk, 2011). When decisions are made under uncertainty, it is inevitable to make mistakes. Nonetheless, the less mistakes are made the better it is.

A report of the Dutch Council of Jurisdiction (Raad voor de rechtspraak) published in the last quarter of 2009 highlights the problem of a lack of statistical knowledge of judges. As said before, statistical knowledge is becoming more important in proving the guilt of a defendant. The report names some fallacies that are commonly made by judges. The consequences of a wrong verdict can be tremendous. An experiment in the report shows that most people do not make a difference between a wrongful conviction and a wrongful acquittal. The authors end their report with a recommendation to improve the statistical knowledge of people working in the legal sector. They argue that judges should have a good understanding of the statistics in order to be able to stand above the forensic experts when it comes to the general aspects of methods of research, including statistics. They argue that because of the increasing importance of statistics in trials decisions under uncertainty should be made by lawyers who are explicitly trained for this (Raad voor de rechtspraak, 2009).

This paper will examine whether the publication of the report in the last quarter of 2009 has caused a kind of shock effect to increase the level of interpretation of forensic evidence by Dutch courts significantly or not. The Dutch Council of Jurisdiction is a very influential institute in the Netherlands for people working in the legal sector. Courts, judges and even

parliament mostly follow her advice. To examine this question the following research question is formulated:

Is there a significant improvement of the level of interpretation of statistical forensic evidence by Dutch courts after the publication of a report of the Dutch council of jurisdiction (Raad voor de rechtspraak) on statistics in criminal trials in 2009?

This will be examined on the basis of jurisprudence. In order to have similar cases only cases with a DNA-match will taken into account. The level of interpretation before 2009 and the period thereafter will be examined. Because it is unknown on which day in the last quarter the report was published, this report has examined the difference between the periods before and after 01-01-2010. This means that some kind of shock effect is measured.
In an article from Sonnemans and Van Dijk (2011) it is suggested that judges and students with an economic or science background interpret evidence in a more correct manner than candidate judges and law students. Although those law students and candidate judges know the basics of statistics, most of them lack the skills to reach correct verdicts when it comes to hard criminal cases is the conclusion of the paper. The paper found that judges are better in interpreting evidence than candidate judges, which suggests that experience matters. Interesting would be to examine if judges at the high courts have a better level of interpreting statistical forensic evidence than their colleagues at the low courts, since they are on average more experienced. To become a judge at one of the high courts in the Netherlands someone needs at least 10 years of experience in the legal sector and experience as a judge is desirable, while for the low court 5 years of experience is sufficient (Raad voor de rechtspraak, 2017). Because of the difference in requirements it is expected to see a better level of interpretation of forensic evidence at high courts compared with low courts in the Netherlands.
Furthermore, it should not matter for the penalty if the judge has some doubts about the guilt of the defendant. De Keijser and Van Knoppen (2004) introduce in their paper the conviction paradox, which means that judges give perpetrators a lower penalty when they are not fully convinced of their guilt. This conviction paradox could lead to a lower standard of proof for the judge and at the same time to a lighter penalty for the crime. The authors show that the conviction paradox does not exist. Interesting would be to examine if the seriousness of the crime and the penalty given to the defendant matter for the level of interpretation of forensic evidence. Since the authors show the conviction paradox is not true there is no reason to

believe the seriousness of the crime and the penalty have any impact on the level of interpretation of the court.

It was expected that an improvement could be seen after the publication of the report. However, this paper does not find a significant difference in the level of interpretation between the period before and after the publication of the report of the Dutch council of jurisdiction. When a dummy variable is used for the date of the verdict the results are still far from significant. Also the dummy variable has barely any impact on the interpretation level. Furthermore, contrary to what was expected, the level of interpretation of the high courts is not significantly better than of their colleagues of the low courts in the Netherlands. This result can be interpreted as very positive, since it benefits the legal security. The results of testing the impact of the seriousness of the crime are in line with the expectations based on the conviction paradox. This paper suggests that there is no evidence for a difference between the level of interpretation of serious and less serious crimes in the Netherlands. Also the penalty given to the defendant has no significant impact on the level of interpretation. Like the first sub-question these results are a very positive sign. It shows that these factors do not influence the level of interpretation of the court, which is good for the legal security in the country.

The paper is organised as follows. To examine the research question and the two sub-questions the theoretical framework will be discussed first. This section will be followed by the presentation of the data and the empirical strategy. Thereafter the results will be presented and discussed. Lastly there is a conclusion and discussion where a summary is given of the results. Furthermore possible policy implications and limitations of this research will be discussed.

Theoretical framework

In this section the statistics behind the interpretation of the statistical forensic evidence will be explained. The likelihood ratio approach and the Bayesian statistics will be discussed in more detail, since the three questions are based on these statistics. Furthermore, there are some common fallacies that are made when interpreting statistical evidence. These fallacies will be mentioned, because nowadays they still exist in the courtroom. Judges need to know these fallacies in order to prevent making those mistakes themselves. Also the conviction paradox of De Keijser and Van Knoppen (2004) will be discussed, to argue if the penalty and seriousness of the crime has any impact on the level of interpretation or not. Lastly, the legal framework in the Netherlands will be shortly explained.

A correct interpretation of evidence is essential to prevent a wrongful conviction. However, many mistakes are still made when interpreting the results of forensic researchers. To prevent mistakes when interpreting forensic evidence, statistical knowledge is essential for people working in the legal sector.

The Dutch Forensic Institute (NFI) uses the likelihood ratio (LR) as the basis of their conclusion and does not report only a p-value of the likelihood ratio to the court. The LR is used to show the probability of the results under two formulated hypotheses and measures the strength of the evidence. In practice the likelihood ratio is the same as 1/coincidental match and therefore the forensic expert only reports this to the court (Nederlands Forensisch Instituut, 2014).

There is broad scientific consensus that likelihoods are a better tool for DNA evidence evaluation compared with mentioning only p-values. In contrast to p-values, likelihood ratios measure the strength of the evidence (Steele & Balding, 2014). The likelihood ratio is the chance to get the found results if hypothesis 1 is true divided by the chance to get the results if hypothesis 2 is correct. The NFI makes two statements: one if there is a match between the compared DNA profiles and the second statement gives the rarity of the match. When there is a full match the chance a random unrelated person also matches the DNA is less than 1 out of 1 billion, but with an incomplete DNA profile the chance is much higher. When the forensic researcher wants to do a DNA test, he first formulates two hypotheses. Mostly he will draw up the following two hypotheses:

- The donor of the cell material is the suspect.
- The donor of the cell material is a random person, unrelated to the suspect.

When there is a match between the DNA of the suspect and the DNA found at the crime scene, the forensic expert would argue that, under the assumption everything went well, the found results are more likely when hypothesis 1 is true compared to the case hypothesis 2 is correct. It is incorrect to interpret his interpretation as if because of the high degree of similarity of the DNA it is more likely the defendant has done the crime instead of somebody else. The judge could use this result together with the associative evidence for his verdict. When hypothesis 1 is much more likely than hypothesis 2, this does not necessarily imply that the defendant is the perpetrator. Despite the fact that a DNA-profile is extremely rare, it is possible that two unrelated persons have the same DNA-profile (Nederlands Forensisch Instituut, 2007). The forensic researcher is limited to give a statement about the likelihood.
So, the forensic expert does not give a certain conclusion. This has several reasons. Firstly, when a DNA-match is extremely rare, this does not necessarily means the match is unique. It cannot be ruled out that someone else has the same DNA as the perpetrator. Secondly, it is possible a family member of the defendant is the perpetrator, which results in a strong match. Lastly, although the chance of mistakes is small, it is possible they take place. Therefore the forensic expert should be cautious with his conclusions (Nederlands Forensisch Instituut, 2007).

Likelihood ratio

As mentioned before, the forensic expert of the Dutch Forensic Institute reports likelihood ratios (LR) to the court. According to Perlin (2010) there are four different ways to present the likelihood ratio, namely: the hypothesis form, the likelihood form, the genotype form and the match form. The likelihood form is used in the Netherlands. The task of the forensic researcher is to determine the probability of obtaining the observed properties of the sample of known origin and the sample of questioned origin under the hypothesis that the two samples have the same origin versus under the hypothesis that the samples have different origins. In scientific literature there is a broad consensus that likelihood ratios are the primary tool for DNA evidence evaluation and that forensic experts should present the strength of the DNA evidence by making use of the likelihood ratio instead of only p-values of the likelihood ratio, since those are not suitable for measuring the strength of DNA evidence.
The following formula is used to calculate the likelihood ratio (LR):

$$\mathrm{LR} = \frac{\mathrm{P}(E|H_1)}{\mathrm{P}(E|H_2)} \qquad (1)$$

E denotes the evidence and H_1 and H_2 represent the competing hypotheses corresponding to the position of the prosecutor and the position of the defence respectively. The hypothesis of the prosecutor is typically unique, but for the hypothesis of the defence case many hypotheses should be considered. A likelihood ratio greater than 1 supports H_1, while a LR of less than 1 supports H_2 (Steele & Balding, 2014).

There are several reasons why the likelihood ratio approach is well-suited for forensic evidence interpretation in the courtroom. While the likelihood ratio approach, sometimes called the logical approach, measures the strength of the evidence and allows a finder of fact to combine the evidence with background information or other evidence in a coherent way, p-values of the likelihood ratio only tells how (un)likely it is to obtain evidence that is equally strong or stronger if the suspect is assumed to be a randomly chosen person. Probably the best feature of the likelihood ratio approach is the fact that the inferential process is out of the hands of the forensic expert, but rests with the court instead, who is not responsible for assessing the strength of the evidence (Kruijver, Meester, & Slooten, 2015). However, there has been resistance against the likelihood ratio approach in the courtroom. The main critique on the approach is that is hard to explain the meaning of a likelihood ratio. People make mistakes easily with the interpretation of likelihood ratios. There are several very common fallacies. It is likely that the application of common sense leads to errors by interpreting forensic statistics. Especially for countries with a jury trial this can result in errors with major consequences. It has been shown that jurors do not handle probabilistic evidence in a correct manner (Buckleton, 2005). However, given the extent to which p-values of the likelihood ratio are misunderstood in the sciences, it seems unlikely that p-values would be well understood in the courtroom (Goodman, 2008; Hubbard & Lindsay, 2008; Windish, Huot & Green, 2007).

Because of the fact that p-values can be highly misleading when interpreted as a measure of the strength of the evidence, Kruijver et al. (2015) argue that p-values of the likelihood ratio should not be used to explain the strength of evidence in the courtroom.

Statistics

There are two important schools of statistics, namely the classical school and the Bayesian school. The Bayesian school is the most important one for forensic statistics. The most important difference between the two schools is that Bayesian statistics uses an a priori likelihood. When a judge sees a suspect for the first time he must have an a priori likelihood that the defendant is guilty of 0.5, otherwise the judge would be biased. However, judges

know that the chance that someone before the court is guilty is bigger than the probability the person is not guilty, so an a priori likelihood of 0.5 is not necessarily the case. This a priori likelihood is very important in the final decision of the judge.

The formula for judges to calculate the chance the defendant is guilty is as follows:

$$\frac{P(g|e)}{P(ng|e)} = \frac{P(g)}{P(ng)} \times \frac{P(e|g)}{P(e|ng)} \quad (2)$$

In the formula g stands for guilty, ng for not guilty, e stands for evidence and P(g|e) + P(ng|e) = 1.

The formula says that the posterior odds = prior odds x strength of the evidence (which is the likelihood).

When the last part is >1, the evidence is incriminating otherwise it is exonerating (Sonnemans & Dijk, 2011). The article of Bonnes, Vergeer & Stoel (2015) provides an example why the initial belief of the judge (the second part of the equation) can matter. In the example based on the evidence and a burden of proof of at least 99.95%, the judge should only convict the suspect when his a priori chance of guilt is 0.9 or higher, otherwise he should acquit the person.

The likelihood ratio, given by the forensic expert, helps by determining the posterior odds. A high likelihood ratio does not necessarily imply that the posterior odds are high as well because prior odds play a role as well. Moreover, high posterior odds do not necessarily imply that the first hypothesis is likely, since the odds only compare just two hypotheses. It is possible that there may exist another hypothesis that is in fact more likely.

Fallacies

Despite the fact that equation 2 is quite clear, the literature shows that there are difficulties for people applying the concepts of Bayesian statistics. Firstly, experiments show that people have difficulties interpreting the individual probabilities of the right-hand side of the formula and with combining the prior odds and the strength of the evidence. People tend to be more conservative than Bayes' theorem when revising judgements in the light of new information. On average people tend to stick too much to their prior belief and thereby they tend to underestimate the strength of the evidence (Bornstein, 2004; Faigman & Baglioni, 1988; Thompson & Schumann, 1987). Also Kahneman and Tversky (1972) show that people have difficulties with the interpretation of probabilities. They show that people tend to use a simplifying heuristic to evaluate probabilities. Also people tend to base their judgements on

the extent to which the evidence being analysed is representative of the category. When the evidence appears representative of the category (for example the defendant is shifty and nervous), people judge the likelihood that the evidence is a product of that category as high. The experiments in the article of Guthrie, Rachlinski and Wistrich (2001) show this kind of behaviour. People undervalue the importance of the frequency with which the underlying category occurs. This frequency is known as the base-rate statistic. Even though the base-rate information is highly relevant, the experiments in the article show that people routinely ignore or discount it when making categorical judgements.

Victims of the inverse fallacy assume that P(A|B) (the conditional probability of A given B) is the same as P(B|A) where A represents the case where the defendant is innocent and B represents the case where the defendant has the matching blood type. Prosecutors made the mistake quite often, hence the name. When people determine the probability of the defendant's guilt by subtracting the incidence rate of a match between the DNA of the defendant and DNA found at the crime scene from one, they also suffer from the prosecutor's fallacy. By doing so you get a misguided assessment because it purports to determine the defendant's probability of guilt based solely on that specific evidence. The strength of all other evidence is ignored. For example, in the case that the DNA of the defendant and the DNA found at the crime scene math on a blood type found in 10% of the population, some people reason that there is a 90% chance the defendant is guilty. As mentioned before, the strength of the other evidence is not considered. So the victims of this fallacy confuse the probability of the DNA evidence given a particular proposition with the probability of the proposition given the DNA evidence (Thompson & Schumann, 1987).

In the experiment of Guthrie et al. (2001) there were as many judges victim of the inverse fallacy as judges who responded correctly to a specific case. Although, according this research, the sample of judges performed better than other decision makers with respect to framing and representativeness, still 40% of the judges gave answers consistent with the inverse fallacy, implying that they overrate the evidence. The question is if judges are less susceptible to the inverse fallacy than other decision makers and people with different backgrounds. The literature is not consistent on this point. A paper of Guthrie, Rachlinski and Wistrich (2002) says that judges are less often victim of the inverse fallacy compared with other people. However, Sonnemans en Van Dijk (2011) show in their experiment that judges make approximately the same number of mistakes as students with an economic or science background. Law students and candidate judges perform even worse. Although those law

students and candidate judges know the basics of statistics, most of them lack the skills to reach correct verdicts when it comes to hard criminal cases is the conclusion of the paper.
This paper will examine if there is a difference in the level of interpretation of forensic evidence by judges of low courts compared with judges of high courts in the Netherlands. On average judges from the high courts are more experienced. To become a judge at one of the high courts in the Netherlands someone needs at least 10 years of experience in the legal sector and experience as a judge is desirable. To become a judge at one of the low courts only 5 years of experience in the legal sector is required (werkenbijderechtspraak, 2017). Hence, it is expected they will interpret statistical evidence better.

Another commonly made misinterpretation of forensic data, first heard voice by a defence attorney, is called the defence attorney's fallacy. Victims of this fallacy disregard associative evidence, regardless of the rarity of the matching characteristics. Those victims reason that the associative evidence is irrelevant because it shows that the defendant and perpetrator belong to the same large group of people. Suppose, for example, that the defendant and the perpetrator share a blood type possessed by 0.1% of a population. In this case victims of the fallacy reason that when the population consists of 50 million people, there would be approximately 50,000 people sharing the same blood type. They will say that there are 49,999 other people who have the same chance being the perpetrator as the defendant. However, there is associative evidence that resulted in the trial of this defendant instead of the other 49,999 people. The associative evidence drastically narrows the group of people who are suspect.
Fenton and Neil (2000) present in their paper a third fallacy, the jury fallacy. This fallacy implies that observers of a trial automatically lose some confidence in a not guilty verdict when hearing about evidence of a previous similar conviction by the defendant. Although this extra information should not matter, the authors show in their article that most people are going to doubt about their first opinion of the defendant.

Conviction paradox

According to article 350 of the Dutch code of criminal procedure the judge has to determine if there is enough evidence to convict the suspect for the facts he is being accused of. When the answer on the question is positive, the judge has to decide which penalty he will give the defendant. It should not matter for the penalty if the judge had some doubt by deciding if the defendant was innocent or not because the proof is a conditio sine qua non. De Keijser and

Van Knoppen (2004) asked themselves the question if this is true in practice. The authors thought that some judges give perpetrators a lighter penalty when they are not fully convinced of their guilt. They call this the conviction paradox. This conviction paradox could lead to a lower standard of proof for the judge and at the same time to a lighter penalty for the crime. In their experiment the authors show that the conviction paradox is not true. In their experiment they presented all the judges similar cases but the evidence was different in the cases from weak to very strong. If a judge decided to convict the defendant, there was no significant difference in penalty between the cases with strong evidence and the cases where the evidence is sufficient for a conviction but less strong.

Interesting would be to examine if there are differences in interpretation of statistical evidence between cases where the defendant is under suspicion of a serious crime like murder or where he is under suspicion for a relatively small crime like theft. Perhaps judges are more careful with interpreting the results of the forensic evidence when the defendant is under suspicion of a serious crime in comparison with cases where suspect is under suspicion of a less serious crime. This paper will also take the penalty given to the defendant into account. Perhaps judges are more careful when interpreting forensic evidence when they give a high penalty to a defendant. Because Keijser and Van Knoppen (2004) show that the conviction paradox is not true, there is no reason to believe that the seriousness of the crime matters for the level of interpretation of the statistical forensic evidence by the court. This question will be examined in the paper. As serious crimes will be considered all the crimes with a maximum penalty of more than 8 years imprisonment.

Legal framework

In the Netherlands DNA investigation in criminal matters is legally allowed according to the decree DNA-research in criminal matters of 2001 (Rijksoverheid, 2014). Next to the decree the public prosecutor is able to give an order to research the DNA of a suspect according to article 151a and 151b of the Dutch criminal law.

Based on these articles DNA research is a legally allowed piece of evidence in criminal trials. However, a judge cannot convict a suspect based on only the DNA investigation. According to article 338 of the Dutch criminal law a judge is only able to blame a suspect when he or she is convinced of the guilt of the suspect beyond reasonable doubt on the basis of pieces of evidence at the investigation at the hearing. The law explicitly says pieces of evidence, which means that according to the Dutch criminal law only a DNA match is not enough to convict a suspect. Therefore associative evidence is needed.

Data

The data that is used in this thesis is retrieved from rechtsspraak.nl, the website of the Dutch administration of justice. On this website you can find jurisprudence. However, it must be noted that not all jurisprudence is published on rechtsspraak.nl. To have similar interpretations of the forensic evidence, only jurisprudence where DNA matches were interpreted are used. The search term "DNA match NFI" is used to find relevant jurisprudence. NFI is the Dutch Forensic Institute (Nederlands Forensisch Instituut) and they provide the forensic evidence for the court. Cases of the Dutch Supreme Court are not included in the database, because she only examines the formal procedures of criminal trials and she does not look another time to the evidence. Because of this only cases of the Dutch *'rechtbanken'* (low courts) and *'gerechtshoven'* (high courts) are included in the database. The codes of all the cases can be found in table 9 in the appendix.

As mentioned earlier, the Dutch Council of Jurisdiction highlighted the lack of statistical knowledge of Dutch judges in a report in the last quarter of 2009. The authors of the report argue that judges should have a good understanding of the statistics in order to be able to stand above the forensic experts when it comes to the general aspects of methods of research, including statistics. For this judges need extra training in statistics. Also they argue that because of the increasing importance of statistics in trials, only judicial experts who are also explicitly trained in statistics should make decisions under uncertainty.

They came with a recommendation to improve the quality of the statistical knowledge of judges through training. Besides that, they recommended to improve the statistical knowledge and decision-making under uncertainty of judges (Raad voor de rechtspraak, 2009).

To examine if the level of interpretation increased significantly after the publication of the report, the following data are included of every case:

A The date of the verdict

It is interesting if there is a significant difference in the level of interpretation before 2010 and the period thereafter as a result of the publication of the report. The exact publication date is unknown, but the report was published in the fourth quarter of 2009. Despite that the Dutch council of jurisdiction is an influential organisation in the legal sector in the Netherlands, it is unknown of her reports are read by all judges within a short period after publication. Therefore, as a kind of robustness check, the difference between the periods before and after

01-01-2011 will be examined as well. To examine if there is a significant difference between the periods two dummy variables will be added to the model: date1 and date2. The first dummy will take the value 0 for verdicts before 01-01-2010 and 1 otherwise and the second dummy will take the value 1 for verdicts from the period 01-01-2011 and 0 for the period before. However, it is not expected that the dummy date2 has a significant impact on the interpretation level.

B Interpretation of the evidence by the judge

Based on the results of the forensic expert the judge can only conclude if there is randomness or not. Unfortunately, judges make very different interpretations of the forensic evidence. For example, sometimes the judge says the defendant is the perpetrator because of the low frequency of the DNA-match. However, this is not the conclusion that can be drawn from the result of the forensic expert. A degree of correctness is made to compare the different interpretations of the evidence by the courts. Three degrees of correctness were made to examine the level of interpretation in a specific case.

Wrong interpretation (0): The judge misinterpreted the forensic evidence, for example by saying the defendant is the perpetrator purely based on the DNA-match as a result of the low frequency.

Sufficient interpretation of the evidence as whole but not interpreting the forensic evidence separately (1): A judge is not able to convict a defendant on the basis of only a DNA-match, because according to Dutch law at least two pieces of evidence are needed. It is often stated by the judge that on the basis of all the mentioned evidence the defendant is found guilty of the crime. However, to follow the mental process of the judge and to control that no mistakes are made it is necessary the interpretation of the forensic evidence is mentioned separately.

Sufficient interpretation of the evidence as a whole and the forensic evidence is interpreted separately (2): Same as under 1 but now the forensic evidence is interpreted separately and in a correct way.

C Fallacy

If the interpretation of the forensic evidence is wrong, the possible fallacy is mentioned.[1]

[1] By using a multinomial logit model, as used by Allen, Bray and Seaks (1997), it is shown that the fallacy variable is far from significant. Also the marginal effects of the date variable have a

D Place of the low court / high court

Since 2013 there are 11 courts and five high courts in the Netherlands. In 2013 the judicial division in the Netherlands changed. The number of low courts was reduced from 19 to 11 and the number of high courts was reduced from 5 to 4 (OM, 2017). For example, the court of Middelburg does not exist anymore since 2013 and is know part of the court Zeeland-West-Brabant. To prevent confusion only the names of the courts after the revision are used, so a case of the court of Middelburg from 2008 is labelled as a case of the court of Zeeland-West-Brabant. This paper will examine if there is a difference in interpretation of forensic evidence between the low and high courts in the Netherlands. To test if there is a difference in the level of interpretation between the low courts and the high courts in the Netherlands the variable court is a dummy variable that takes the value 0 for verdicts of the low court and 1 for judgements of the high court.

E Duration of imprisonment

The duration of imprisonment (penalty) is stated in months and the conditional penalty is not taken into account. The penalty variable will be used to test if the penalty given to the defendant has an impact on the level of interpretation by the court.

F Seriousness of the crime the defendant is accused of

Of every case the crime of which the defendant is accused of is mentioned. The stated crime is the primary accusation. Serious crimes are crimes for which the maximum penalty is at least 8 years imprisonment and are labelled with a 1. The other crimes are less serious crimes and are labelled with a 0. The Dutch criminal law of 8 July 2015 will be used to determine the maximum penalties for a specific crime. As mentioned earlier this data will be used to examine if there is a difference in the interpretation of forensic evidence between very serious crimes, such as murder, and less serious crimes like pickpocketing.

Descriptive statistics

DNA matches are becoming more common in the courtroom because of improved technologies. Therefore the database includes more jurisprudence from the period 2010 till the end of 2016 compared with the period prior to that. Of the period 2006-2009 40 cases are included and the period 2010-2016 consists of 80 cases. Table 1 shows the number of cases in each year that are included. In 2016 around 20% of the criminal cases was treated by both the

low court and the high court[2], so the number of high court cases in the sample is quite representative (Raad voor de rechtspraak, 2017).

Year/type of court	Low court	High court	Total
2006	1 (0.83%)	0	1 (0.83%)
2007	3 (2.5%)	0	3 (2.5%)
2008	11 (9.17%)	4 (3.33%)	15 (12.5%)
2009	16 (13.33%)	5 (4.17%)	21 17.5%)
2010	22 (18.33%)	6 (5%)	28 (23.3%)
2011	7 (5.83%)	4 (3.33%)	11 (9.17%)
2012	22 (18.33%)	4 (3.33%)	26 (21.67%)
2013	4 (3.33%)	1 (0.83%)	5 (4.17%)
2014	2 (1.67%)	2 (1.67%)	4 (3.33%)
2015	1 (0.83%)	2 (1.67%)	3 (2.5%)
2016	3 (2.5%)	0	3 (2.5%)
Total	92 (76.67%)	28 (23.33%)	120 (100%)

Table 1: Number of cases distributed over the years and type of court

In less than 20% of the cases the court interpreted the forensic evidence wrongly, as shown in table 2. However, in the cases the court did not interpret the forensic evidence in a wrong manner, the judge did not write down the conclusion based on the forensic conclusion sufficiently in more than half of the cases (43.33% and 40% respectively).

Year/level of interpretation	0	1	2
2006	0	1	0
2007	0	1	2
2008	3	3	9
2009	3	10	8
2010	5	12	11
2011	2	5	4
2012	4	15	7
2013	2	0	3
2014	0	2	2
2015	1	1	1
2016	0	2	1
Total	20 (16.67%)	52 (43.33%)	48 (40%)

Table 2: The degree of interpretation by the court distributed over the years

[2] In 2016 the low courts treated 176,000 criminal cases of which 34,600 cases were also treated by the high court.

In 96 of the 120 cases the defendant was suspected of a crime of which the maximum penalty was at least 8 years imprisonment. The average penalty for the defendants of serious crimes is 94.9 months imprisonment, while the average penalty of all the cases where the defendant is convicted is 79.9 months.[3]

Table 3 shows the number of cases at each court. The court of Gelderland has the most cases, but is due to the fact that this court has several locations.

Court	Number of cases
Rechtbank Amsterdam	12 (10%)
Rechtbank Den Haag	6 (5%)
Rechtbank Gelderland	17 (14.17%)
Rechtbank Limburg	5 (4.17%)
Rechtbank Midden-Nederland	12 (10%)
Rechtbank Noord-Holland	10 (8.33%)
Rechtbank Noord-Nederland	4 (3.33%)
Rechtbank Oost-Brabant	13 (10.83%)
Rechtbank Overijssel	4 (3.33%)
Rechtbank Rotterdam	3 (2.5%)
Rechtbank Zeeland-West-Brabant	6 (5%)
Gerechtshof Amsterdam	8 (6.67%)
Gerechtshof Arnhem-Leeuwarden	4 (3.33%)
Gerechtshof Den Bosch	13 (10.83%)
Gerechtshof Den Haag	3 (2.5%)

Table 3: the number of cases at each court

[3] Life imprisonment is counted as 360 months, the maximum temporary imprisonment in the Netherlands.

Empirical strategy

The goal of this paper is studying the effect of the recommendations stated in the report of the Dutch council of jurisdiction that was published in the last quarter of 2009 on the level of interpretation of statistical forensic evidence by Dutch judges. First it will be examined if there is a difference between the level of interpretation before and after the publication of the report. Secondly, the level of interpretation between the low courts and high courts in the Netherlands will be examined. It can be said that the judges at the high courts are on average more experienced than their colleagues at the low courts. Afterwards it will be examined if the seriousness of the crime matters for the level of interpretation of the court.

An ordered probit model will be used to test the central question as well as the two sub-questions. This model will be used because the data have a natural ordering and unlike count data, they do not have natural numerical values. Because there are no natural numerical values, ordinary least squares (OLS) is inappropriate. With the ordered probit model the probabilities of each outcome conditional on the independent variables are modelled using the cumulative normal distribution (Stock & Watson, 2015). An undesirable consequence of applying a linear regression model is that it implicitly assumes that the difference between a wrong interpretation and a sufficient interpretation without interpreting the forensic evidence separately is the same as the difference between the last one and a sufficient interpretation where the forensic evidence is interpreted separately. There is no reason for expecting these differences to be the same. This assumption is not imposed in the ordered probit model (Daykin & Moffatt, 2002).

The ordered probit model is based on the assumption that the latent variable y_i^* depends linearly on x_i according to:

$$y_i^* = x_i\beta + \varepsilon_i \qquad (i=0,1,2) \qquad (3)$$

where y_i^* is the level of interpretation of the court, x_i the vector of independent variables, ß is a vector of regression coefficients that need to be estimated and ε_i is the error term, which is assumed to be normally distributed with zero mean and unit variance.

The better the level of interpretation by the court, the higher y_i will be.

$y=0$ if $-\infty < y^* \leq \kappa_1$

$y=1$ if $\kappa_1 < y^* \leq \kappa_2$

$y=2$ if $\kappa_2 < y^* < \infty$

The parameters k_j are known as "cut-points", or sometimes called "threshold parameters". The threshold values are unknown parameters that need to be estimated. The parameters of the model will be estimated by the method of maximum likelihood estimator (MLE). It is worth nothing that the numbers 0, 1 and 2, which represent the level of interpretation by the court are totally arbitrary, they are just labels for different categories.

But how would the probabilities of the various outcomes change when the value of one of the variables influencing the outcomes changes? Let F be the cumulative distribution function of ε_i, then

$$\Pr(y=i) = \Pr(k_{i-1} < y_i \leq k_i) = \Pr(y_i \leq k_i) - \Pr(y_i \leq k_{i-1})$$
$$= F(k_i - x_i\beta) - F(k_{i-1} - x_i\beta) \quad (4)$$

The marginal effects of the changes in the explanatory variables are given by

$$\frac{\partial P(y=i)}{\partial xi} = (f(k_{i-1} - x_i\beta) - f(k_i - x_i\beta))\beta \quad (5)$$

where f is the density function of ε_i. An increase in $x_i\beta$ will lead to larger values of the index y*, resulting in a larger outcome of y. The probability of the outcome y=0 will decrease, the probability of the outcome y=i will increase and that of y=j for j=2,…,i can increase or decrease, as at the same time $P(y \leq j-1) = F(k_{i-1} - \beta * DATE) - F(k_{i-1} - \beta * COURT)$ decreases and $P(y \geq j+1) = 1 - (F(k_{i-1} - \beta * DATE) - F(k_{i-1} - \beta * COURT))$ increases (Heij et al., 2004).

This paper will use an ordered probit model with two regressors: date and court,. Accordingly, the ordered probit model with the two regressors is:

$$\Pr(y = i \mid date, court) = \phi(ß_0 + ß_1*DATE + ß_2*COURT) \quad (6)$$

where y is the level of interpretation by the court, ϕ is the cumulative standard normal distribution function, $ß_0$ is a constant and the other $ß_n$ are the vectors of the regression coefficients. The variable date refers to the date of the verdict. The variable court is a dummy variable that takes the value 0 if a low court has made the verdict and the value 1 when it has been made by a high court.

If the variable date is positive and significant, it means that in more recent verdicts the forensic evidence is interpreted in a more correct manner. The only reason for this is probably

the impact caused by the publication of the report of the Dutch council of jurisdiction in 2009. It is possible that the level of interpretation differs greatly between the low and high courts in the Netherlands. On average judges of the high courts are more experienced than their colleagues of the low courts. When there is a positive and significant difference, high courts on average interpreted the forensic evidence in a better way. Likewise the model tests if the gravity of the possible punishment and the actually given penalty matter for the level of interpretation of the court.

To measure the fit of the model the pseudo-R_2 will be used. The pseudo-R_2 measures the fit of the model using the likelihood function. It compares the value of the likelihood of the estimated model to the value of the likelihood when none of x's are included as regressors (Stock & Watson, 2015).

In the last quarter of 2009 the report of the Dutch council of jurisdiction was published. To test whether there is a difference in the level of interpretation between the period until 01-01-2010 and the period thereafter will be examined. Therefore the dummy variable date1 will be included in the model. The variable will take the value of 0 for the period until 01-01-2010 and the value 1 of the period thereafter. However, although it is expected that the report has an immediate impact on the interpretation level of Dutch courts, also the dummy variable date2 will be added to the model. The dummy variable takes the value of 0 for the period until 01-01-2011 and 1 for the period thereafter. The variable will be added to the model to control for the possibility that the report has no immediate impact but that this takes at least several months. But to emphasize again, it is expected that the report has a quick impact on the statistical knowledge of judges, so that the dummy variable date2 has no influence on the level of interpretation of the court.

$$\Pr(y = i \mid \text{date1}, \text{date2}, \text{court}) = \phi(ß_0 + ß_1*\text{DATE1} + ß_2*\text{DATE2} + ß_3*\text{COURT}) \quad (7)$$

To test the last sub-question the following model will be used:

$$\Pr(y = i \mid \text{date1}, \text{court}, \text{seriousness}, \text{penalty}) = \phi(ß_0 + ß_1*\text{DATE1} + ß_2*\text{COURT} + ß_3*\text{SERIOUSNESS} + ß_4*\text{PENALTY}) \quad (8)$$

where the dummy variable seriousness takes the value 1 if the primary accusation is a crime for which the maximum penalty is at least 8 years imprisonment according to Dutch law and 0 otherwise. The dummy variable seriousness is added in response to the conviction paradox introduced by De Keijser and Van Knoppen (2004). It will examine if there is a significant causal relation between the seriousness of the crime of which the defendant is suspected of and the level of interpretation of the court. Also the variable penalty is added to the ordered probit model, to see if a higher penalty results in a better level of interpretation of the forensic evidence. This is an additional test to examine if both the maximum possible penalty and the actually given penalty have an impact on the interpretation level of Dutch courts, only one of them have or neither of the two variables.

The difference between the variables penalty and seriousness is that the first variable is more related to the result of among others the interpretation of the statistical evidence and that the seriousness variable is more related to one of the possible causes of the level of interpretation. This mean that the penalty variable is more related to the prior odds of formula 2 while the seriousness variable has more to do with the posterior odds of the same formula.

Results

The first question is to examine is if there is a difference in the level of interpretation of statistical evidence by Dutch courts between the period before and after the introduction of the report of the Dutch council of jurisdiction. The report was published in the last quarter of 2009, so the period before and after 01-01-2010 is examined. It is expected that the level of interpretation has been increased in the second period compared with the first stage.
The results of the first model are summarized in table 11, which can be found in the appendix, and the marginal effect of the variable date is summarized in table 4. As shown the variable date has almost no effect on the interpretation level by Dutch courts. A later verdict has even a slightly negative effect on a good level of interpretation. This would suggest that the publication of the report did not improve the statistical knowledge of Dutch judges, but maybe even the opposite has occurred. However, this result is close to 0 and far from significant, with a significance level of 0.511. Therefore, this result suggests that the date of the verdict has no influence on the level of interpretation of the statistical evidence in a criminal trial.

y	Average marginal effect	Significance
0	0.000023	0.511
1	0.0000126	0.515
2	-0.0000356	0.509

Table 4: Average marginal effect DATE of ordered probit model table 11

The court variable has a more positive impact on the latent variable, meaning that judges of the high courts are on average better in interpreting the forensic evidence in a criminal trial. However, it must be noted that the dummy variable court is not statistical significant, with a significance level of 0.404. Through this it is difficult to interpret the results of table 11. Next to the insignificance of the variables also pseudo R^2 is very low. The variables explain just 0.5% of the interpretation level of the courts.

In table 12 the dummy variable date1 will be used. As an extra test another dummy variable is added to the model, namely date2. As mentioned earlier the dummy variable date1 compares the periods before and after 01-01-2010, while date2 examines if there is a difference in the

level of interpretation between the period before and after 01-01-2011. It is assumed that the publication of the report had an immediate impact on the statistical knowledge of Dutch judges and this should have increased the level of interpretation. However, it may be possible that some judges did not read the report until 2010. Those judges did not have improved statistical knowledge at the beginning of 2010. Therefore the variable date2 was added to the model. Hence, it is expected that the dummy variable date2 has no significant impact on the latent variable. Table 5 shows indeed that date2 is insignificant, with a significance level of nearly 1. Therefore the dummy date1 will be used for further analysis. The variable date1 suggests that the interpretation of forensic evidence in verdicts after 01-01-2010 is on average less good than in verdicts in the period before. However, it must be said that also the dummy variable date1 is far from significant, with a significance level of 0.701. Also the court variable is still not significant. By using the dummy variables date1 and date2 the pseudo R^2 of the new model has decreased to 0.0035.

The marginal effect of the dummy variable date1, as shown in table 5, suggests indeed that the level of interpretation of verdicts since 2010 is on average less good compared with the period before. While the effect of a later period is still positive for y=1, this changes to a negative effect for y=2. However, as said before, with significance levels of around 0.700 the effects are far from significant.

As mentioned earlier, this is a surprising result. It was expected that the level of interpretation would increase after publication of the report. The report should have confronted judges with commonly made mistakes in interpreting statistical forensic evidence, but the results of the first two ordered probit models suggest that the opposite has occurred. However, the marginal effects of the dummy variable date1 are, like the variables in both models, insignificant. Therefore, we cannot draw conclusions upon these results.

The results of the two models suggest that there is no significant improvement of the level of interpretation of Dutch judges since the publication of the report on statistics in criminal trials by the Dutch council of jurisdiction in the last quarter of 2009. It is possible that the statistical knowledge of judges in the Netherlands was already of a very good level before the publication of the report. Although the average knowledge of statistics is possibly of a good level, too many mistakes are found in the data. Table 1 showed that the interpretation of the forensic evidence was insufficient in almost 17% of the verdicts. Despite the fact that not all insufficient interpretations necessarily have to lead to wrong convictions or acquittals, this number is still quite high in my opinion.

y	Average marginal effect	Significance
0	0.0259839	0.697
1	0.0151093	0.714
2	-0.0410932	0.702

Table 5: Average marginal effect DATE1 of ordered probit model table 12

In both models the dummy variable court is positive with an estimated value of around 0.200. This value indicates that on average judges at the high courts in the Netherlands are better in interpreting the statistical forensic evidence in a criminal trial compared with their colleagues at the low courts. The marginal effects, showed in table 6, also confirm this. The chance that the level of interpretation (y) is 2 is almost 8% higher when it is a verdict of the high court compared with the situation the trial took place at a low court. Again, it must be said that in both models the variable court is not significant. The significance levels of the variable in the models are comparable, with values both around 0.400. These results suggest that there is no significant difference in the level of interpretation of statistical forensic evidence between the low and high courts in the Netherlands. This is not the result that was expected. As mentioned earlier, to become a judge at one of the high courts in the Netherlands someone needs at least 10 years of experience in the legal sector and experience as a judge is desirable. To become a judge at one of the low courts only 5 years of experience in the legal sector is required (Raad voor de rechtspraak, 2017). Because of the difference in requirements it was expected to see a better level of interpretation of forensic evidence at high courts compared with low courts in the Netherlands. However, it can be seen as a very positive sign that there is no significant difference in the level of interpretation between the courts. The knowledge of statistical forensic evidence of the judges at both courts is therefore probably comparable, which benefits the legal certainty in the Netherlands.

y	Average marginal effect	Significance
0	-0.0513388	0.405
1	-0.0280678	0.411
2	0.0794065	0.401

Table 6: Average marginal effect variable COURT of ordered probit model table 12

To test if the seriousness of the crime and the given penalty have an impact on the level of interpretation by the court the dummy variable seriousness and the variable penalty are added to the model. The seriousness variable takes the value 1 for crimes for which the maximum penalty is at least 8 years imprisonment according to the Dutch criminal law of July the 1st 2015 and 0 otherwise. The seriousness variable is positive with an estimated value of 0.090. When a defendant is suspected of a serious crime the forensic evidence is on average better interpreted by the court compared with less serious crimes. Also the marginal effects of table 7 imply that the more serious crimes are on average better interpreted, because the chance the statistical forensic evidence is interpreted in a good manner (y=2) is 3.5% higher when the defendant was suspected of a more serious crime. However, like the other variables also the dummy variable seriousness is insignificant with a significance level of no less than 0.737.

The penalty variable has a slightly negative effect on the level of interpretation, but the coefficient is close to 0. The marginal effects of the penalty variable are close to 0 as well and they are all insignificant. This means that the penalty given to the defendant has almost no impact on the level of interpretation of the statistical forensic evidence by Dutch courts.

The pseudo R^2 of the ordered probit model of table 13 has slightly increased compared with the first model. However, because the dummy variable is insignificant and the increase of the pseudo R^2 is that small, it cannot be concluded that the seriousness of the crime of which the defendant is suspected has any impact on the level of interpretation by Dutch courts. The result of the seriousness variable is comparable with the results of the paper of De Keijser and Van Knoppen (2004) and their conviction paradox. The authors showed in their paper that there is no evidence that judges lower the penalty they give to a defendant when they are not fully convinced of his guilt. This paper shows that there is no evidence for the hypothesis that judges are more careful with the interpretation of forensic evidence of more serious crimes. Also there is no evidence that the penalty given to the defendant has any impact on the level of interpretation. It is good to observe that there is no evidence for a difference in interpretation, because that would suggest that judges make a distinction in the seriousness of the crime when interpreting statistical evidence in a criminal trial or a more careful when the penalty they want to give is higher. It is essential for the Rule of law that judges are not prejudiced, because this could decrease the chance of a fair trial.

y	Average marginal effect	Significance
0	-0.0223162	0.737
1	-0.0122107	0.738
2	0.0345269	0.737

Table 7: Average marginal effect variable SERIOUSNESS of ordered probit model table 13

y	Average marginal effect	Significance
0	0.0001584	0.633
1	0.0000867	0.635
2	-0.0002451	0.632

Table 8: Average marginal effect variable PENALTY of ordered probit model table 13

Conclusions

The aim of this paper has been to examine if there was a significant difference between the level of interpretation of statistical forensic evidence by Dutch courts before and after the publication of a report of the Dutch council of jurisdiction on statistics in criminal trials in the last quarter of 2009. The Dutch council of jurisdiction is a very influential organisation in the Netherlands, especially in the legal sector, and she worried about the statistical knowledge of Dutch courts. Hence, it was expected that an improvement could be seen after the publication of the report. However, this paper shows by using an ordered probit model with 120 verdicts from the period 2006-2016 that there is no statistical significant evidence to conclude that a difference between the periods exits. When the dummy variable date1 is used for date, the variable was still insignificant.

Also it was examined if there was a difference in the level of interpretation between the low and high courts and if the seriousness of the crime of which the defendant was suspected had an impact on the level of interpretation. This paper could not find a significant difference for the first question. This probably means that the statistical knowledge of judges at the low courts and high courts in the Netherlands is comparable. This was not expected, since on average judges at the high courts are more experienced than their colleagues at the low courts. This result can be interpreted as very positive, because it benefits the legal security. The results for the second question were insignificant too. The results suggest that the seriousness of the crime and the penalty given to the defendant have no impact on the level of interpretation. This is comparable with the conclusions of the paper of De Keijser and Van Knoppen (2004) and their conviction paradox. Also this result is positive, since it is essential for the Rule of law that judges are not prejudiced, because this could decrease the chance of a fair trial. Lastly, the height of the penalty given to the defendant has no significant impact on the level of interpretation.

The results do not suggest that there is a difference in the level of interpretation before and after the publication of the report in the last quarter of 2009. This could mean that the statistical knowledge of Dutch judges is of a very good level. However, table 2 shows that in 16.67% of the examined verdicts the forensic evidence was interpreted in a wrong manner. This is still a large number. Judicial errors can have huge consequences and should be prevented as much as possible. Therefore this paper recommends further improving the statistical knowledge of judges in the Netherlands, for example by providing training. I

further support the recommendation of Sonnemans en Van Dijk (2011) that courts should use judges who are specifically trained in making decisions under uncertainty and have knowledge about the statistical methods in trials where statistical forensic evidence is important.

This research suffers several shortcomings. Firstly, only 120 randomly chosen verdicts are examined, which is a small sample. Furthermore only cases where DNA evidence played an important role in the case were taken into account. For further research it would be recommended to take a larger sample of cases and to examine other statistical forensic evidence besides DNA evidence as well. Thirdly, not all the jurisprudence is published on rechtspraak.nl. This could have influenced the results. Next to these shortcomings this paper used only four variables for the models. Further research could examine the impact of more variables on the level of interpretation of the court. It could be interesting to include several characteristic variables in the models. However, this could be difficult since most characteristics of the defendant are not described in the verdicts because of privacy regulations. The major shortcoming of this paper could be the incompleteness of the verdicts. Judges could interpret the statistical forensic evidence in a correct manner but it possible that they do not write this down correctly. It could be useful to emphasize the importance of a correct formulation of the interpretation of the evidence in order to be able to check the court.

Forensic evidence is becoming more important nowadays in court cases because of improved technologies. Further research should be done to the interpretation level of courts in the coming years when hopefully more advanced technologies are available. Besides this looking to the level of interpretation of statistical forensic evidence in other Western countries could be very useful for Dutch courts. To point out the differences and similarities in the manner of interpretation between different countries could help to improve the statistical knowledge of judges in the Netherlands.

Bibliography

Allen, S.D., Bray, J. & Seaks, T.G. (1997). A Multinominal Logit Analysis of the Influence of Policy Variables and Board Experience on FOMC Voting Behavior. *Public Choice, 92* (1), 27-39.

Bonnes, J., Vergeer, P. & Stoel, R. (2015). Rationele besliskunde in het strafrecht. *Nederlands Jursistenblad* , 364-371.

Bornstein, B. (2004). The Impact of Different Types of Expert Scientific Testimony on Mock Jurors' Liability Verdicts. *Psychology, Crime and Law , 10* (4), 429-446.

Buckleton, J. (2005). A Framework for Interpreting Evidence. In J. Buckleton, C. Triggs, & S. Walsh, *Forensic DNA evidence interpretation* (pp. 37-73). Boca Raton: CRC Press.

Daykin, A. & Moffatt, P. (2002). Analyzing ordered responses: A review of the ordered probit model. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences , 1* (3), 157-166.

Faigman, D. & Baglioni, A. (1988). Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence. *Law and Human Behavior , 12* (1), 1-17.

Fenton, N. & Neil, M. (2000). Jury Observation Fallacy and the Use of Bayesian Networks to Present Probabilistic Legal Arguments. *Mathematics Today Bulletin for the IMA , 36*, 180-187.

Garoupa, N. & Rizzolli, M. (2012). Wrongful Convictions Do Lower Deterrence. *Journal of Institutional and Theoretical Economics , 168* (2), 224-231.

Goodman, S. (2008). A dirty dozen: twelve P-value misconceptions. *Seminars in hematology, 45* (3), 135-140.

Guthrie, C., Rachlinski, J. & Wistrich, A. (2001). Inside the Judical Mind. *Cornell Law Faculty Publications , 86* (5), 777-830.

Guthrie, C., Rachlinsky, J. & Wistrich, A. (2002). Judging by Heuristic: Cognitive Illusions in Judicial Decision Making. *Judicature , 86* (1), 44-50.

Heij, C., Boer, P. de, Hans Franses, P., Kloek, T. & Dijk, H. van (2004). *Econometric Methods with Applications in Business and Economics.* New York: Oxford University Press.

Hubbard, R. & Lindsay, R.M. (2008). Why P-values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18* (1), 69-88.

Kahneman, D. & Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. *Cognitive psychology , 3* (3), 430-454.

Keijser, J. de & Koppen, P. van (2004). Compensatoir straffen: Over de relatie tussen bewijs, overtuiging en straf. In J. d. Keijser, & H. Elffers, *Het maatschappelijk oordeel van de strafrechter. De wisselwerking tussen rechter en samenleving* (pp. 133-183). Den Haag: Boom Juridische Uitgevers.

Kruijver, M., Meester, R. & Slooten, K. (2015). p-Values should not be used for evaluating t he strength of DNA evidence. *Forensic Science International: Genetics* , 226-231.
Lando, H. (2006). Does Wrongful Conviction Lower Deterrence? *The Journal of Legal Studies , 35* (2), 327-337.

Meester, R., Collins, M., Gill, R. & Lambalgen, M. van (2006). On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Law, Probability and Risk* , 233-250.

Mungan, M. (2011). A Utilitarian Justification for Heightened Standards of Proof in Criminal Trials. *Journal of Institutional and Theoretical Economics , 167* (2), 352-370.

Nederlands Forensisch Instituut. (2007). *Interpretatie van DNA-bewijs I match en berekende frequentie.* Den Haag: Nederlands Forensisch Instituut.

Nederlands Forensisch Instituut. (2014, October). *De reeks waarschijnlijkheidstermen van het NFI en het Bayesiaanse model voor interpretatie van bewijs.* Opgeroepen op April 22, 2017, van https://www.forensischinstituut.nl/binaries/Vakbijlage%20waarschijnlijkheidstermen_tcm35-56319.pdf

NRC. (2010, April 14). *NRC.nl*. Opgeroepen op April 10, 2017, van https://www.nrc.nl/nieuws/2010/04/14/lucia-de-b-vrijgesproken-door-hof-11876689-a981692

OM. (2017, May 10). *Herziening gerechtelijke kaart*. Opgeroepen op May 10, 2017, van Website public prosecutor: https://www.om.nl/organisatie/herziening/

Perlin, M. (2010). *Explaining the Likelihood Ratio in DNA Mixture Interpretation.* Pittsburgh: Cybergenetics.

Posner, R. (1999). An Economic Approach to the Law of Evidence. *Stanford Law Review , 51* (6), 1477-1546.

Raad voor de rechtspraak. (2009). *Kansrekening en strafrechtspraak: fouten bij beslissen onder onzekerheid.* Den Haag: Sdu uitgevers BV.

Raad voor de rechtspraak (2017a). *Jaarverslag 2016.* Opgeroepen op May 30, 2017, van http://www.jaarverslagrechtspraak.nl/files/rechtspraak2016/Jaarverslag_2016_web.pdf

Raad voor de rechtspraak (2017b). *Rechter of Raadsheer worden?*. Opgeroepen op June 16, 2017, van werkenbijderechtspraak.nl: https://www.werkenbijderechtspraak.nl/rechter-of-raadsheer-worden/

Rijksoverheid. (2014, November 1). *Besluit DNA-onderzoek in strafzaken*. Opgeroepen op May 02, 2017, van http://wetten.overheid.nl/BWBR0012791/2014-11-01

Sonnemans, J. & Dijk, F. van (2011). Errors in Judicial Decisions: Experimental Results. *The Journal of Law, Economics, & Organization , 28* (4), 687-716.

Steele, C. & Balding, D. (2014). Statistical Evaluation of Forensic DNA Profile Evidence. *Annual Review of Statistics and Its Application* , 361-384.

Stock, J. & Watson, M. (2015). Regression with a Binary Dependent Variable. In *Introduction to Econometrics* (pp. 437-444). Essex: Pearson Education Limited.

Thompson, W. & Schumann, E. (1987). Interpretation of Statistical Evidence in Criminal Trials. *Law and Human Behavior , 11* (3), 167-187.

Windish, D.M., Huot, S.J. & Green, M.L. (2007). Medecine residents' understanding of the biostatistics and results in the medical literature. *Jama, 298* (9), 1010-1022.

Appendix

Case	Date verdict	Place of low court/high court
ECLI:NL:RBDOR:2006:AZ0766	24-10-06	Rechtbank Rotterdam
ECLI:NL:RBBRE:2007:BB3032	06-09-07	Rechtbank Zeeland-West-Brabant
ECLI:NL:RBSGR:2007:BB9089	29-11-07	Rechtbank Den Haag
ECLI:NL:RBSHE:2007:BC1123	24-12-07	Rechtbank Den Haag
ECLI:NL:RBZLY:2008:BC3544	31-01-08	Rechtbank Overijssel
ECLI:NL:GHSHE:2008:BC5105	26-02-08	Gerechtshof Den Bosch
ECLI:NL:GHLEE:2008:BC9327	11-04-08	Gerechtshof Arnhem-Leeuwarden
ECLI:NL:RBHAA:2008:BD4753	12-06-08	Rechtbank Noord-Holland
ECLI:NL:RBSGR:2008:BD7186	15-07-08	Rechtbank Den Haag
ECLI:NL:GHAMS:2008:BF0883	22-08-08	Gerechtshof Amsterdam
ECLI:NL:GHSHE:2008:BF8472	27-08-08	Gerechtshof Den Bosch
ECLI:NL:RBBRE:2008:BC5372	29-08-08	Rechtbank Zeeland-West-Brabant
ECLI:NL:RBAMS:2008:BF8865	14-10-08	Rechtbank Amsterdam
ECLI:NL:RBROE:2008:BG1077	20-10-08	Rechtbank Limburg
ECLI:NL:RBUTR:2008:BG4634	19-11-08	Rechtbank Midden-Nederland
ECLI:NL:RBHAA:2008:BG5400	26-11-08	Rechtbank Noord-Holland
ECLI:NL:RBGRO:2008:BG5720	01-12-08	Rechtbank Noord-Nederland
ECLI:NL:RBUTR:2008:BG6736	05-12-08	Rechtbank Midden-Nederland
ECLI:NL:RBSHE:2008:BG6203	09-12-08	Rechtbank Oost-Brabant
ECLI:NL:RBZUT:2009:BH1011	27-01-09	Rechtbank Gelderland
ECLI:NL:RBSHE:2009:BH0887	27-01-09	Rechtbank Oost-Brabant
ECLI:NL:RBAMS:2009:BH1786	04-02-09	Rechtbank Amsterdam
ECLI:NL:RBDOR:2009:BH2082	05-02-09	Rechtbank Rotterdam
ECLI:NL:RBHAA:2009:BH3054	16-02-09	Rechtbank Noord-Holland
ECLI:NL:RBUTR:2009:BH3279	17-02-09	Rechtbank Midden-Nederland
ECLI:NL:RBSHE:2009:BI0918	15-04-09	Rechtbank Oost-Brabant
ECLI:NL:RBALK:2009:BI2163	23-04-09	Rechtbank Noord-Holland
ECLI:NL:RBAMS:2009:BI3717	13-05-09	Rechtbank Amsterdam
ECLI:NL:GHSHE:2009:BJ0938	08-06-09	Gerechtshof Den Bosch
ECLI:NL:RBGRO:2009:BI8719	18-06-09	Rechtbank Noord-Nederland
ECLI:NL:RBZUT:2009:BJ1981	08-07-09	Rechtbank Gelderland
ECLI:NL:RBUTR:2009:BK7518	15-07-09	Rechtbank Midden-Nederland
ECLI:NL:GHSHE:2009:BP7605	09-09-09	Gerechtshof Den Bosch
ECLI:NL:GHSHE:2009:BJ7934	17-09-09	Gerechtshof Den Bosch
ECLI:NL:GHSHE:2009:BJ7936	17-09-09	Gerechtshof Den Bosch
ECLI:NL:RBZUT:2009:BJ9770	09-10-09	Rechtbank Gelderland
ECLI:NL:RBSHE:2009:BK0716	21-10-09	Rechtbank Oost-Brabant
ECLI:NL:GHAMS:2009:BX5727	08-12-09	Gerechtshof Amsterdam
ECLI:NL:RBARN:2009:BK6465	14-12-09	Rechtbank Gelderland
ECLI:NL:RBMID:2009:BK6995	19-12-09	Rechtbank Zeeland-West-Brabant
ECLI:NL:RBZUT:2010:BK9749	19-01-10	Rechtbank Gelderland
ECLI:NL:RBROT:2010:BL0531	25-01-10	Rechtbank Rotterdam

ECLI:NL:RBUTR:2010:BL0892	27-01-10	Rechtbank Midden-Nederland
ECLI:NL:RBUTR:2010:BL0887	27-01-10	Rechtbank Midden-Nederland
ECLI:NL:RBSGR:2010:BL1698	02-02-10	Rechtbank Den Haag
ECLI:NL:RBSHE:2010:BL2130	05-02-10	Rechtbank Oost-Brabant
ECLI:NL:RBZUT:2010:BL5483	24-02-10	Rechtbank Gelderland
ECLI:NL:GHSHE:2010:BL8618	22-03-10	Gerechtshof Den Bosch
ECLI:NL:RBARN:2010:BM0871	12-04-10	Rechtbank Gelderland
ECLI:NL:RBZUT:2010:BM1196	15-04-10	Rechtbank Gelderland
ECLI:NL:GHAMS:2010:BM2026	22-04-10	Gerechtshof Amsterdam
ECLI:NL:RBHAA:2010:BM9869	29-04-10	Rechtbank Noord-Holland
ECLI:NL:RBUTR:2010:BM8182	03-06-10	Rechtbank Midden-Nederland
ECLI:NL:GHARN:2010:BN0027	30-06-10	Gerechtshof Arnhem-Leeuwarden
ECLI:NL:RBBRE:2010:BN3602	27-07-10	Rechtbank Zeeland-West-Brabant
ECLI:NL:RBGRO:2010:BN5491	26-08-10	Rechtbank Noord-Nederland
ECLI:NL:RBAMS:2010:BN5929	03-09-10	Rechtbank Amsterdam
ECLI:NL:RBSHE:2010:BN8104	24-09-10	Rechtbank Oost-Brabant
ECLI:NL:RBSHE:2010:BN9570	06-10-10	Rechtbank Oost-Brabant
ECLI:NL:RBSHE:2010:BN9482	06-10-10	Rechtbank Oost-Brabant
ECLI:NL:RBZUT:2010:BO1068	20-10-10	Rechtbank Gelderland
ECLI:NL:RBARN:2010:BO1350	22-10-10	Rechtbank Gelderland
ECLI:NL:GHSGR:2010:BO1756	27-10-10	Gerechtshof Den Haag
ECLI:NL:GHSHE:2010:BO4267	17-11-10	Gerechtshof Den Bosch
ECLI:NL:RBMAA:2010:BO4733	23-11-10	Rechtbank Limburg
ECLI:NL:RBAMS:2010:BO8429	17-12-10	Rechtbank Amsterdam
ECLI:NL:RBZLY:2010:BP2042	17-12-10	Rechtbank Overijssel
ECLI:NL:GHSHE:2010:BO7930	20-12-10	Gerechtshof Den Bosch
ECLI:NL:RBZUT:2011:BP8108	18-03-11	Rechtbank Gelderland
ECLI:NL:GHSGR:2011:BP8649	23-03-11	Gerechtshof Den Haag
ECLI:NL:RBALK:2011:BQ0906	05-04-11	Rechtbank Noord-Holland
ECLI:NL:GHSHE:2011:BR2732	22-07-11	Gerechtshof Den Bosch
ECLI:NL:RBUTR:2011:BT1898	19-09-11	Rechtbank Midden-Nederland
ECLI:NL:RBAMS:2011:BT6540	29-09-11	Rechtbank Amsterdam
ECLI:NL:RBHAA:2011:BT8676	19-10-11	Rechtbank Noord-Holland
ECLI:NL:GHSGR:2011:BU1302	25-10-11	Gerechtshof Den Haag
ECLI:NL:RBAMS:2011:BU5090	17-11-11	Rechtbank Amsterdam
ECLI:NL:RBAMS:2011:BU6199	28-11-11	Rechtbank Amsterdam
ECLI:NL:GHAMS:2011:BU8995	21-12-11	Gerechtshof Amsterdam
ECLI:NL:RBZLY:2012:BV7599	03-01-12	Rechtbank Overijssel
ECLI:NL:RBUTR:2012:BV1662	17-01-12	Rechtbank Midden-Nederland
ECLI:NL:RBSHE:2012:BV1247	19-01-12	Rechtbank Oost-Brabant
ECLI:NL:RBSGR:2012:BV1573	23-01-12	Rechtbank Den Haag
ECLI:NL:RBAMS:2012:BV2312	27-01-12	Rechtbank Amsterdam
ECLI:NL:RBZLY:2012:BV6625	09-02-12	Rechtbank Overijssel
ECLI:NL:RBSHE:2012:BV6234	21-02-12	Rechtbank Oost-Brabant
ECLI:NL:RBSGR:2012:BV6655	22-02-12	Rechtbank Den Haag
ECLI:NL:RBSHE:2012:BW0369	30-03-12	Rechtbank Oost-Brabant

ECLI:NL:RBBRE:2012:BW2076	12-04-12	Rechtbank Zeeland-West-Brabant
ECLI:NL:RBUTR:2012:BW4606	01-05-12	Rechtbank Midden-Nederland
ECLI:NL:RBALK:2012:BW5810	15-05-12	Rechtbank Noord-Holland
ECLI:NL:GHLEE:2012:BW5989	16-05-12	Gerechtshof Arnhem-Leeuwarden
ECLI:NL:RBALK:2012:BW7670	06-06-12	Rechtbank Noord-Holland
ECLI:NL:RBMAA:2012:BW8507	14-06-12	Rechtbank Limburg
ECLI:NL:RBUTR:2012:BX5072	20-08-12	Rechtbank Midden-Nederland
ECLI:NL:RBUTR:2012:BX5060	20-08-12	Rechtbank Midden-Nederland
ECLI:NL:GHAMS:2012:BX5854	27-08-12	Gerechtshof Amsterdam
ECLI:NL:GHAMS:2012:BX5850	27-08-12	Gerechtshof Amsterdam
ECLI:NL:RBAMS:2012:BX7583	29-08-12	Rechtbank Amsterdam
ECLI:NL:RBSHE:2012:BX8769	01-10-12	Rechtbank Oost-Brabant
ECLI:NL:RBZUT:2012:BY2675	06-11-12	Rechtbank Gelderland
ECLI:NL:RBSHE:2012:BY5361	07-12-12	Rechtbank Oost-Brabant
ECLI:NL:RBAMS:2012:BY6272	11-12-12	Rechtbank Amsterdam
ECLI:NL:RBAMS:2012:BY6308	13-12-12	Rechtbank Amsterdam
ECLI:NL:GHSHE:2012:BY6981	20-12-12	Gerechtshof Den Bosch
ECLI:NL:RBONE:2013:CA1920	22-03-13	Rechtbank Gelderland
ECLI:NL:RBGEL:2013:BZ9309	03-05-13	Rechtbank Gelderland
ECLI:NL:RBGEL:2013:CA1264	29-05-13	Rechtbank Gelderland
ECLI:NL:GHSHE:2013:4255	17-09-13	Gerechtshof Den Bosch
ECLI:NL:RBZWB:2013:8071	11-11-13	Rechtbank Zeeland-West-Brabant
ECLI:NL:RBGEL:2014:375	23-01-14	Rechtbank Gelderland
ECLI:NL:GHARL:2014:1723	06-03-14	Gerechtshof Arnhem-Leeuwarden
ECLI:NL:GHAMS:2014:3605	01-05-14	Gerechtshof Amsterdam
ECLI:NL:RBNHO:2014:7173	17-07-14	Rechtbank Noord-Holland
ECLI:NL:RBLIM:2015:1588	25-02-15	Rechtbank Limburg
ECLI:NL:GHAMS:2015:622	26-02-15	Gerechtshof Amsterdam
ECLI:NL:GHSHE:2015:2136	03-06-15	Gerechtshof Den Bosch
ECLI:NL:RBNHO:2016:232	18-01-16	Rechtbank Noord-Nederland
ECLI:NL:RBGEL:2016:5353	10-10-16	Rechtbank Gelderland
ECLI:NL:RBLIM:2016:10468	02-12-16	Rechtbank Limburg

Table 9: The sample of verdicts with their official code, date of the verdict and place of the low court/high court. The verdicts are ordered by date.

Variable	Coefficient	Significance
Interpretation level	-53.10257	0.997
Date1	-15.83909	0.999
Court	17.00202	0.999
Constant	18.08073	0.998
Log-likelihood	-6.2358727	
Pseudo R^2	0.9027	

Table 10: Multinomial logit model of the fallacy variable

Variable	Coefficient	Significance
Date	-0.0000926	0.511
Court	0.2066367	0.404
Thresholds		
k_1	-2.650655	-
k_2	-1.42435	-
Log-likelihood	-122.779	
Pseudo R^2	0.0042	

Table 11: Ordered probit model of the level of interpretation by the court

Variable	Coefficient	Significance
Date1	-0.1062899	0.701
Date2	0.0045928	0.986
Court	0.196832	0.425
Thresholds		
k_1	-0.9990578	-
k_2	0.2257482	-
Log-likelihood	-122.87084	
Pseudo R^2	0.0035	

Table 12: Ordered probit model of the level of interpretation by the court with date dummies

Variable	Coefficient	Significance
Date1	-0.0885427	0.699
Court	0.1882831	0.455
Seriousness	0.0898925	0.737
Penalty	-0.0006382	0.633
Thresholds		
k_1	-0.9645212	-
k_2	0.2620503	-
Log-likelihood	-122.70973	
Pseudo R^2	0.0048	

Table 13: Ordered probit model of the level of interpretation by the court with seriousness dummy and penalty variable