Eur J Law Econ (2018) 45:29–57
https://doi.org/10.1007/s10657-016-9539-y

An economic approach on countering the misuse of the right to challenge judges: an experiment

Joep Sonnemans[1] · Frans van Dijk[2] · Bart Donders[3] · Eddy Bauw[4]

Published online: 7 May 2016

Abstract Parties can challenge a judge (request a recusal) when they have reasons to believe that a judge is not impartial. In practice this procedure is sometimes abused by lawyers who, for example, want to delay proceedings. Countries have taken different measures to deter the improper use of the procedure to request a recusal, like fines for dismissed requests, or immediately dismissing evidently unfounded requests. In a laboratory experiment we examine the effects of a summary review whether a challenge is evidently unfounded, with or without fines. We find that a review without fine improves legal protection in practice as well as efficiency by reducing unfounded challenges and increasing challenges that have a substantial chance of success. Overall the number of challenges declines. With a fine, challenges decline, but also legal protection.

Keywords Challenge judges · Improper use · Experiment

JEL Classification K41 · C91

✉ Joep Sonnemans
j.h.sonnemans@uva.nl

[1] CREED, Tinbergen Institute, University of Amsterdam, Amsterdam, The Netherlands

[2] Netherlands Council for the Judiciary, The Hague, The Netherlands

[3] University of Amsterdam, Amsterdam, The Netherlands

[4] Utrecht Centre for Accountability and Liability Law, Utrecht University, Utrecht, The Netherlands

1 Introduction

Article 6 ECHR (as does Article 47 Charter of Fundamental Rights of the European Union) stipulates that in the determination of his civil rights and obligations or of any criminal charge against him, everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law. From this fundamental right follows, firstly, the primary responsibility of a judge to ensure his impartiality and to recuse himself when his impartiality can reasonably be questioned, and secondly, when the judge does not do this on his own motion, the necessity of a legal system to have a mechanism to challenge the jurisdiction of the judge or judges during court proceedings because of real or reasonably apprehended bias or lack of impartiality. This article focuses on the latter. The terminology differs between countries with different legal traditions and between regional and international courts. In this article we will refer to this mechanism as "challenging the (impartiality of the) judge(s)" and "challenging procedure". The mechanism is triggered by a "request" or "motion" "to recuse" or "for recusal" by one of the parties or by their legal representation on their behalf. We will use these different wordings without meaning any difference in substance. A distinction can be made between objective and subjective impartiality. Objective impartiality concerns the absence of factors that can be objectively established and throw doubt on the impartiality of the judge, especially conflicts of interest stemming from e.g. family or other social or business ties between the judge and a party or his legal representation. Subjective impartiality relates to personal attitudes or opinions of the judge that lead to doubt about his impartiality. While breaches of objective impartiality can in most countries be established more easily and objectively on the basis of existing guidelines or codes of conduct for judges and the facts of the case, infringement of subjective impartiality is much harder to establish, and therefore requires a full judicial procedure to decide charges of subjective partiality. For both categories it is sufficient for a motion to recuse to succeed, to establish that, despite the fact that in all legal systems the personal impartiality of a judge must be presumed until there is proof of the contrary,[1] there is an apprehension of bias, i.e. a reasonable and informed person would be concerned that the judge might be biased.

The challenging procedures differ across countries (Giesen et al. 2012). Differences are about, for instance, whether or not the main case is halted once a motion for recusal has been made, which judge or judges hear the motion, whether or not (financial) sanctions follow a rejected motion. A motion for recusal is in some jurisdictions[2] (in first instance) heard by the judge whose impartiality is questioned, in others (most European countries) the motion is heard by other judges from the same or in some cases from another court. If the motion is not followed by recusal, the judge continues. Otherwise, another judge takes his place. In many countries

[1] See for Europe the judgment of the European Court for Human Rights of 23 June 1981 (7238/75, 6878/75, (1981) 4 EHRR 1), Le Compte, Van Leuven and De Meyere v. Belgium.

[2] In the US, this is the case in a majority of states as well as in the federal courts (Buhai 2011). Another example is Canada (Bryden and Hughes 2011).

among which Belgium, France and the UK there are concerns about the improper use of this mechanism (Giesen et al. 2012). Improper use may aim at retarding the main procedure, at replacing the judge when a party has the impression that the judge will not rule in favor of him on legal and/or factual grounds, or at venting frustration with the legal system in general or the course of events in the case without offering any indication of partiality. In several countries, these concerns have led to a higher bar for charges of impartiality by, for instance, introducing a court fee or requiring that charges can only be made by a lawyer, and the introduction of financial sanctions in case of improper use. For instance, financial sanctions can be imposed in England, Belgium, Italy, Spain and Switzerland, while in France a financial sanction must be imposed if in criminal cases a charge is dismissed. Another approach is to summarily dismiss charges that evidently have no ground. See again Giesen et al. (2012). In other countries barriers are resisted. The Netherlands are a case in point. In 2013, 479 challenges were made in the first instance court, of which 12 were upheld, while in the appeal courts 147 challenges were made of which 9 were upheld (Raad voor de rechtspraak 2014, Tables 31 and 32).[3] These figures point to a very inefficient procedure, as there is no reason to assume that the decisions on challenges were biased against the challengers. It should be noted that not only improper use of the challenging procedure may cause this huge difference between challenges and upheld challenges. For instance, lack of information and emotional reactions to unexpected and/or undesired decisions of the judge play a role, especially for parties that do not have (adequate) legal counsel. This, however, does not alter the inefficiency of the procedure. Lawyers and judges alike consider barriers and, in particular, financial barriers an unwarranted infringement of the right of an impartial trial, although the general population believes otherwise (Van Rossum et al. 2012). Also, financial sanctions are considered to be ineffective by these actors. Empirical research about actual challenge behavior of parties is scarce. In Sect. 2 we will discuss the literature known to us. To our knowledge, no empirical research has been undertaken to establish the impact of barriers. In the absence of facts, everybody can pick his own argument, congenial to his normative position.

Given the divergent approaches and opinions, the subject merits research. The data for the Netherlands indicate that challenge procedures that have no barriers are extremely inefficient. The inefficiency is likely to increase, as in many countries the role of the judge in a case is changing from a passive to an active role, increasing the potential for conflicts between judge and parties (see Sect. 3). The issue is, therefore, how to increase the efficiency of challenge procedures, while maintaining the right to an impartial trial. We will focus on a financial sanction in the form of a fine and will address two questions. First, are fines effective in discouraging unfounded challenges? Second, do they lead to well-founded challenges not being

[3] We do not have data for other countries with a similar legal system. For Canada which has a quite different legal system data are available. Recusal is up to the judge, either on his own motion or because parties object, Bryden and Hughes summarize their data as follows: ".. most Canadian provincial and territorial court judges will recuse themselves more than once, but less than five times in a typical year, and that normally they will recuse themselves of their own motion …" (Bryden and Hughes 2011).

made? We use experimental methods in an economic framework to address these questions.

Section 2 addresses briefly what is known about challenge behavior, while Sect. 3 examines changes in the role of the judge that will have an effect on the use of challenge procedures. Section 4 sets out the model, Sect. 5 the predictions and Sect. 6 the experimental design. The results are presented in Sects. 7 and 8 concludes.

2 Challenge behavior in practice

We know only of a study in the Netherlands (Van Rossum et al. 2012). This study as well as the comparative study mentioned before (Giesen et al. 2012) were commissioned by the Netherlands Council for the Judiciary in response to the sharp increase of the number of challenges, experienced in that country. The number increased from 258 in 2007 to 607 in 2011, while the number of confirmed challenges increased from 16 to 36 (Van Rossum et al. 2012, p. 26).[4] The confirmation rate remained roughly the same. The number of the challenges is a very small percentage of the total volume of cases: 0.015 % in 2007, increasing to 0.034 % in 2011 (Van Rossum et al. 2012, p. 28). As pointed out in the study, the number of challenges is, however, high, when compared with the number of judges in the Netherlands (2500). Challenges have become a common phenomenon. As to the results of the study, numbers are presented about the grounds that were adduced to substantiate challenges (Van Rossum et al. 2012, p. 33). Challenges of objective impartiality were scarce (2.5 % in 2011). Thus, nearly all challenges were of a subjective nature. A major category concerns procedural decisions by the judges (e.g., planning of the hearings and whether or not to allow witnesses to be heard). This category accounts for 21.5 % of all challenges, and is relatively stable over time, according to the authors. A rapidly growing category is the treatment of parties or conduct of the case (disrespectful behavior, utterances that give an impression of partiality, suggestions that the judge already made up his mind, etc.). In 2011 the proportion was 33 %. Another category pertains to earlier decisions of the judge that suggest that he is not impartial (10 %), doubts about his professionalism (19 %), information deficiencies (3 %), distrust (2.8 %) and rest categories of unknown and miscellaneous reasons. The report gives also some anecdotal evidence that the grounds are not always the real reasons to challenge a judge. This happens, for instance, when a lawyer (unreasonably) wants to get more time. The authors found that, while in the decisions about challenges it is concluded in 12 % of the cases that the procedure has been abused, there is no shared definition of what amounts to abuse. What is often treated as abuse are challenges that are not aimed at the judge (but at the judiciary in general) or repeated challenges based on the same grounds. Strategic use aimed at getting more time or at getting a more favorable judge is not treated as abuse of the procedure. Challenges aimed at such

[4] As noted already, in 2013 the volume was 626, while only 21 challenges were confirmed (Raad voor de rechtspraak 2014).

objectives have not been identified. They can be found, in particular, in the categories procedural decisions and treatment of parties or conduct of the case, although it should be noted that the study finds that especially in situations where parties represent themselves anger and frustration play an important role. These people generally want to signal their dissatisfaction. Lawyers take a more rational approach by weighing advantages and disadvantages, and behave strategically. Both groups note that no financial costs have to be incurred and that that makes the decision to challenge easier. The authors find it surprising that the mechanism is not used more often by lawyers. While they found that lawyers use the mechanism strategically, they also found that lawyers are reluctant to use the instrument. Expressed reasons are that it prolongs the procedure, that a challenge puts a strain on the relationship with the judge, but also that lawyers still adhere to the social norm that judges are in principle impartial and that a challenge is only fitting in extreme cases. Whether this social norm will continue to be shared, can be doubted.

3 The role of judge and challenging behavior

There is reason to believe that unfounded challenges will increase. Generally, judges conducted and in some countries still conduct procedures in a passive manner. The judge behaves like a "sphinx". To quote from a recent report of the European Network of Councils for the Judiciary about judicial reform: "… in several countries the parties often only give long explanations about their view of the case, without any dialogue or questions from the judge and then the judge only fixes the date when the judgment will be pronounced. At the end of the hearing, parties have no idea which direction the verdict will take. Very much importance is given to the briefs, but the hearing would be much more interesting for the judges and the parties, if there would be more "discussion" or "dialogue"." (ENCJ 2013, p. 17). As the judge gives no indication of his thinking about the case, even if only by his questions, parties have no reason to doubt his subjective impartiality or to worry about the likely direction the verdict will take. This passive, neutral role is giving way to a much more active role (Bauw 2011). The report of the ENCJ describes this trend and advocates its further adoption as part of essential judicial reform. Case management is an important instrument. This is defined by the ENCJ as "…the judge taking the lead in resolving a legal conflict in a fair, expeditious and efficient manner. Case management applies to all areas of law. Within the law, the judge determines the procedure in cooperation with the parties and their legal representation, and ensures that this procedure is adhered to. The judge ensures that the procedure is commensurate with the complexity, size and relevance of the conflict. Therefore, it is the responsibility of the judge not only to decide the case, but also to direct it." (ENCJ 2013, p. 14). Pre-trial conferences to establish the proper method to resolve the case and to sort out differences of opinion about procedure are an essential component of case management. Obviously, the judge has to give insight in his thinking about the case. In addition, he runs the risk of giving too much away of his views or giving wrong impressions by expressing himself in an unfortunate manner. According to the ENCJ, judges struggle with this

new role: "A typical case is Belgium, where pre-trial conferences are short, and the only purpose is to create a more proactive approach and prepare questions for the parties in the final hearing. However this is more an exception, than the rule. Most of the Belgian judges clearly fear to give, by their questions, a statement about their position in the case, which would permit one of the parties to claim that the judge isn't neutral." (ENCJ 2013, p. 15).

The ENCJ also promotes the simplification of procedure, for instance, by restricting the number of procedural steps in a case. In its view, "it is entirely reasonable to require parties to supply the court with all relevant information up front, instead of holding back information for strategic reasons (see above on pre-trial conferences). Repeated exchange of arguments on paper could be disallowed, and replaced by a swift hearing, immediately followed by an oral or written verdict. … Methods currently used in on line dispute resolution may provide the courts with tools to have parties present and discuss their disputes in a more informal and interactive manner." (ENCJ 2013, p. 18). These developments will lead to more conflict in the court room. As the judge puts more pressure on parties to proceed in an expeditious manner, some parties and their legal representation will disagree with the judge and will try to gain time by the remaining means such as a charge of partiality. Also, as the judge reveals more of his thinking, parties will challenge his impartiality, either to put pressure on him or in the hope that the challenge succeeds and a new judge will think differently. It seems that these developments require the mechanisms to challenge judges to include barriers for unfounded challenges, while upholding the fundamental right to an impartial trial.

4 Model

There are several rational reasons for a party or lawyer to challenge. (1) There may be reasons to believe that the present judge is biased to his disadvantage and a replacement judge will rule more favorably; (2) there are no signals that the judge is biased, but there are indications that the judge will rule against a party on legal or factual grounds, for instance when the judge gives his interpretation of a relevant law or when he critically examines a witness. Replacement of the judge may lead to a new judge who has a different opinion; (3), a lawyer can request a challenge because, even if the challenge will be dismissed with certainty, the short delay gives him more time for preparation; (4), a challenge may signal the aggressiveness of the lawyer to the judge, to the adversary and to his present and prospective clients; (5), the replacement of the judge and the restart of the trial can be very advantageous for the party for whom the status quo is favorable, as the trial will be delayed substantially. For such a party, even a very small possibility of success makes a challenge worthwhile in expectation.

To model the situation in a format that can be studied in a laboratory, we have to simplify; see Table 1 for a summary of the sequence of events. Consider the following court case. Two business partners, Mr. Red and Mr. Blue, have decided to split up their company. However, they are unable to come to an agreement on how to split up the remaining capital (100 points in the experiment) and go to court. Mr. Blue is the

Table 1 The sequence of events in a period

1. Parties learn their color in this period (Red or Blue)
2. In part 2 only: the size of the fine (0,4 or 12) is announced
3. An Unbiased, Red or Blue judge is randomly chosen, each with probability 1/3; the bias of the judge is unknown to the parties
4. Parties observe 9 signals from this judge
5. Parties decide simultaneously to challenge or not
6. In part 2 only: unfounded challenges are dismissed and either go unpunished or are fined.
    a. When both parties don't challenge, the verdict and payoffs are announced; end of period
    b. When one or both parties challenge, than either:
        i. The judge is unbiased and is not replaced; the verdict and payoffs are announced; end of period
        ii. The judge is biased and will be replaced with a new judge (Unbiased, Red or Blue with probability 1/3).
            A. If this was the 4th judge, the new judge cannot be challenged; the verdict and payoffs are announced; end of period
            B. Otherwise, we return to step 4.

claimant and Mr. Red is the defendant. A judge will divide the 100 points between the parties. Judges can be biased in favor of the Blue party, the Red party or be unbiased. The unbiased judge will divide the 100 points evenly, while the Blue (Red) judge will appoint 75 points to Blue (Red) and the remaining 25 points to Red (Blue). The odds for each type of judge to be handling the case is always $1/3$ and it is never clear beforehand which type of judge has been appointed to the case. However, during the session the judge will give 9 signals that indicate potential partiality. For the unbiased judge each signal is Blue (or Red) with probability 50 %. For the Blue (Red) judge the probability of a Blue signal is 75 % (25 %) and a Red signal 25 % (75 %). After the signals are observed a request to challenge can be made. When the judge is not challenged, or when the challenging request is turned down because the judge was in fact impartial, the game ends by the judge dividing the 100 points. When a challenging request is made and the judge was Blue or Red, the judge is replaced and the procedure starts again. For practical reasons, we limited the number of times a judge can be replaced; the fourth judge cannot be challenged.

The payoffs for the Red and Blue parties in the control treatment (without summary review whether or not a challenge is evidently unfounded) are as follows. Representing the gains from a signal to a client (see point 4 above) or the extra preparation time (2), a party that requests a challenge earns 2 points regardless of the judge turned out to be partial or not. However, these points are only awarded when a subject requests a challenge of the first judge, i.e. in the first stage of the period.[5]

[5] This setup is somewhat unnatural in two aspects: in practice the gain in extra preparation time will be independent of which party challenge, and when a party is quite sure the other party will challenge it pays (2 points) to challenge too. We have chosen this simple setup for practical reasons.

Only one party, generally the defendant, gains from delaying the court by the replacement of a judge (5), and the other party, the claimant, loses. In the experiment it was assumed that every delay was a disadvantage for the Blue party, while the Red party would gain. This was represented by the fact that the Blue party had to pay the Red party 5 points whenever a judge was replaced by a new one after a successful challenge. Contrary to the gains from the signal to a client, this delaying effect also took place for the second and later judges/stages.

The final payoffs for the Blue party consists of the points awarded by the final judge (50 when this judge is unbiased, 25 when this judge is Red and 75 when this judge is Blue), *minus* the number of replaced judges multiplied by 5 points, plus 2 points only if Blue challenged the first judge. For the Red party the final payoffs are the points awarded by the final judge (50 when this judge is unbiased, 75 when this judge is Red and 25 when this judge is Blue), *plus* the number of replaced judges multiplied by 5 points, plus 2 points only if Red challenged the first judge.

The experimental treatments all three contain a summary review whether a challenge is evidently unfounded. A challenge is evidently unfounded if less signals or only marginally more signals of bias are in favor of the opposing party than of the challenger. A challenge after four or less signals in favor of the opposing party is always regarded as evidently unfounded and a request after 5 signals is evidently unfounded with a probability of 50 %. This probability reflects the uncertainty involved at the margin in judging whether a challenge meets the criterion of being evidently unfounded. In practice, parties cannot be certain about the outcome of the decision.

Challenges that are ruled to be evidently unfounded are summarily dismissed, either without further consequence for the challenger or at a cost to him. The level of the fine is announced at the beginning of a period and is 0, 4 or 12 points.

Note that the definition of the problem and the choice of parameters are heavily influenced by practical considerations. In the real world the percentage of biased judges is very low and signals in one or the other direction are rare, while in our experiment 66 % of the judges are biased and the parties always observe 9 signals. If we would use a more realistic percentage of biased judges and fewer signals we would need many more periods and participants, but the basic underlying problem is the same. In the real world the costs and benefits can vary widely (for example the benefit or cost of delay will differ between cases) but in the experiment we had to choose specific numbers. The different treatments (no fine and fines of different sizes) are chosen such that the theoretical predictions would be interesting and reasonable. This is completely in line with standard experimental economics methodology (Fréchette and Schotter 2015).

5 Predictions

We calculate the optimal behavior in this game under the assumption of risk neutrality. In the control condition, all judges are challenged, either by the Blue party if there are four or less Blue signals, or by the red party if there are 5 or more Blue signals. The first judge in a trial are challenged by *both* parties in the Nash

equilibrium, because of the 2 points one can earn by challenging the first judge (if the judge is likely to be biased in your favor, the other party will request a challenge and you can earn 2 points by *also* challenging). However, challenging when the judge seems to be on your side is a very risky strategy. For example, if the first judge provides 9 blue signals it is in expectation advantageous for Blue to challenge only if he expects that Red will challenge with a probability higher than 94.4 %.[6] In the treatment conditions, it is always best to challenge when it is certain that the review whether or not a challenge is evidently unfounded will be passed (0–3 signals of your own color). In the border cases of 4 or 5 blue signals, optimal behavior depends on the fine and the color: the Blue party will not challenge in the case with 4 Blue signals when the fine is 4 or 12, and the Red party will not challenge with 5 Blue signals when the fine is 12. As Table 2 shows, we expect as many challenges and replacements of judges in the control treatment and the fine = 0 treatment, but fewer when the fines are higher.

These predictions are under the assumption of risk neutrality. Risk neutrality is a reasonable assumption in the field when the decision makers encounter the same problem on a regular basis. Risk aversion will lead to fewer challenges of the first judge in the control treatment when the signals indicate a judge in your favor (see previous paragraph and footnote 2). In all other cases a reasonable level of risk aversion has no effect on the prediction, with only one exception: in the fine = 4 treatment a risk-averse red party will not challenge when there are 5 blue signals.[7] This very limited effect of risk-attitude on the predictions can intuitively be explained by the fact that every decision is a choice between *two* risky situations: both the bias of the current judge and a new judge is uncertain.

There is abundant evidence for social preferences in laboratory experiments: decision-makers care about relative earnings. In games like the ultimatum or dictator game many decision-makers appear to be inequity-averse (Fehr and Schmidt 1999). In other cases, where the situation is more like a competition, decision-makers typically want to earn more than the other and "to win" (e.g. Bault et al. 2008). Our experimental setup, with two antagonistic parties meeting in court, is more like the second case and may enhance competitive attitudes. Competitive preferences would influence specifically the Red party, because a successful challenge transfers 5 points from the Blue to the Red party.

Finally, there are interesting empirical studies that show that fines may have an effect in the opposite direction than we expect. For example, Gneezy and Rustichini (2000) report a field experiment at day-care centers where some parents used to arrive late to collect their children. When introducing a monetary fine for parents who arrived more than 10 min late to pick up their children, they found that the

[6] With 8, 7 and 6 Blue signals the probabilities of a Red challenge should be at least 94.1, 93.3 and 90.5 %, respectively. For the Red party it is only in expectation profitable to challenge with 0, 1, 2 and 3 Blue signals when the probability of a Blue challenge are at least are 84.1, 83.3, 80.7 and 70.0 %, respectively.

[7] We calculated optimal behavior under the assumption that all participants have the same utility function $U(x) = x^{(1-\rho)}$. We find that risk aversion leads to less challenges in the fine = 4 treatment: for $\rho > 0.2$ the red party will *not* challenge in case of 5 blue signals. In the other cases and treatments even extreme ρ's (e.g. 0.8) do not change the predictions.

Table 2 Predictions for risk-neutral parties

	Control	Treatment		
		Fine = 0	Fine = 4	Fine = 12
Number of Blue Signals				
0–3	B	B	B	B
4	B	B	–	–
5	R	R	R	–
6–9	R	R	R	R
Expected earnings Red	60.02	58.89	59.03	56.32
Expected earnings Blue	43.98	42.84	42.35	44.74

A B (R) indicates that the Blue (Red) party will challenge. The predictions are not different for the first judge or different judges, with one exception: In the control treatment both parties will challenge the first judge in the Nash equilibrium, because of the 2 points premium (see main text)

number of late-coming parents increased significantly. After the fine was removed no reduction occurred. The explanation of this result is that the fine changed the parents' perception of the social interaction in which they are involved. Without the fine the costs of coming late were not exactly defined. There were only abstract costs of breaking a social norm, i.e. the norm that it is socially unacceptable to let the teachers wait. By contrast, after the fine had been introduced the consequences of coming late were defined precisely. Probably parents now interpret the fine as the only cost for letting the teachers wait. Hence, the effect of a fine could lead to a different outcome than the deterrence hypothesis would predict. Similar behavior could take place in our experiment, where the social norm might be that the impartiality of judges is not to be put in doubt. While in our experiment the judges are virtual and not participants in the experiment, which makes the social context of the decisions less central, the experiment is deliberately set in a judicial frame. This frame triggers the legal and social norms of the courts, as participants imagine these.

6 Experimental design

The experiment was run in the laboratory of CREED at the University of Amsterdam in January 2013. In total, 80 students participated as subjects in four sessions and earned on average 30 euro in about 2 h. A within-subject design was employed: in the first part (10 periods) there are no checks whether or not the challenges are evidently unfounded (control) and in the second part (30 periods) unfounded challenges are refused and the requester has to pay a fine of 0, 4 or 12 points (treatment).[8] Each period could consist of a maximum of five sequential stages in which all subjects that were still in play had to make a decision whether or

[8] A possible disadvantage of using a within-subject design is order effects because of learning. However, in this case we start with the control treatment which represents the current situation and the summary review whether challenges are evidently unfounded and fines are introduced only after the subjects have experience in the control treatment, like in the real world.

not to request a challenge. Before starting experiment subjects' risk preference was measured using the Holt and Laury test.[9]

Subjects were given a monetary incentive to perform: in each period depending on their choices and the choices of their opponent they could earn points, i.e. 1000 points equalled € 10. At the end of the experiment their earnings of all periods and the Holt and Laury test were totaled and paid out in cash anonymously.

In the experiment the following court case was sketched. Two business partners, Mr. Red and Mr. Blue, have decided to split up their company. However, they are unable to come to an agreement on how to split up the remaining 100 points and go to court. In each period two parties of opposing parties were randomly and anonymously assigned to play against each other. All participants were informed that they would never play the same opponent in consecutive periods.

6.1 Procedure

At the beginning of the experiment each of the subjects was assigned to a private cubicle and computer according to a randomly drawn code. Communication between subjects was not allowed. When all the subjects were seated the experiment started with the Holt and Laury (2002) test in which subjects have to choose ten times between two lotteries.

Subsequently, the instructions[10] of the first part of the experiment (control treatment) were shown on the screen and were handed out on paper. Subjects had to answer three questions that checked their understanding of the experiment, i.e. they had to calculate a party's payoff for an outlined situation that could occur during the experiment. It was stressed to the participants that the situations were random and that the mentioned choices are not necessarily wise. They were also given the opportunity to ask questions, which an experimenter answered privately.

The control treatment consisted of 10 periods with re-matching after each period.[11] At the beginning each period the subjects learned their color for that period and the nine signals, represented by red and blue balls, were shown. Hence, the participants could make inferences on the partiality of the sitting judge and the corresponding potential verdict. Based on this information both opposing lawyers simultaneously had to make a decision on whether or not to request a challenging procedure.

When both lawyers decided not to request a challenge, the judge would come to a verdict and the 100 points were distributed accordingly. When one or both lawyers decided to challenge, the partiality of the judge was checked. In case the judge turned out to be neutral, that judge would remain on the case and would give the verdict. However, in case the sitting judge turned out to be partial, this judge would be replaced by a new one. This would entail that the period would proceed to the second stage, which starts of—as do all stages—with nine signals from the new

[9] See Holt and Laury (2002) for more information on the design and procedure of this test.

[10] See Appendix 1 for the instructions used in the experiment.

[11] In each session the subjects were divided in two groups and re-matching was within these groups. For the most conservative statistical analyses we use these 8 groups as independent observations.

judge. In effect, this would return the game to the starting point when both parties again have to decide whether to request a challenge. In turn, the period could go on for five stages, i.e. the participants were informed that after the replacement of four judges, the fifth judge cannot be challenged and will immediately come to a verdict. When a judge has come to a verdict the period ends and the earnings—the division of the 100 points according to the verdict and gains or losses from the signals to clients and the delay of the court—are revealed.

After the end of part one of the experiment subjects received the instructions for the second part of the experiment. These were also shown on the computer screen. Again subjects had to answer control questions to check the understanding of the experiment.

The second part of the experiment consisted of 30 periods, with random re-matching after each period. At the start of each period the subjects learned the fine for an evidently unfounded challenging request (0, 4 or 12), which would stay constant during that period. As in the control treatment, subjects had to decide whether or not to put forward a challenging request after learning the nine signals. When both lawyers decided not to request a challenge, the verdict of the judge was revealed. When one or both lawyers decided to put forward a challenge, this request was firstly tested whether or not it is evidently unfounded. Hence, if the challenge was found to be unfounded the requester would have to pay a fine and the challenge of the sitting judge was refused. Subsequently, the judge would come to a verdict and both parties' earnings for that period were revealed. However, if the request was found not to be unfounded, the partiality of the judge was checked. As in part one, if the judge turned out to be neutral, that judge would remain on the case and would give the verdict, and in case the sitting judge turned out to be partial, the judge would be replaced by a new one and the period would proceed to stage two. As in part one, the fifth judge could not be challenged.

After the end of part two, the earnings of the subjects were totaled and revealed to the participant. This also entailed that the computer randomly picked which lottery was played out of the ten chosen lotteries chosen earlier by the subjects in the Holt and Laury test. Subsequently, both the earnings from the Holt and Laury lottery and the rest of the experiment were revealed privately to the subject. Lastly, the participants were asked to give some personal information regarding their age, gender and study. Before leaving the lab all participants were paid out privately.

7 Results

To give a first impression, we consider the overall fraction of challenges in the treatments, compared with the predictions of Table 2 (Fig. 1). We have to distinguish between the first judge and subsequent judges, because of the 2 points premium for challenging the first judge, and between the behavior of Blue and Red parties. The introduction of a review whether or not a challenge is evidently unfounded without a fine decreases the challenges by Red parties substantially (see Table 3 for statistical tests and p values). For Blue parties we do not find a statistically significant effect for the first judge, but for subsequent judges the

number of challenges significantly *increases*. The test communicates apparently also that challenges based on solid information are allowed. Interestingly, it can be

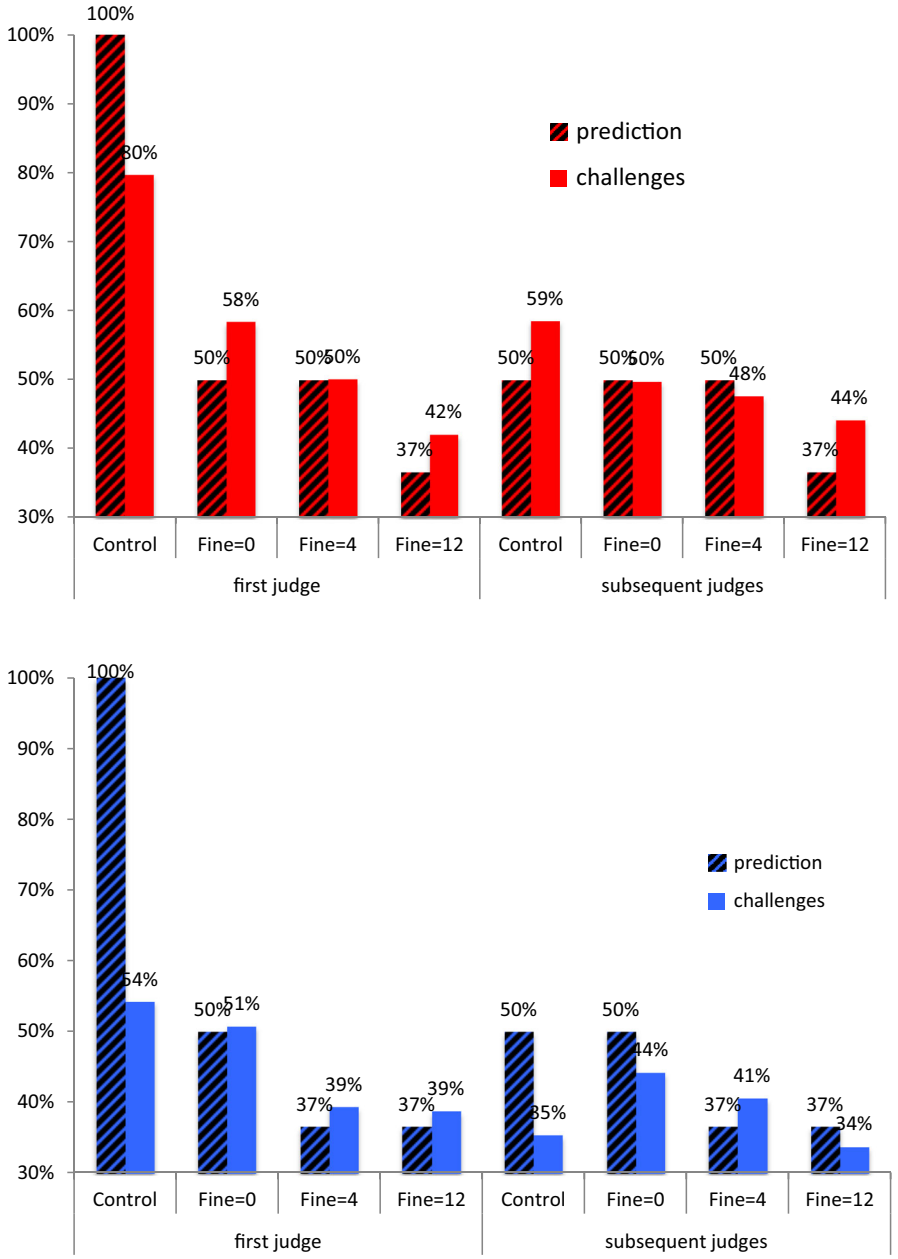

Fig. 1 The overall fraction of challenges by the Red players (*top*) and Blue players (*bottom*), with next to it the predictions of Table 2

Table 3 First four rows: percentage of challenges with between brackets the number of choices

Treatment	Red player			Total	Blue player			Total
	Number of Blue signals				Number of Blue signals			
	0–4 Improper	5 Uncertain	6–9 Proper		0–3 Proper	4 Uncertain	5–9 Improper	
First judge								
Control	62.5 % (184)	94.4 % (54)	94.4 % (162)	79.8 % (400)	86.8 % (136)	50.0 % (48)	34.8 % (216)	54.3 % (400)
Fine = 0	23.9 % (205)	91.8 % (49)	97.8 % (138)	58.4 % (392)	92.4 % (145)	63.3 % (60)	14.5 % (187)	50.8 % (392)
Fine = 4	7.0 % (201)	75.5 % (53)	95.7 % (162)	50.1 % (416)	97.9 % (143)	29.3 % (58)	3.2 % (215)	39.4 % (416)
Fine = 12	1.9 % (203)	51.1 % (47)	96.5 % (142)	42.1 % (392)	94.6 % (147)	7.1 % (56)	4.7 % (189)	38.8 % (392)
C versus 0	**0.012**	0.498	0.173	**0.012**	0.398	0.310	**0.017**	0.208
C versus 4	**0.012**	0.116	0.735	**0.012**	**0.043**	0.484	**0.012**	**0.012**
C versus 12	**0.012**	**0.018**	0.400	**0.012**	0.091	**0.017**	**0.012**	**0.017**
0 versus 4	**0.012**	0.207	0.225	**0.025**	0.128	0.069	**0.036**	**0.017**
0 versus 12	**0.012**	**0.012**	0.465	**0.012**	0.398	**0.012**	**0.036**	0.093
4 versus 12	0.063	**0.028**	0.686	**0.017**	0.144	**0.043**	0.484	0.779
Subsequent judges								
Control	22.0 % (241)	85.7 % (63)	97.8 % (182)	58.6 % (486)	83.7 % (178)	15.9 % (63)	5.3 % (245)	35.4 % (486)
Fine = 0	12.9 % (232)	95.7 % (46)	94.4 % (144)	49.8 % (422)	91.9 % (173)	40.7 % (59)	2.1 % (190)	44.3 % (422)
Fine = 4	3.4 % (239)	76.5 % (51)	97.3 % (183)	47.7 % (473)	97.1 % (173)	28.8 % (66)	2.1 % (234)	40.6 % (473)
Fine = 12	4.3 % (209)	39.7 % (58)	96.4 % (165)	44.2 % (432)	92.6 % (149)	10.0 % (60)	0.9 % (223)	33.8 % (432)
C versus 0	**0.025**	0.115	0.263	0.069	0.123	**0.025**	0.183	**0.017**
C versus 4	**0.012**	0.484	0.866	**0.025**	**0.017**	0.203	0.069	0.208
C versus 12	**0.017**	**0.012**	0.498	**0.012**	**0.036**	0.401	**0.018**	0.779
0 versus 4	**0.017**	**0.018**	0.080	0.484	**0.018**	**0.025**	0.500	0.093
0 versus 12	**0.050**	**0.012**	0.499	0.161	0.674	**0.018**	0.715	**0.036**
4 versus 12	0.575	0.093	0.225	0.327	0.176	0.063	0.345	**0.050**

Bottom six rows: statistical tests (two-sided p values of Wilcoxon tests based upon the 8 independent observations of the matching groups, printed in bold when $p < 0.05$)

concluded that a review without fine leads to an improvement of efficiency (less unfounded challenges) as well as a higher degree of legal protection (more not-unfounded challenges).

Introducing a fine of 4 decreases the number of challenges of both the Blue parties (in line with the prediction; $p < 0.05$ for first and $p < 0.10$ for subsequent judges) and Red parties (against the prediction, $p < 0.05$ for the first judge and n.s. for subsequent judges), compared with a 0 fine. Increasing the fine from 4 to 12 decreases the Red challenges of the first ($p < 0.05$) but not subsequent judges, and keeps the Blue challenges of the first judge unchanged and decreases the challenges of subsequent judges ($p = 0.05$).

Next we study the occurrence of challenging for the signals where challenges would be not founded, unfounded or uncertain. Figure 2 and Table 3 shows averages and statistical tests for the cases with a signal that indicate a negatively biased judge.

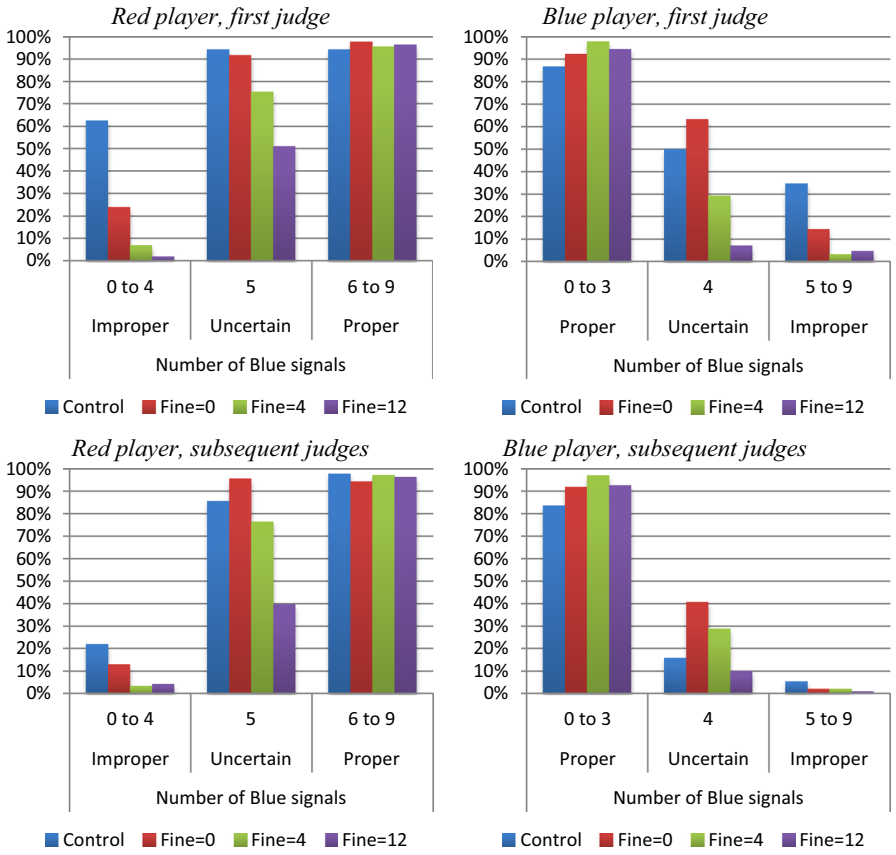

Fig. 2 Percentage of challenge requests by Red (*left*) and Blue (*right*) players for the first (*top*) or subsequent judges in a period

When a challenge is well-founded, both the red party and the blue party are very likely to challenge in all treatments (94–98 % for red and 87–97 % for blue, see Fig. 2 or Table 3). This is in line with our prediction: introducing a review with or without fine does not influence the challenging behavior when the judge is clearly biased.

In the control treatment a party who receives favorable signals about the first judge, but is quite sure that the other party will challenge, should challenge too because of the 2 points premium. As explained in footnote 3, this can be quite risky, especially for the blue party. Indeed, the blue party is on average more reluctant than the red party to take that risk and challenges a favorable judge (borderline favorable 26 vs 79 %, 6–9 favorable signals 38 vs 57 %, see Table 3: Wilcoxon test on matching group level, two-sided $p = 0.012$ and $p = 0.018$).

In case of a favorable signal subsequent judges should not be challenged, according to our prediction. Indeed, blue parties do rarely challenge when there are 5 or more blue signals. However, red parties challenge subsequent judges with 0–4 blue signals in 22 % of the cases in the control treatment (see Fig. 2) and in 46 % when there are exactly 4 Blue signals (not in table). These challenges are considered evidently unfounded in the fine-treatments and become rare when the fine is larger than 0.

Now we turn to the borderline "uncertain" cases where a challenge is considered unfounded with 50 % chance. We first consider the Red parties when there are 5 Blue signals. The prediction is that the Red party will only cease challenging when the fine is increased to 12 points. For the first judge the percentages of challenges do not significantly differ between the control and the fine = 0 or fine = 4 treatments, but we find significantly fewer challenges in the fine = 12 treatment than the other three treatments (see Table 3). This change is in line with our predictions, however, there are still a considerable number of challenges when the fine is 12 (51 % for the first and 40 % for subsequent judges). This suggests competitive social preferences of at least some subjects.

When there are 4 Blue signals the Blue parties are predicted to challenge only in the control and the fine = 0 treatments. Surprisingly, in the control treatment they challenge in only 50 % (first judge) and 16 % (subsequent judges) of these cases. Challenges are more common in the fine = 0 treatment and increase to 63 % (n.s.) and 41 % (statistically significant), respectively. Defining in what cases challenging is unfounded implicitly defines challenging founded in all other situations and this may have increased the occurrence of challenging. However, learning is an alternative explanation of this result. Challenges decrease again with the introduction of fines: for the first judge from 63 % (fine = 0) to 29 % (fine = 4, comparison fine is 0 and 4: $p = 0.069$) to 7 % (fine = 12, comparison fine = 4 and 12: $p = 0.043$) and for subsequent judges from 41 % (fine = 0) to 29 % (fine = 4, comparison fine is 0 and 4: $p = 0.025$) to 10 % (fine = 12, comparison fine = 4 and 12: $p = 0.063$).

Finally, we turn to the effect of the risk attitude on the challenge decisions. Based upon the Holt and Laury test we divide the participants in risk-averse ($\geq$6 A-choices, N = 43), risk-neutral (5 A-choices, N = 9) and risk loving ($\leq$4 A choices, N = 19) subjects. Four subjects never switched from the A to the B-choice and 5 subjects made inconsistent choices (switched more than once): these participants are excluded for this analysis. We find no systematic effect of risk attitude on challenge decisions (see

appendix 2). In the specific case where we predicted fewer challenges for risk averse participants (fine = 4, Red party, 5 Blue signals) the data are in the predicted direction (the first judge is challenged in 69 % of the cases by risk-averse subjects and 90 % and 100 % by respectively risk-neutral and risk-loving subjects; for the subsequent judges these numbers are 63.3, 100 and 94.4 % respectively) but these differences are not statistically significant (note however that the number of observations and thus the power of the tests are small).

Interestingly, it can be concluded that a summary review without fine leads to an improvement of efficiency (less unfounded challenges) as well as a higher degree of legal protection (more founded challenges). A small fine of 4 effectively reduces evidently unfounded challenges to close to zero and also reduces uncertain challenges that may or may not be deemed evidently unfounded by a judge. A fine of 12 leads only to a further reduction of uncertain challenges. It can be concluded that with the introduction of a fine a trade-off is created between efficiency and legal protection. This is especially the case with a fine of 12.

8 Conclusion

In this paper we have addressed the strategic (mis)use of the procedures for challenging judges because of alleged impartiality. Such procedures exist in most legal systems. The strategic use of these procedures, to gain time or to get a judge who might think differently about specific legal matters than the current judge, is an issue in several jurisdictions. Procedures for challenging judges exist to provide legal safeguards against the eventuality of judges that are not impartial. Legal protection is paramount. However, challenges prolong procedures and consume resources. An efficient mechanism is, therefore, also important. Ideally, such a mechanism would deter challenges for strategic reasons and without (substantial) evidence, but would not deter parties from challenging judges when (substantial) evidence is available. Evidently, there will be a trade-off between legal protection and efficiency, when it comes to determining the burden of proof. How this trade-off actually works out is an empirical matter that is difficult to research empirically, other than by experimental means.

We designed an experiment to capture key aspects of the interaction in a simplified and abstract manner. It was set up in such a way that rationally without counter measures judges are frequently challenged. This provides the opportunity to test procedural remedies. Subjects actually challenged judges very often. When there are strong indications of partiality and thus challenges are founded on fact (six or more signals out of nine are against a party, and thus three or less in favour), challenges are nearly always made: 94.4 % of initial judges and 97.8 % of subsequent judges who are called upon after a successful, initial challenge and possibly further successful challenges, in case of parties that benefit from delay ("red" parties, generally the defendants) or very often (86.8 % of first, and 83.7 % of subsequent judges) for the opposing parties that are disadvantaged by delay ("blue" parties, generally the claimants). Also, many challenges are made especially by the red parties (62.5 % of first and 22.0 % of subsequent judges)

and, to a lesser degree, by the blue parties (34.8 % of first, 5.3 % of subsequent judges), even if the signals are in favour of these parties themselves (five or more signals in favour of a party and thus four against it). In between are uncertain situations with weak evidence of partiality (five signals against a party and four in favour of it). Then, red parties challenged 94.4 % of first judges (85.7 % of subsequent), blue parties only 50.0 % (15.9 %).

A summary review whether a challenge is evidently unfounded was introduced. If that was found to be the case, the challenge was dismissed without further consequences for the claimant. The outcome of the test is determined with certainty if the signals point in a clear direction as demarcated above. In the intermediate situations the decision can fall both ways. The impact of the introduction of the review is that the number of unfounded challenges is reduced from 62.5 to 23.9 %[12] (22.0–12.9 % for subsequent judges) for red parties and from 34.8 to 14.5 % (5.3–2.1 % for subsequent judges which is not significant) for blue parties. Firmly founded challenges are not affected significantly. Uncertain challenges by red parties are not affected, but these challenges by blue parties of first judges increase from 50.0 to 63.3 % (not significant) and of subsequent judges from 15.9 to 40.7 % (significant). The total number of challenges by red parties of first judges declines (79.8–58.4 %), while the challenges of subsequent judges do not significantly decline; the challenges by blue parties of first judges remain constant and increase for subsequent judges from 35.4 to 44.3 %. Overall, the total number of challenges declines, as mentioned, for first judges and remains constant for subsequent judges. To conclude, the review mechanisms leads to a reduction of unfounded challenges and an increase of challenges that stand a good chance of success. Overall, the number of challenges declines. The review mechanism serves both purposes: legal protection and efficiency.

Next, fines were attached to the review. If a challenge was ruled to be evidently unfounded, a fine was imposed. The impact of a small fine is a sharp decline of unfounded challenges relative to a review without fine: for red parties from 23.9 to 7.0 % of first judges, and 12.9–3.4 % of subsequent judges. And for blue parties from 14.5 to 3.2 % of first judges and constant for subsequent judges. The number of warranted challenges does not change significantly. The uncertain challenges, however, decrease: for red parties from 91.8 to 75.5 % of first judges (not significant) and from 95.7 to 76.5 % of subsequent judges, and for blue parties from 63.3 to 29.3 % of first judges (only marginally significant, $p = 0.07$) and from 40.7 to 28.8 % of subsequent judges. Overall, the total number of challenges declines, both for first and for subsequent judges. There is a trade-off between legal protection and efficiency.

Increasing fines threefold results, in particular, in a vast reduction of uncertain challenges. For all situations except blue parties and subsequent judges the number of challenges is far below the number without any intervention. For blue parties the number a challenges of subsequent judges is at the same (low) level as initially. Consequently, the total number of challenges decreases further. It was pointed out by an anonymous referee that, if courts are allowed to keep the proceeds from the

[12] All differences are significant at $p < 0.05$ unless explicitly mentioned otherwise.

fines, this may lead to or—we may add -raise the suspicion that it will lead to a stronger incentive to dismiss challenges. This potential incentive problem was not included in the experiment, but the issue can be easily avoided by allocating the proceeds elsewhere.

The results are clear and consistent. Introduction of a review to filter out evidently unfounded challenges increases the effectiveness of legal protection, as more parties with a substantial chance of success use the opportunity, and it reduces the number of challenges. This dominates the control condition, not even taking into account that it saves the judiciary time when challenges are summarily dismissed. Attaching a fine to the review increases the efficiency of the challenge mechanism, but reduces the effectiveness of the legal protection offered by the mechanism. These effects are more extreme for higher fines.

A clear policy advice can be given when strategic use of challenging procedures leads to inefficiencies. A review without a fine should be considered if the policy maker attaches foremost value to legal protection. This actually increases the effectiveness of legal protection, while at the same time increasing efficiency. A high fine should be considered, if efficiency considerations are dominant, and one therefore wants to confine legal protection to challenges that are without any uncertainty founded. There is a trade-off between legal protection and efficiency. More emphasis on legal protection requires a lower fine. It is up to the policy maker to decide on the balance. It should be noted that we focused on informed, strategic litigation, generally requiring knowledgeable council. Self-represented litigants are likely to display much more erratic, ill-informed and emotion based behaviour, and, while our policy advice may well affect the decisions of these litigants in the desired direction, the experiment does not address that behaviour. Still, in the vast majority of major criminal and civil cases lawyers play a central role, and judicial procedures must be capable of dealing with their strategic behaviour in an efficient manner.

Appendix 1: Translation (from Dutch) of the instructions

The experiment of today

The experiment is a coproduction of CREED (University of Amsterdam) and the Council for the Judiciary. The experiment is about challenging judges. Challenging is a court procedure, which is designed to check the objectivity of a judge to judge. It thus helps to ensure the independence and impartiality of the judge. Someone who is a party in a lawsuit must be able to trust that the sitting judge is impartial. A judge must be able to make an objective judgment. If there is reason to believe that a judge is partial, the party that believes to be harmed can request a challenge. When the

challenging chamber establishes that the judge is indeed biased, this judge will be replaced.

If during the experiment you have a question, please raise your hand.

Before we begin the actual experiment, we ask you to choose ten times between two lotteries **A** and **B**. You should always choose one of these lotteries. At the end of the experiment, one of your choices is played out. Which one is chosen will be determined by chance. The questions all have this form:

Option A: 4/10 chance of 2.00, 6/10 chance of 1.60
or
Option B: 4/10 chance of 3.85, 6/10 chance of 0.10

where the **probabilities** in the lotteries will vary each question.

Preliminary questionnaire

Below you are shown 10 times a paired lottery, **A** and **B**. De numbers always represent euros. You have to choose one of these lotteries. After the 10th choice one choice will be played out. Which one is chosen will be determined by chance.

1	○	Option **A**: **1/10** chance of **2.00**, **9/10** chance of **1.60**	*or*	○ Option **B**: **1/10** chance of **3.85**, **9/10** chance of **0.10**
2	○	Option **A**: **2/10** chance of **2.00**, **8/10** chance of **1.60**	*or*	○ Option **B**: **2/10** chance of **3.85**, **8/10** chance of **0.10**
3	○	Option **A**: **3/10** chance of **2.00**, **7/10** chance of **1.60**	*or*	○ Option **B**: **3/10** chance of **3.85**, **7/10** chance of **0.10**
4	○	Option **A**: **4/10** chance of **2.00**, **6/10** chance of **1.60**	*or*	○ Option **B**: **4/10** chance of **3.85**, **6/10** chance of **0.10**
5	○	Option **A**: **5/10** chance of **2.00**, **5/10** chance of **1.60**	*or*	○ Option **B**: **5/10** chance of **3.85**, **5/10** chance of **0.10**
6	○	Option **A**: **6/10** chance of **2.00**, **4/10** chance of **1.60**	*or*	○ Option **B**: **6/10** chance of **3.85**, **4/10** chance of **0.10**
7	○	Option **A**: **7/10** chance of **2.00**, **3/10** chance of **1.60**	*or*	○ Option **B**: **7/10** chance of **3.85**, **3/10** chance of **0.10**
8	○	Option **A**: **8/10** chance of **2.00**, **2/10** chance of **1.60**	*or*	○ Option **B**: **8/10** chance of **3.85**, **2/10** chance of **0.10**
9	○	Option **A**: **9/10** chance of **2.00**, **1/10** chance of **1.60**	*or*	○ Option **B**: **9/10** chance of **3.85**, **1/10** chance of **0.10**
10	○	Option **A**: **10/10** chance of **2.00**, **0/10** chance of **1.60**	*or*	○ Option **B**: **10/10** chance of **3.85**, **0/10** chance of **0.10**

Instructions experiment

Shortly, a summary of these instructions will also be distributed on paper for use during the experiment. The experiment consists of two parts, respectively 10, and 30 periods. Every period you are matched to another participant. You will never be linked to the same participant two periods in a row. We first explain the situation.

The situation

Subject of the experiment: challenge of judges This experiment is about challenging judges. A challenge is a court procedure, which is designed to check the objectivity of the judge. It thus helps to ensure the independence and impartiality of the judge. Someone who is a party in a lawsuit must be able to trust that the sitting judge is impartial. A judge must be able to objectively come to a verdict. When there is reason to believe that a judge is partial, the party that believes to be harmed can request a challenge. When the challenging chamber establishes that the judge is indeed biased, this judge must subsequently be replaced.

The impression is that in practice lawyers may not only challenge to check the alleged partiality of the judge. There could also be different reasons to request a challenge. Two of those other reasons are:

- **Signal to the client:**
  The lawyer can also request a challenge to give a signal to the client that he is an aggressive lawyer that is prepared to do everything for his client.
- **Delay:**
  In some cases, lawyers are thought to use the procedure in order to slow down the judicial process, because this could turn out to be beneficial for their party. For instance, it could be that a lawyer needs time to gather additional evidence or the objective could be to exhaust the counterpart. By starting a challenging procedure the lawyer can delay justice with a few days or weeks. If the judge actually is found to be partial and must be replaced, the whole process will have to be redone and a new judge will have to orientate herself. This may result in a delay of several months or, in complex cases, even a year.

Design of the experiment

The lawyers In this experiment, the following trial is covered. Two entrepreneurs, Mr. Red and Mr. Blue, have set up a successful business over the past 3 years. They have decided to split up and to divide the company. However, they do not agree on the division of property worth 100 points and it comes to a lawsuit. In this lawsuit and throughout the experiment, you are either lawyer of the Red Party or the lawyer of the Blue Party. The lawyer of the other party is another participant in this experiment. Each point you earn for your client earns you money: 1000 points equals 10 euro.

The judge The distribution of 100 points is determined by the judge assigned to the case. In this experiment, this sitting judge is represented by a basket of red and blue balls. The division between red and blue balls in the basket determines the bias/partiality of the judge and the resulting verdict. For each red ball in the basket the judge will give 1 point to the Red party and for every blue ball the judge will give 1 point to Blue. In total there are thus 100 (red and blue) balls in the basket. There are three types of judges (baskets of balls) and corresponding verdicts:

- A **neutral judge** has 50 red and 50 blue balls in the basket and divides the property in 50 points per party.
- A **red judge** has 75 red and 25 blue balls and will provide that Red gets 75 points and Blue 25 points.
- A **blue judge** has 25 red and 75 blue balls in the basket and will provide that Red gets 25 points and Blue 75 points.

The probability that one of these three types of judges will be on a case is always $1/3$. It is never clear what kind of judge is assigned to the case. However, during the court case the judge gives signals as an indication of his possible partiality and thus the final verdict. In this experiment, these signals are nine random draws with replacement from the basket with 100 balls belonging to the judge that is currently the case.

Instructions part 1

Part one of the experiment consists of ten periods, and is further described below:

Challenging procedure

After the drawn balls are shown to both lawyers, you and your opponent, you will simultaneously be asked to decide whether you want to challenge or not. If both parties decide not to challenge, the verdict of the judge will follow. Once you or your opponent decides to challenge the judge the contents of the basket of balls will be revealed and the potential partiality of the judge is visible to both parties. Should the judge turn out to be neutral, then the judge remains and the verdict will subsequently follow. However, if the judge is found to be partial, the judge will be replaced for a new judge (and thus a new basket of red and blue balls). This new judge will again be neutral, biased for Red or for Blue with a probability of $1/3$. From the new basket of balls belonging to the new judge again balls will be drawn with replacement and shown to the lawyers, so the game is back to the initial situation.

Attention! Per trial only four judges can be replaced; hence the fifth judge cannot be challenged and he will come to his verdict immediately.

The gains and losses of a challenge for both parties

- **Signal to the client**
  This is an advantage for the lawyer that is requesting the challenge; he will earn 2 points. These point will be awarded regardless of the challenge being successful. However, these revenues are only possible when the first judge of each period is challenged.
- **Delay of the lawsuit by disqualification**
  In practice, a successful challenge (leading to a replacement judge) will result in justice being seriously delayed, because the complete trial has to be conducted again. This is often advantageous for one particular party, while it is at the expense of the position of the other party. In this experiment, such a delay will be in favor of Red and is disadvantageous for Blue. Therefore, in this experiment Blue will have to pay 5 points to Red when a judge is replaced by a successful challenge. These gains and losses from delaying the court case occur after each successful challenge, so also for a successful challenge of the second or later judges in the court case.

Next part of the procedure

Once a decision is made, the proceeds of the period are made known to the participant. Then comes the next period. Here you will be up against another opponent. You never play two periods in a row against the same opponent. Naturally, in the new period a new judge is put on the case.

The sequence of a period Below is a point-by-point breakdown of the sequence of a period:

1. Each period begins with a new judge and a new opponent (another participant than in the previous period), and your role (lawyer of Red or Blue) is revealed.
2. The draw of the balls is visible to both parties.
3. Both lawyers decide whether to challenge or not:

    - Both lawyers do not challenge: the verdict follows (step 4).
    - One or both lawyers request a challenge: **+2 points** for the requesting party(ies), ***Attention***, this benefit occurs only for the first judge;

        - (Im)partiality of the judge is revealed:

            - Judge is neutral/impartial: the verdict follows (step 4).
            - Judge is partial: **+5 points** for Red and **−5 points** for Blue;

            A new judge is appointed and we go back to step 1.

4. The verdict: earnings of the period are made public. New period with a new judge and opponent (step 1).

We are now going to ask some questions to make sure that you have understood everything.

Practice questions part 1

Below we have some questions to be sure that you have understood everything. The situations outlined are random and the choices are not necessarily wise! Please raise your hand if you do not understand something or have a question.

A.
You are the lawyer of party Red. There are 4 red and 5 blue balls in the random draw. You decide to challenge, but the judge proves to be neutral. What are in this period the total earnings for party Red?

points

B.
You are the lawyer of party Red. There are 5 red and 4 blue balls drawn. The lawyer of party Blue decides to challenge the judge and he turn out to be partial. The new judge consists of a draw of 3 red and 6 blue balls and is not challenged. However the judge turnes out to be partial in favor of blue. What are this period's total earnings for party Red?

points

C.
You are the lawyer party Blue. There is 1 red and 8 blue balls in the draw. You decide to challenge the judge and he turns out to be partial. With a new judge you decide to challenge again with a draw of 4 red and 5 blue balls. Again the judge proves to be biased. With the new judge your opponent decides to challenge, but the judge proves to be neutral. What are this period's total earnings for Party Blue?

points

Instructions part 2

(These instructions will shortly be distributed on paper to be used during the experiment.)

Part 2 of the experiment consists of thirty periods and is to a large extent similar to part 1. However, there is an addition to Part 1. The addition consists of the preliminary review of the properness of the challenging request. If the request is proper the challenge goes to the challenge chamber which determines whether the judge is partial and the judge should be replaced. If the request is not proper (obviously unfounded) a **penalty** is awarded to the requester.

The gains and losses of a challenge for both parties

- **Signal to the client**
  This is an advantage for the lawyer that is requesting the challenge; he will earn 2 points. These points will be awarded regardless of the challenge being

successful. However, these revenues are only awarded when the first judge of each period is challenged.

- **Delay of the lawsuit by disqualification**
  In practice, a successful (leading to a replacement judge) will result in justice being seriously delayed, because the complete trial has to be conducted again. This is often advantageous for one particular party, while it is at the expense of the position of the other party. In this experiment, such a delay will be in favor of Red and is disadvantageous for Blue. Therefore, in this experiment Blue will have to pay 10 points to Red when a judge is replaced by a successful challenge. These gains and losses from delaying the court case are awarded after each successful challenge. Thus also for a successful challenge of the second or later judges in the court case.
- **A fine in case of an improper request**
  The properness test is as follows. When in the draw of nine balls at least six balls of the color of the opponent party are present the request is proper. When less than five balls of the opponent party are present the request is improper. In the borderline case when exactly 5 balls have the opponent's color the request will considered proper with a probability of 50 %.
  An improper request will not be considered and the judge will remain on the trial. In addition, the submitter of the improper request must pay a penalty. The level of the penalty may change each period and will be announced prior to the start of the period. The penalty for improper challenging requests will apply to both the Blue and Red party and will also be imposed for second or later improper request of a period.

Similar to part 1, the judge will be replaced if he turns out to be partial. If the judge is impartial or if he is not properly challenged, the verdict will follow from that judge. Just like in part 1 the probability of a neutral judge, a judge that is partial in favor of Red or a judge in favor of Blue remains $1/3$.

The sequence of a period Below is a point-by-point breakdown of the sequence of a period:

1. Each period begins with a new judge and a new opponent (another participant than in the previous period) and your role (lawyer of Red or Blue) and the level of the fine is revealed.
2. The draw of the balls is visible to both parties.
3. Both lawyers decide whether to challenge or not:
   Improper requests are refused and fined. A request is proper with at least six balls of the color of the opponent, is improper with four or fewer balls of the color of your proponent and by exactly 5 balls of the color of your proponent the request is improper with a probability of 50 %.

   - Both lawyers do not challenge (or all requests are improper): the verdict follows (step 4).

- One or both lawyers request a proper challenge: **+2 points** for the requesting party(s), ***Attention***, this is only for the first judge.

  - The potential partiality of the judge is revealed:

    - Judge is neutral/impartial: the verdict follows (step 4).
    - Judge is partial: **+5 points** for Red and **−5 points** for Blue;

    A new judge is appointed and we go back to step 1.

4. The verdict: earnings of the period are made public. New period with a new judge and opponent and potentially also a new level of the fine (step 1).

We are now going to ask some questions to make sure that you have understood everything.

Practice questions part 2

Below we have some questions to be sure that you understand everything. The situations outlined are random and the choices are not necessarily wise! Please raise your hand if you do not understand something or have a question.

A.
You are the lawyer of party Blue. There are 1 red and 8 blue balls in the draw. You decide to request a challenge. Will that request be considered by the challenging chamber?

○ Yes
○ No

B.
You are the lawyer of party Red. There are 4 red and 5 blue balls in the draw. The penalty is 1 point. You decide to request a challenge and this will not be considered. The judge turns out to be partial for Blue. What are this period's total earnings for Red?

[ ] points

C.
You are the lawyer of party Red. There are 6 red and 3 blue balls in the draw and the penalty is 3 points. The lawyer of Blue decides to request a challenge, which will be considered. The judge turns out to be partial. With the new judge no request to challenge is put forward with a draw of 4 red and 5 blue balls. However, the judge turns out to be partial in favor of Red. What are this period's total earnings for party Blue?

[ ] points

Appendix 2

See Table 4.

Table 4 Challenges by risk attitude

Treatment	Risk att	Number of Blue signals									
		0–3		4		5		6–9		Total	
		Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
First judge red player											
Control	Loving	53.8 %	26	91.7 %	12	90.0 %	10	93.5 %	46	81.9 %	94
	Neutral	41.2 %	17	100.0 %	5	60.0 %	5	94.7 %	19	71.7 %	46
	Averse	57.0 %	79	77.3 %	22	100.0 %	35	96.2 %	79	80.5 %	215
	n/a	78.6 %	14	55.6 %	9	100.0 %	4	88.9 %	18	80.0 %	45
	Total	56.6 %	136	79.2 %	48	94.4 %	54	94.4 %	162	79.8 %	400
Fine = 0	Loving	24.4 %	45	23.5 %	17	100.0 %	12	94.6 %	37	55.9 %	111
	Neutral	17.4 %	23	20.0 %	5	75.0 %	4	100.0 %	15	48.9 %	47
	Averse	19.7 %	61	35.3 %	34	87.5 %	24	98.6 %	69	60.1 %	188
	n/a	25.0 %	16	25.0 %	4	100.0 %	9	100.0 %	17	67.4 %	46
	Total	21.4 %	145	30.0 %	60	91.8 %	49	97.8 %	138	58.4 %	392
Fine = 4	Loving	6.1 %	33	13.3 %	15	100.0 %	8	89.2 %	37	48.4 %	93
	Neutral	0.0 %	12	16.7 %	6	90.0 %	10	100.0 %	13	56.1 %	41
	Averse	6.0 %	84	10.0 %	30	69.0 %	29	95.6 %	90	48.9 %	233
	n/a	7.1 %	14	0.0 %	7	50.0 %	6	100.0 %	22	53.1 %	49
	Total	5.6 %	143	10.3 %	58	75.5 %	53	95.1 %	162	50.0 %	416
Fine = 12	Loving	0.0 %	32	0.0 %	10	66.7 %	6	94.1 %	34	43.9 %	82
	Neutral	0.0 %	15	0.0 %	7	42.9 %	7	100.0 %	15	40.9 %	44
	Averse	2.3 %	87	2.9 %	35	42.9 %	28	96.2 %	78	39.5 %	228
	n/a	7.7 %	13	0.0 %	4	83.3 %	6	100.0 %	15	55.3 %	38
	Total	2.0 %	147	1.8 %	56	51.1 %	47	96.5 %	142	42.1 %	392
First judge blue player											
Control	Loving	93.3 %	30	69.2 %	13	21.4 %	14	64.1 %	39	67.7 %	96
	Neutral	91.7 %	12	28.6 %	7	27.3 %	11	21.4 %	14	43.2 %	44
	Averse	87.5 %	80	45.5 %	22	25.0 %	24	34.8 %	89	54.4 %	215
	n/a	64.3 %	14	50.0 %	6	40.0 %	5	10.0 %	20	35.6 %	45
	Total	86.8 %	136	50.0 %	48	25.9 %	54	37.7 %	162	54.3 %	400
Fine = 0	Loving	92.3 %	39	72.2 %	18	9.1 %	11	5.9 %	34	51.0 %	102
	Neutral	100.0 %	15	40.0 %	5	16.7 %	6	25.0 %	12	55.3 %	38
	Averse	95.9 %	73	66.7 %	33	4.5 %	22	19.0 %	79	52.2 %	207
	n/a	72.2 %	18	25.0 %	4	10.0 %	10	23.1 %	13	40.0 %	45
	Total	92.4 %	145	63.3 %	60	8.2 %	49	16.7 %	138	50.8 %	392
Fine = 4	Loving	100.0 %	37	25.0 %	12	0.0 %	8	4.9 %	41	42.9 %	98
	Neutral	100.0 %	14	14.3 %	7	0.0 %	7	5.3 %	19	34.0 %	47
	Averse	100.0 %	73	26.7 %	30	0.0 %	33	3.5 %	85	38.0 %	221
	n/a	84.2 %	19	55.6 %	9	0.0 %	5	5.9 %	17	44.0 %	50
	Total	97.9 %	143	29.3 %	58	0.0 %	53	4.3 %	162	39.4 %	416

Table 4 continued

Treatment	Risk att	Number of Blue signals									
		0–3		4		5		6–9		Total	
		Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Fine = 12	Loving	95.0 %	40	12.5 %	8	0.0 %	9	3.7 %	27	47.6 %	84
	Neutral	91.3 %	23	8.3 %	12	0.0 %	7	0.0 %	11	41.5 %	53
	Averse	97.2 %	72	3.0 %	33	4.0 %	25	7.2 %	83	36.6 %	213
	n/a	83.3 %	12	33.3 %	3	0.0 %	6	4.8 %	21	28.6 %	42
	Total	94.6 %	147	7.1 %	56	2.1 %	47	5.6 %	142	38.8 %	392
Subsequent judges red player											
Control	Loving	3.8 %	52	57.1 %	14	92.3 %	13	97.8 %	46	53.6 %	125
	Neutral	9.5 %	21	83.3 %	6	80.0 %	10	100.0 %	17	59.3 %	54
	Averse	18.2 %	88	37.1 %	35	85.7 %	35	98.0 %	101	61.0 %	259
	n/a	23.5 %	17	37.5 %	8	80.0 %	5	94.4 %	18	58.3 %	48
	Total	13.5 %	178	46.0 %	63	85.7 %	63	97.8 %	182	58.6 %	486
Fine = 0	Loving	7.7 %	52	25.0 %	16	100.0 %	8	94.1 %	34	43.6 %	110
	Neutral	5.3 %	19	20.0 %	5	100.0 %	7	100.0 %	11	47.6 %	42
	Averse	5.1 %	78	27.6 %	29	92.3 %	26	95.9 %	73	51.5 %	206
	n/a	16.7 %	24	44.4 %	9	100.0 %	5	88.5 %	26	56.3 %	64
	Total	7.5 %	173	28.8 %	59	95.7 %	46	94.4 %	144	49.8 %	422
Fine = 4	Loving	0.0 %	41	0.0 %	15	94.4 %	18	97.8 %	45	51.3 %	119
	Neutral	0.0 %	9	0.0 %	4	100.0 %	1	100.0 %	12	50.0 %	26
	Averse	4.8 %	104	4.9 %	41	63.3 %	30	99.0 %	105	46.4 %	280
	n/a	5.3 %	19	0.0 %	6	100.0 %	2	90.5 %	21	45.8 %	48
	Total	3.5 %	173	3.0 %	66	76.5 %	51	97.8 %	183	47.8 %	473
Fine = 12	Loving	17.6 %	17	23.1 %	13	50.0 %	18	97.1 %	35	59.0 %	83
	Neutral	0.0 %	16	0.0 %	6	100.0 %	2	100.0 %	19	48.8 %	43
	Averse	2.9 %	105	0.0 %	33	30.3 %	33	95.9 %	97	39.6 %	268
	n/a	0.0 %	11	0.0 %	8	40.0 %	5	92.9 %	14	39.5 %	38
	Total	4.0 %	149	5.0 %	60	39.7 %	58	96.4 %	165	44.2 %	432
Subsequent judges blue player											
Control	Loving	80.8 %	52	0.0 %	9	0.0 %	13	3.9 %	51	35.2 %	125
	Neutral	100.0 %	20	16.7 %	6	16.7 %	6	5.9 %	17	46.9 %	49
	Averse	83.3 %	90	17.9 %	39	2.7 %	37	5.3 %	95	33.7 %	261
	n/a	75.0 %	16	22.2 %	9	14.3 %	7	10.5 %	19	33.3 %	51
	Total	83.7 %	178	15.9 %	63	4.8 %	63	5.5 %	182	35.4 %	486
Fine = 0	Loving	87.5 %	32	43.5 %	23	9.1 %	11	2.7 %	37	38.8 %	103
	Neutral	92.9 %	14	0.0 %	4	0.0 %	5	0.0 %	14	35.1 %	37
	Averse	94.4 %	108	46.4 %	28	4.2 %	24	1.3 %	78	49.2 %	238
	n/a	84.2 %	19	25.0 %	4	0.0 %	6	0.0 %	15	38.6 %	44
	Total	91.9 %	173	40.7 %	59	4.3 %	46	1.4 %	144	44.3 %	422

Table 4 continued

Treatment	Risk att	Number of Blue signals									
		0–3		4		5		6–9		Total	
		Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Fine = 4	Loving	97.5 %	40	30.0 %	20	6.3 %	16	2.1 %	47	38.2 %	123
	Neutral	100.0 %	20	16.7 %	6	0.0 %	6	0.0 %	23	38.2 %	55
	Averse	97.8 %	93	29.4 %	34	0.0 %	23	0.0 %	93	41.6 %	243
	n/a	90.0 %	20	33.3 %	6	33.3 %	6	5.0 %	20	44.2 %	52
	Total	97.1 %	173	28.8 %	66	5.9 %	51	1.1 %	183	40.6 %	473
Fine = 12	Loving	96.4 %	28	12.5 %	16	0.0 %	16	0.0 %	34	30.9 %	94
	Neutral	87.0 %	23	33.3 %	6	0.0 %	4	0.0 %	22	40.0 %	55
	Averse	92.0 %	88	3.6 %	28	0.0 %	32	0.0 %	88	34.7 %	236
	n/a	100.0 %	10	10.0 %	10	33.3 %	6	0.0 %	21	27.7 %	47
	Total	92.6 %	149	10.0 %	60	3.4 %	58	0.0 %	165	33.8 %	432

References

Bault, N., Coricelli, G., & Rustichini, A. (2008). Interdependent utilities: How social ranking affects choice behavior. *PLoS ONE, 3*(10), e3477.

Bauw, E. (2011). Wat te doen met wraking? *Ars Aequi, 60*(3), 204–208.

Bryden, P., & Hughes, J. (2011). The Tip of the Iceberg: A Survey of the Philosophy and Practice of Canadian Provincial and Territorial Judges Concerning Judicial Disqualification. *Alberta Law Review, 48*, 3.

Buhai, S. L. (2011). Federal judicial disqualification: A behavioral and quantitative analysis. *Oregon Law Review, 90*, 69.

ENCJ (2013). *Judicial Reform in Europe part II, Guidelines for effective justice delivery*. www.encj.eu.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, 817–868.

Fréchette, G. R., & Schotter, A. (Eds.). (2015). *Handbook of experimental economic methodology*. Oxford: Oxford University Press.

Giesen, I., Kristen, F., Enneking, L., de Kezel, E., van Lend, L., & Willemsen, P. (2012). *De wrakingsprocedure. Een rechtsvergelijkend onderzoek naar de mogelijkheden tot herziening van de Nederlandse wrakingsprocedure.* Memorandum 5-2012. Raad voor de rechtspraak. www.rechtspraak.nl.

Gneezy, U., & Rustichini, A. (2000). Fine is a price. *Journal of Legal Studies, 29*, 1.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*(5), 1644–1655.

Raad voor de rechtspraak, Jaarverslag 2014, tables 31 and 32.

van Rossum, W. L., Tighelaar, J., & Ippel, P. (2012). *Wraking bottom-up.* Research Memorandum 6-2012. Raad voor de rechtspraak. www.rechtspraak.nl.