Proof Beyond a Reasonable Doubt: Laboratory Evidence

Florian Baumann[1] Tim Friehe[2]

June 2015

Abstract

We investigate how third-party punishers and potential violators decide under evidentiary uncertainty in a take game. In line with the legal requirement and in contrast to economic models, neither the sanction nor the harm level affects the punishment probability, but the quality of evidence does have an impact. In our experimental setup, potential violators' decisions are strongly influenced by the expected punishment probability but not by the level of the sanction.

JEL-Codes: K42, D81, C91

Keywords: crime; experiment; reasonable doubt; standard of proof; third-party punishment

[1] Duesseldorf Institute for Competition Economics, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany. fbaumann@dice.hhu.de. +49 (0)221 8114318.

[2] Corresponding author: University of Marburg, Am Plan 2, 35037 Marburg, Germany. tim.friehe@uni-marburg.de. +49 (0) 6421 28 21701.

1. Introduction

When deciding whether to convict a criminal suspect, legal decision-makers ask for *proof beyond a reasonable doubt*.[3] The famous ratio first formulated by Blackstone (1769)—better that ten guilty persons escape, than that one innocent suffer—is often referred to in this context (see, e.g., Epps 2015). This standard of proof does not vary with the aspects of the case; instead, it is solely concerned with the quality of the evidence. For example, jury instructions in New York State refer to the absence of any reasonable doubt, concerning the existence of any element of the crime or the defendant's identity as the person who committed the crime (NY Courts 2015), without asking jury members to see the quality of the evidence in light of the circumstances of the case including the severity of the criminal act or the punishment applied in the event of a conviction. However, from an economic point of view, the evidence required by a legal decision-maker for convicting a criminal defendant might well be tailored to the case at hand. Andreoni (1991) provides a theory in which decision-makers choose between convicting a suspect or not under evidentiary uncertainty about the suspect's guilt, knowing the fixed sanction in the event of a conviction and the harm level of the criminal act. He establishes that the standard of proof is increasing in the sanction and decreasing in the harm level. This results because a higher sanction increases the decision-maker's intrinsic costs of wrongful convictions (i.e., a legal error of type I), while higher harm makes wrongful acquittals more costly (i.e., a legal error of type II). As a result of the implied change in the standard of proof, the probability of wrongful convictions should decrease in the sanction and increase in the harm level, whereas the probability of wrongful acquittals should increase in the sanction and decrease in the harm level. Moreover, the implied change in the standard of proof influences the level of deterrence. Andreoni's contribution is the impetus for our experimental investigation. We explore the interaction between evidentiary uncertainty and the sanction level for acts of varying severity regarding the standard of proof required by decision makers. In doing so, we respond to Miceli (2009) who notes the lack of evidence on how (legal) decision-makers actually interpret *reasonable doubt*, i.e., more closely aligned with either the legal

[3] The "beyond a reasonable doubt" standard is a constitutional requirement (see, e.g., Pollock 2012), asserted by the Supreme Court, for example in the case *In re Winship* (In re Winship, 397 U.S. 358, 90 S. Ct. 1068, 25 L. Ed. 2d 368 (1970)).

understanding (as exemplified by the jury instructions above) or the economic one (as formalized by Andreoni, for instance).

Analyzing data from a laboratory experiment, this paper presents results on how third-party punishers deal with evidentiary uncertainty in a take game when the aspects of the case vary. We consider two sanction and two harm levels in a 2x2 between-subject design. We are interested in the implications of the aspects of the case for the likelihood of legal errors, the individual punishment decision, and the frequency of takings. By relying on experimental data, we can control other influences apart from that of the sanction and the harm level and track judicial decisions as actual legal errors.

All adjudicative procedures must deal with the possibility of errors, making the research at hand practically relevant to a large number of settings. It may be expected that the issue of interpreting a standard of proof gets particularly important when laymen are assigned the task of deciding cases as is true, for example, in criminal jury cases in the USA. Our focus on the decision of a third party, whether or not to impose a given sanction, is all the more practically relevant in settings in which the court's discretion, with regard to the level of the sanction, is seriously restricted as is the case for judicial decision-makers under sentencing guidelines in the USA (e.g., Miceli 2008, Schanzenbach and Tiller 2007). Moreover, standards of proof are set in other contexts as well, for example, when standards are set for promotion and employee discipline in the internal organization of a firm or when defect rates are determined for rejection of a supplier's shipment (Kaplow 2011).

The theoretical literature on standards of proof is quite extensive (see, e.g., Davis 1994, Friedman and Wickelgren 2006, Kaplow 2011, Lando 2009, Miceli 1990, Mungan 2011, Ognedal 2005, Rizzollo and Saraceno 2013, Rubinfeld and Sappington 1987, Yilankaya 2003). In some treatises (e.g., Rubinfeld and Sappington 1989), the standard of proof and the penalty are both considered as the court's policy instruments. The aspect that we revisit using data from the laboratory—namely, how decisions regarding whether to convict or not are influenced by the level of the sanction and the level of harm in the presence of evidentiary uncertainty—is theoretically analyzed in Andreoni (1991), as described before, and the closely related contribution by Feess and Wohlschlegel (2009). The latter establish,

inter alia, that the deterrence-maximizing sanction might be increasing with the quality of the legal system (quality of information).[4]

Empirically, the effect of evidentiary uncertainty has been studied in voluntary contribution mechanism (VCM) experiments (Ambrus and Greiner 2012, Grechenig et al. 2010). In contrast to our setup, individuals who decide on punishment usually take part in the VCM as well. For example, Grechenig et al. (2010) find that people punish extensively despite evidentiary uncertainty and that a deterioration in evidentiary quality sometimes even increases punishment. This is at odds with our results for third-party punishment. Obviously, third-party punishment is more relevant to our interests when compared to peer punishment. Rizzolli and Stanca (2012) consider a take game and study how exogenously imposed legal errors influence deterrence (i.e., focus on how errors influence potential offenders and disregard the source and the severity of legal errors). Sonnemans and van Dijk (2012) are interested in judicial decisions when there is a possibility of legal error. In their design, subjects were asked to identify the correct decision in 30 abstract cases, being given the prior probability of guilt of one half and the outcome of an investigation (which is eitherincriminating, exonerating or neutral). Sonnemans and van Dijk (2012) are interested in how subjects process the information available and how they arrive at verdicts, finding that indeed a great number of people use the information provided rationally but may not acquire the optimal amount of information. Our analysis is complementary in that we consider how the standard of proof required for conviction is influenced by aspects of the case at hand, namely the level of harm and the level of the sanction. There are also a number of key differences in the experimental design because judicial decision-makers in our setting decide without preset priors and with the knowledge that their verdict has payoff consequences for other experimental subjects. Van Dijk et al. (2014) consider the distinction between individual decision-makers and groups when it comes to judicial error, finding that group decisions tend to reduce the occurrence of legal errors.

Most closely related to our research question about how evidentiary uncertainty interacts with the sanction and harm level regarding the standard of proof is Feess et al. (2014) who

[4] See Lando (2005) for a related contribution about the size of the sanction and the quality of the evidence.

conduct a study designed and implemented independently from ours. In fact, Feess et al. and the present work are the only contributions in which the risk of legal errors depends endogenously on the behavior of potential violators—practically speaking, the most relevant scenario. Our paper is complementary to Feess et al. due to important differences in experimental design (to be explained at the end of Section 2.1).

The remainder of the paper is organized as follows. Section 2 presents the design of our experiment and the data collected. Section 3 portrays our results from analyzing the experimental data. Section 4 concludes.

2. **Experimental design and data**

2.1 Experimental design

In our experiment, subjects were endowed with 20 points each (1 point = 40 Euro cents), and three subjects, assigned to different roles (A, B, and C), formed a group.[5] In brief, our main experiment included the following three stages (which are explained in detail below): In stage 1, player A had to decide whether or not to take points from player B in order to increase his/her payoff. In stage 2, player C received a noisy signal about player A's choice and had to decide whether or not to deduct points from player A; this punishment had no influence on player C's material payoffs. Player B remained passive in stages 1 and 2. In stage 3, beliefs were elicited from all participants about both the taking and punishment probability. After the main experiment, participants took part in a risk elicitation procedure (Holt and Laury 2002) and filled out a questionnaire.

We considered two treatment dimensions in our 2x2 design: the harm *h* created by player A's taking of points from player B may be either high or low, $h \in \{10; 20\}$; the punishment *s* player C may impose on player A may be either high or low, $s \in \{10; 20\}$.

[5] The instructions for the participants may be found in the Supplementary Material.

Next, we describe the three stages of our main experiment in more detail. In stage 1, player A could take *h* points from player B to receive 5 points.[6] The level of *h* is meant to represent the external cost of player A's taking which may be either high (*h=20*) or low (*h=10*). However, the level of harm always exceeds player A's private gain by a wide margin, clearly conveying that taking should be seen as a norm infraction. Player A knows that the choice between taking and not taking points from player B influences the signal received by player C in stage 2. More precisely, in stage 2, player C learned the color of a ball which was randomly drawn out of urn BLACK when player A took points from player B and out of urn WHITE otherwise. The probability that the draw was a black ball was weakly greater when player A actually took points from player B in comparison to when player A did not take points. We distinguished six different urn compositions as displayed in Table 1, thereby varying the informativeness of the signal received by player C. Urn composition (1) represents the case of a signal that allows to infer player A's choice without error (a black ball implies that player A took points from player B and a white ball established player A's innocence). The signal's precision is decreasing in the number of the urn composition. The signal is still informative for urn compositions (2) to (5). Urn composition (6) represents the case of a completely uninformative signal. Knowing these facts, each player A indicated his/her decision about taking for each of the possible urn combinations (1) to (6) where we made use of the strategy method (Selten 1967). Because we are interested in decision-making when there is an informative signal, we did not include urn composition (6) in our data analysis; however, it was included in the instructions to make them easier to understand.

Table 1: BLACK and WHITE urn compositions

Composition	Urn BLACK	Urn WHITE
(1)	10 black & 0 white balls	0 black & 10 white balls
(2)	9 black & 1 white balls	1 black & 9 white balls
(3)	8 black & 2 white balls	2 black & 8 white balls
(4)	7 black & 3 white balls	3 black & 7 white balls
(5)	6 black & 4 white balls	4 black & 6 white balls
(6)	5 black & 5 white balls	5 black & 5 white balls

[6] We rely on the take game to bring norm violations into the lab (as in, e.g., Falk and Fischbacher 2002, Rizzolli and Stanca 2012, Schildberg-Hörisch and Strassmair 2012).

In stage 2, player C could punish player A at no personal material cost by subtracting *s* points from player A's payoff. The level of the sanction may either be low (*s = 10*) or high (*s = 20*). However, the level of the sanction always exceeds player A's private gain from taking (that is fixed at 5 points). Player C had to decide whether to punish or not based on the information inferred from the signal about player A's choice (that is, the color of the ball drawn) and own priors. Player C chose between no punishment and punishment contingent on the color of the ball for all six urn compositions, implying a total of 12 decisions; one decision for every possible combination of color of the ball (black or white) and urn composition, where we again made use of the strategy method.

In stage 3, subjects stated their (incentivized) beliefs regarding how many players A outside of their group took points and how many players C assigned punishment for each contingency. For each correct belief, subjects earned 4 points. In other words, we do not present priors as Sonneman and van Dijk (2012), for example, as it is impossible in our design given that infractions are determined by players A themselves. Nevertheless, by adding the third stage, we have information about the subjects' priors which is critical for understanding their decision-making. As stated above, Players B do not make decisions in stages 1 and 2 but state their beliefs in stage 3.

After having explained our experimental design in detail, we now briefly relate it to the design used by Feess et al. (2014). A key difference is that Feess et al. consider within-subject variation with regard to the level of the sanction, which may draw subjects' attention to this key variable, and do not incorporate variation with respect to the external costs of the act.[7] Importantly, we elicit beliefs about the *a priori* violation probability and about the expected punishment probability, allowing us to include these variables in our empirical analysis. In addition, the violation considered by Feess et al. consisted of reducing a donation to a charity, and they allowed all subjects to participate as both potential violators and judges. In our design, subjects participated in only one role in order to limit empathy effects (i.e., third-party punishers being lenient on takers). Indeed, the low level of

[7] Feess et al. (2014) incorporate three different sanction levels but do not consider a variation in the harm level due to the act (in contrast to our 2x2 setup).

punishment in Feess et al. (2014) is consistent with this possibility. We will relate our findings to their results below.

2.2 Implementation and data

In total, 504 subjects (mean age 24, 56 % female) participated in our experiment at the University of Bonn which was computerized using zTree (Fischbacher 2007). ORSEE (Greiner 2003) was used for recruitment. Data was collected from June to October 2014. The 504 subjects formed 168 groups. We analyzed data from 55 groups for treatment (*s=10*, *h=10*), 51 for the case in which (*s=20*, *h=10*), 32 for (*s=10*, *h=20*), and 30 for treatment (*s=20*, *h=20*).

The experiment was a stand-alone one (i.e., not preceded by any other experiment) and concluded by a long questionnaire. Payoffs were assessed as follows: first, one urn composition was drawn randomly; next, player A's decision for that composition was implemented which determined whether urn BLACK or urn WHITE was relevant; next, the color of the ball was determined by a draw of nature according to the composition of the relevant urn (either BLACK or WHITE) and the decision of player C (either punish or not punish) was implemented. In addition, subjects received payments for correct beliefs in stage 3 and the subsequent risk elicitation procedure. Answering the questionnaire was rewarded with an additional 20 points. On average, participants earned €18.05 in the experiment, which—due to the long questionnaire after the experiment—lasted in total an average of 80 minutes.

Table 2 presents the summary statistics of the data we use in our analysis. Concentrating on urn compositions (1) to (5), we observe 5 taking decisions by players A (which gives a total of 168 x 5 = 840) and 10 punishment decisions by players C (5 for a black ball, 5 for a white one).[8]

[8] In our analysis, we do not use the information on players B which is therefore not reported in Table 2.

Table 2: Summary of data

Variable	N	Mean	Standard deviation	Min.	Max.
Choice variables					
Taking	840	0.412	0.492	0	1
Punishment black	840	0.648	0.478	0	1
Punishment white	840	0.177	0.382	0	1
Treatment variables and signal precision					
Fine high	840	0.482	0.500	0	1
Harm high	840	0.369	0.483	0	1
High precision	840	0.400	0.490	0	1
Beliefs					
Belief punish black	840	0.720	0.280	0	1
Belief punish white	840	0.269	0.278	0	1
Belief taking	840	0.483	0.359	0	1
Personal characteristics					
Risk aversion	336	6.173	1.719	0	10
Male	336	0.455	0.499	0	1
Probabilities of legal error					
Probability legal error of type II	346	0.474	0.434	0	1
Probability legal error of type I	494	0.223	0.330	0	1

Players A took points from players B in 41.2 % of the cases. Players C punished player A in 64.8 % of the cases in which a black ball was drawn and in 17.7 % of cases in which a white ball was drawn. *Fine high* (*Harm high*) is a treatment dummy variable which takes the value of one if $s = 20$ ($h = 20$). We aggregate the information on the urn compositions in the dummy variable *High precision* which takes the value one when the urn composition is either (1) or (2) and zero otherwise. Beliefs of players C (players A) about the behavior of player A (player C) are taken from the answers to the question of "How many of the players A (C) in this session but not in your group decided to take points (punish in the event of a black / white ball)" which individuals answered in stage 3 of the experiment. As expected the belief about the probability of punishment indicates a higher average value for a black than a white ball. Players C expect an average taking rate of 48.3 %. The measure of *Risk aversion* results from an elicitation procedure as in Holt and Laury (2002) with higher values

indicating a higher degree of risk aversion (i.e. *Risk aversion* is the number of risk-averse choices in the Holt and Laury procedure). *Male* is a dummy variable indicating male subjects Finally, we calculated values for the probabilities of legal errors given the decisions taken by players A and C. Players A decided to take points from player B in 346 cases. The probability that player A remained unpunished in these cases (legal error of type II) is 47.4 %.[9] Players A decided against taking in the remaining 494 cases. For these cases, the probability that player A nevertheless received punishment (legal error of type I) is on average 22.3 %.[10]

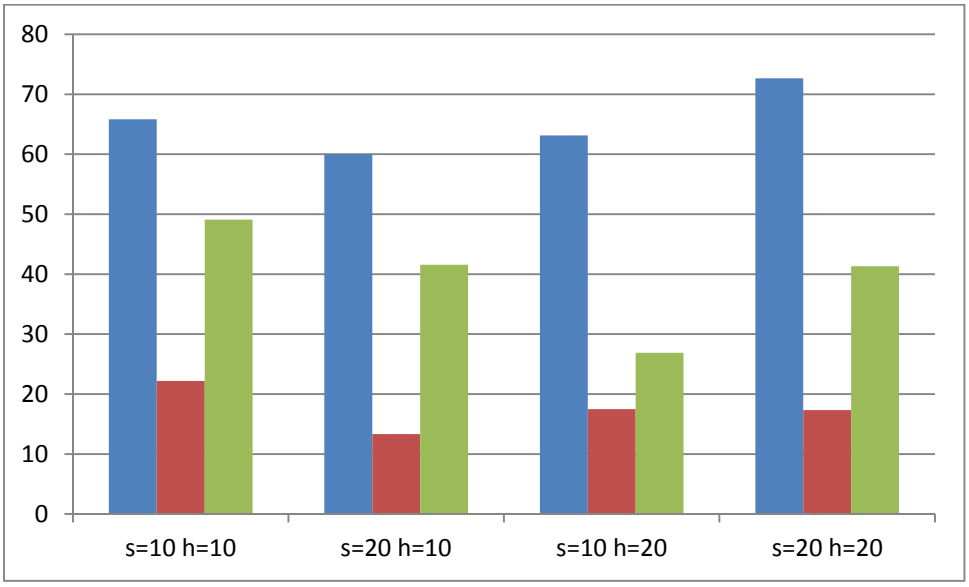
3. Results

The share of players C who punished players A in the four different treatments is shown in Figure 1. The averages indicated by the blue bars for the case that a black ball was drawn and the red bars for the case that a white ball was drawn suggest that the share of players C who punish decreases when the level of the sanction is raised and the level of harm is fixed at 10, whereas it seems to be rather the other way around when the level of harm is 20. When we hold the level of the sanction fixed at 20, the average share of players C who punish seems to increase with the level of harm. The share of taking (green bar) was the highest in treatment (*s=10, h=10*), very similar in (*s=20, h=10*) and (*s=20, h=20*), and the lowest in (*s=10, h=20*).

[9] If player A decided to take points from player B, the probability of a legal error of type II is calculated as (1- Punish black) x Number of black balls in urn BLACK/10 + (1- Punish white) x Number of white balls in urn BLACK/10, where Punish black and Punish white describe the decisions by player C and each urn contains 10 balls. Likewise, if player A decided against taking points, the probability of a legal error of type I results as Punish black x Number of black balls in urn WHITE/10 + Punish white x Number of white balls in urn WHITE /10.

[10] In our data, 24 of the 168 players C decided never to punish player A. In the subsequent questionnaire, several of them stated that their reluctance to punish was due to the fact that punishment had no effect on their own payoffs or that ex post punishment could not alter player A's choice.

Figure 1: Average share of punishers when the ball was black (blue bar) or white (red bar); average share of takers (green bar)



Both punishment and taking decisions are endogenous in our experimental design and together imply a likelihood of legal error which depends on the treatment as displayed in Figure 2.

Figure 2: Probability of legal errors of type II (blue bar) and type I (red bar)

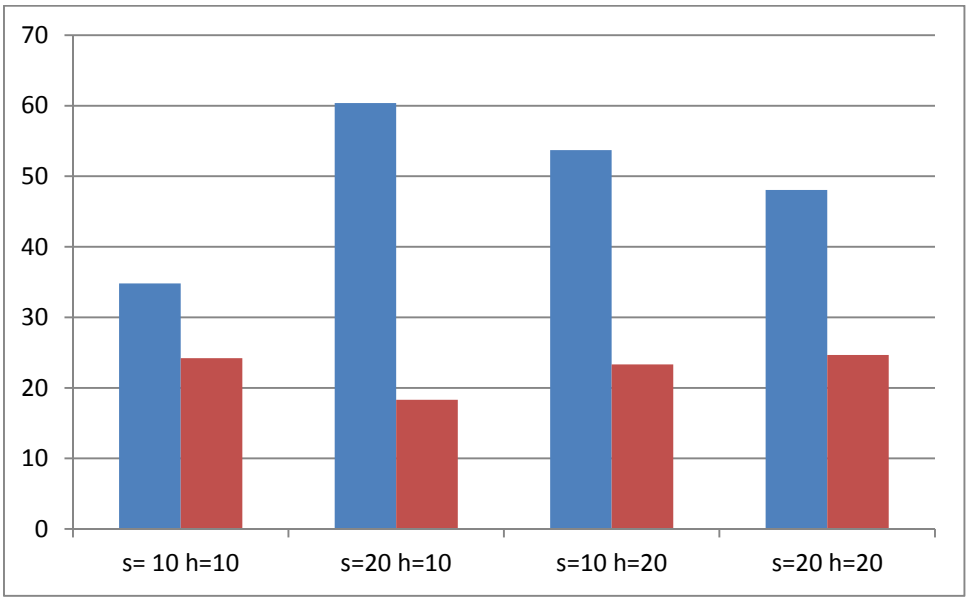


Figure 2 shows that the probability of legal error varies strongly across treatments and that the two errors respond differently to the comparative statics regarding sanctions and the harm level. For example, moving from the scenario (*s=10*, *h=10*) to (*s=20*, *h=10*) brings about

a strong increase in the probability of wrongful acquittals and a decrease in the probability of wrongful convictions. This is in line with the predictions of the model by Andreoni (1991) where the differences are due to higher (intrinsic) costs of a type I error for the legal decision-maker when the exogenous sanction is higher. However, a similar result cannot be found for the high harm level. All in all, Figure 2 shows that the probability for a wrongful acquittal is, in most treatments, higher than the probability for a wrongful conviction. This differs from the finding by Sonnemans and van Dijk (2012) that errors are biased toward wrongful convictions.

To understand what influences the probability of error, we turn to a regression analysis (Table 3). For the low-harm baseline, the probability of an error of type II (i.e., a wrongful acquittal) is significantly higher when the sanction is high, confirming the impression from the descriptive analysis above and in line with Andreoni (1991). However, also paralleling the descriptive analysis, this impact of a higher sanction is not present when the level of harm is high. The probability of an error of type I (i.e., a wrongful conviction) is not significantly affected by the aspects of the case. Intuitively, the probability of error is decreasing in the precision of the signal.

Table 3: Tobit regressions for error probabilities of types I and II

	Probability Type II Error	Probability Type I Error
Fine high	0.564**	-0.054
	(2.43)	(-0.44)
Harm high	0.419	0.005
	(1.40)	(0.03)
Fine high * Harm high	-0.685*	0.132
	(-1.67)	(0.72)
High precision	-0.763***	-0.409***
	(-6.93)	(-6.92)
Constant	0.491***	0.222**
	(3.53)	(2.23)
N	346	494
F-stat.	16.67	12.63

Notes: The dependent variable is defined as the probability that taking points will not be punished (error type II) or as the probability that a player A who did not take points will be punished (error type I). *Fine high* and *Harm high* are dummy variables. High precision is a dummy variable that is equal to one when the urn composition is either (1) or (2). t-values in parentheses; standard errors clustered at the group-level; *** 0.01, ** 0.05, * 0.10 significance level.

Next, we explore the standard of proof used by third parties. Since the level of punishment is exogenously given, third parties only choose whether or not to punish. Table 4 shows that the aspects of the case do not significantly influence player C's probability of punishment (regressions are run separately for the signal of a black and a white ball). This is aligned with the legal concept of *proof beyond a reasonable doubt* that emphasizes solely evidentiary quality. Importantly, individual punishment decisions clearly exhibit the *in dubio pro reo* principle. A high-precision signal significantly increases (decreases) the likelihood of punishment for a black (white) ball. Players C in our experiment are thus primarily concerned about the quality of the evidence, and less about the circumstances of the case. Intuitively, players C who believe that many players A take points are more likely to punish even when they observe a white ball. The measure of player C's moral expectations of others cannot explain punishment decisions. We can compare our results to those put forward by Feess et al. (2014) who consider three sanction levels in their within-subject analysis. In line with the findings in our study, they find no significant difference between the punishment probabilities for the low and the medium sanction level but indicate that their high sanction level leads to a significant decrease in the punishment probability.

Table 4: Probit estimates of punishment

	Punish when ball is black			Punish when ball is white		
	All	Only Damage Low	Only Damage High	All	Only Damage Low	Only Damage High
Fine high	0.002	-0.058	0.110	-0.045	-0.081	0.032
	(0.03)	(-0.79)	(1.38)	(-1.08)	(-1.58)	(0.50)
Harm high	0.055			-0.002		
	(0.97)			(-0.06)		
High precision	0.222***	0.207***	0.247***	-0.124***	-0.070**	-0.224***
	(8.24)	(6.43)	(5.25)	(-4.60)	(-2.15)	(-4.76)
Belief taking	0.123	-0.000	0.064	0.144**	0.184**	0.131
	(0.17)	(-0.00)	(0.57)	(2.44)	(2.47)	(1.47)
Risk aversion	-0.014	-0.004	-0.032	0.004	-0.016	0.048*
	(-0.85)	(-0.19)	(-1.21)	(0.30)	(-1.33)	(1.91)
Male	-0.034	0.006	-0.072	-0.055	-0.119**	-0.03
	(-0.61)	(0.07)	(-0.89)	(-1.26)	(-2.23)	(-0.04)
N	840	530	310	840	530	310

Notes: The dependent variable is equal to one when player C choses punishment for the given color of the ball. *Fine high* and *Harm high* are dummy variables. High precision is a dummy variable that is equal to one when the urn composition is either (1) or (2). *Belief taking* is the share of players A who take according to the beliefs of player C. Marginal effects; z-values in parentheses; standard errors clustered at the individual-level; *** 0.01, ** 0.05, * 0.10 significance level.

Another key question is how potential violators behave under evidentiary uncertainty. This is studied in Rizzolli and Stanca (2012) for exogenously specified error probabilities. Column 1 of Table 5 shows for our setup that higher evidentiary quality deters taking. In other words, evidentiary uncertainty that results in a high probability of legal errors (see Table 3) deteriorates deterrence (as argued in theoretical contributions such as Polinsky and Shavell 2007). Also, there is more taking when players A expect to receive punishment even when a white ball is drawn. One key issue of the literature on law enforcement is the deterrent effect of punishment. Interestingly, we find that a higher level of the sanction does not produce additional deterrence. This concurs with recent empirical evidence for Germany (Entorf and Spengler 2015) but stands in sharp contrast to experimental findings by Friesen (2012) who used fixed detection probabilities. In Feess et al. (2014) taking rates decrease only when the sanction is raised to its highest level. In the first column of Table 5 we find a (weakly) significant negative effect of a high harm level on the taking probability which may

indicate moral concerns. Distinguishing the cases with a low and a high harm level in columns 2 and 3 of Table 5, we find that a higher sanction level even increases the probability of taking for a high harm level, i.e., possibly legitimizes taking (the p-value is 0.055). One interpretation is that in such cases, law enforcement is sufficiently strict to crowd out moral concerns (as in Schildberg-Hörisch and Strassmair 2012). Players A may feel that taking points from player B is inadequate behavior when the level of social harm therefrom is high and the sanction is low. Possibly, such moral considerations are crowded out when the level of the sanction is more in line with harm.[11]

Table 5: Probit estimates of taking

	All	Only Damage Low	Only Damage High
Fine high	0.027 (0.49)	-0.067 (-0.97)	0.172* (1.92)
Harm high	-0.098* (-1.68)		
High precision	-0.109*** (-3.16)	-0.087** (-2.04)	-0.121** (-2.04)
Belief punish black	-0.197* (-1.92)	-0.329*** (-2.75)	-0.041 (-0.22)
Belief punish white	0.328*** (3.49)	0.362*** (3.10)	0.286** (2.02)
Risk aversion	-0.029 (-1.51)	-0.018 (-0.77)	-0.036 (-1.29)
Male	0.006 (0.11)	0.003 (0.04)	0.026 (0.26)
N	840	530	310

Notes: The dependent variable is equal to one when player A choses to take points. *Fine high* and *Harm high* are dummy variables. High precision is a dummy variable that is equal to one when the urn composition is either (1) or (2). *Belief punish black/white* is the share of players C who punish when the ball color is black/white according to the beliefs of player A. Marginal effects; z-values in parentheses; standard errors clustered at the individual-level; *** 0.01, ** 0.05, * 0.10 significance level.

[11] Using logit instead of probit does not significantly affect any of the results in Tables 4 and 5.

4. Discussion and conclusion

This paper considers potential offenders' and third-party punishers' response to evidentiary uncertainty and how it is moderated by the facts of the case, namely the level of the sanction and the level of harm. We rely on data from a laboratory experiment with a 2x2 between-subject design. In our experiment, third-party punishers respond strongly to the quality of evidence, confirming *in dubio pro reo*. This is reassuring and contrasts with some results obtained for peer punishment. The standard of proof used by third parties is significantly related to neither the level of the sanction nor that of harm as required by the legal understanding of *proof beyond a reasonable doubt*. However, our empirical analysis establishes that the probability of a wrongful acquittal is increasing with the sanction level when the level of harm is low. Potential violators also strongly respond to the quality of evidence. Moreover, we find that the probability of taking decreases with the expected punishment probability and a proxy of the individual's moral concerns but not with the level of the fine.

The standard of proof actually applied by individuals who (have to) take the position of judicial decision-makers is important in a number of settings. Understanding how it is derived and how it changes with the circumstances is thus of great practical importance. The present study makes a contribution to this field of inquiry but has its limitations. Our participants are students (including law students), allowing for the possibility that professionals decide differently (although the results by Sonnemans and van Dijk (2012) do not point in this direction). Moreover, we consider a 2x2 design, allowing for the possibility that different behavior would result at other sanction levels, for example. In summary, our research presents a first piece of the puzzle and hopefully motivates further research.

Acknowledgements

We gratefully acknowledge the very helpful comments received in different stages of the project by James Andreoni, Eberhard Feess, Louis Kaplow, Michael Kurschilgen, and Hannah Schildberg-Hörisch. Lisa Bruttel contributed greatly, in particular, to the design stage of the experiment.

References

Ambrus, A., and B. Greiner, 2012. Imperfect public monitoring with costly punishment – An experimental study. American Economic Review 102, 3317-3332.

Andreoni, J., 1991. Reasonable doubt and the optimal magnitude of fines: Should the penalty fit the crime? Rand Journal of Economics 22, 385-395.

Blackstone, W., 1769. Commentaries on the laws of England. Vol. 4. Oxford: Clarendon Press.

Davis, M.L., 1994. The value of truth and the optimal standard of proof in legal disputes. Journal of Law, Economics, and Organization 10, 343-359.

Entorf, H., and H. Spengler, 2015. Crime, prosecutors, and the certainty of conviction. European Journal of Law and Economics 39, 167-201.

Epps, D., 2015. The consequences of error in criminal justice. Harvard Law Review 128, 1065-1151.

Falk, A. and U. Fischbacher, 2002. "Crime" in the lab – detecting social interaction. European Economic Review 46, 859-869.

Feess, E., Schramm, M., and A. Wohlschlegel, 2014. The impact of fine size and uncertainty on punishment and deterrence: Evidence from the laboratory. Available at SSRN: http://ssrn.com/abstract=2464937.

Feess, E., and A. Wohlschlegel, 2009. Why higher punishment may reduce deterrence. Economics Letters 104, 69-71.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171-178.

Friedman, E., and A. Wickelgren, 2006. Bayesian juries and the limits to deterrence. Journal of Law, Economics, and Organization 22, 70-86.

Friesen, L., 2012. Certainty of punishment versus severity of punishment: An experimental investigation. Southern Economic Journal 79, 399-421.

Grechenig, K., Nicklisch, A., and C. Thöni, 2010. Punishment despite reasonable doubt – a public goods experiment with sanctions under uncertainty. Journal of Empirical Legal Studies 7, 847-867.

Greiner, B., 2003. An online recruitment system for economic experiments, in: Kremer, K., and Macho, V. (Eds.), Forschung und wissenschaftliches Rechnen. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung, 79-93.

Holt, C.A., and S.K. Laury, 2002. Risk aversion and incentive effects. American Economic Review 92, 1644-1655.

Kaplow, L., 2011. On the optimal burden of proof. Journal of Political Economy 119, 1104-1140.

Lando, H., 2005. The size of the sanction should depend on the weight of the evidence. Review of Law & Economics 2, 277-292.

Lando, H., 2009. Prevention of crime and the optimal standard of proof. Review of Law & Economics 5, 33-52.

Miceli, T., 1990. Optimal prosecution of defendants whose guilt is uncertain. Journal of Law, Economics, and Organization 6, 189-201.

Miceli, T., 2008. Criminal sentencing guidelines and judicial discretion. Contemporary Economic Policy 26, 207-215.

Miceli, T., 2009. Criminal procedure. In: Garoupa, N. (Ed.). Criminal law and economics. Edward Elgar.

Mungan, M., 2011. A utilitarian justification for heightened standards of proof in criminal trials. Journal of Institutional and Theoretical Economics 167, 352-370.

NY Courts, 2015. Reasonable Doubt – CJI2d Instructions of General Applicability. (Downloaded April 15 2015 from http://www.nycourts.gov/judges/cji/1-General/CJI2d.Presumption.Burden.Reasonable_Doubt.pdf).

Ognedal, T. (2005). Should the standard of proof be lowered to reduce crime? International Review of Law and Economics 25, 45–61.

Polinsky, A.M., and S. Shavell, 2007. The theory of public enforcement of law. In: Polinsky, A.M., Shavell, S. (eds.), Handbook of Law and Economics 1. Elsevier, Amsterdam.

Pollock, J.M., 2012. Criminal law. Routledge.

Rizzolli, M., and M. Saraceno, 2013. Better that ten guilty persons escape: punishment costs explain the standard of evidence. Public Choice 155, 395-411.

Rizzolli, M., and L. Stanca, 2012. Judicial errors and crime deterrence: Theory and experimental evidence. Journal of Law and Economics 55, 311-338.

Schanzenbach, M., and E.H. Tiller, 2007. Strategic judging under the US Sentencing Guidelines: Positive political theory and evidence. Journal of Law, Economics, and Organization 23, 24-56.

Schildberg-Hörisch, H., Strassmair, C., 2012. An experimental test of the deterrence hypothesis. Journal of Law, Economics, and Organization 28, 447-459.

Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Sauermann, H. (ed.) Beiträge zur experimentellen Wirtschaftsforschung, 136–168. Tübingen: Mohr.

Sonnemans, J., and F. van Dijk, 2012. Errors in judicial decisions: Experimental results. Journal of Law, Economics, and Organization 28, 687-716.

Van Dijk, F., Sonnemans, J., and E. Bauw, 2014. Judicial error by groups and individuals. Journal of Economic Behavior and Organization 108, 224-235.

Yilankaya, O., 2003. A model of evidence production and optimal standard of proof and penalty in criminal trials. Canadian Journal of Economics 35, 385-409.

Supplementary Material

Translated version of the instructions for the treatment harm low and sanction low (h=10 and s=10)

General explanations

Welcome to this economic experiment.

In the following pages, we explain how you can earn money from your decisions in this experiment. Please read the instructions carefully. If you have any questions, please raise your hand and we will come to your seat.

During the experiment, you are not allowed to talk to the other participants, use cell phones, or start any programs on the computer. Disregarding any of these rules will lead to your exclusion from the experiment and from all payments.

During the experiment, your gains and losses are counted in points instead of in Euros. Your total income will be calculated in points first. At the end of the experiment, your total points will be converted into Euros:

1 point=40 cents.

At the end of the experiment, you will receive the income that results from your decisions in cash.

In the following section, we will describe the exact experimental procedure.

The experiment

Summary:

There are three roles in this experiment: A, B, and C. All participants receive the same endowment of points. Participant A decides whether to take points from participant B. Participant B is passive. Participant C can deduct points from player A (points that A would receive for answering a post-experiment questionnaire) without thereby gaining any points and without being able to observe A's choice. Participants A, B, and C then state their beliefs about the behavior of other groups.

After the experiment, you will be asked to complete a questionnaire. In addition to some questions pertaining to the experiment, you will fill out a scientific questionnaire, for which you will receive an additional 20 points.

Procedure in detail:

There are three roles in this experiment: A, B, and C. Your role will be communicated to you via the screen before the experiment starts. You will not learn the identities of the other participants in your group. Likewise, the participants in your group will not learn your identity. Each group has real participants A, B, and C.

The experiment unfolds in four stages. In stage 1, only participant A makes a decision. In stage 2, only participant C makes a decision. In stage 3, all participants (A, B, and C) make decisions. Payoffs are assessed in stage 4.

Stage 1:

All participants receive an endowment of 20 points.

Participant A can decide whether or not to deduct 10 points from player B in order to gain 5 points.

Participant A's decision also specifies from which urn a ball will be drawn in stage 4. There are two urns containing 10 balls each. Urn BLACK is relevant if participant A deducts points from participant B. Urn WHITE is relevant if participant A does not deduct points from participant B. The number of black balls is weakly higher in urn BLACK than in urn WHITE. The remaining balls are white. The possible compositions for urns BLACK and WHITE are laid out in Table 1.

Composition	Urn **BLACK**	Urn **WHITE**
(1)	10 black & 0 white balls	0 black & 10 white balls
(2)	9 black & 1 white balls	1 black & 9 white balls
(3)	8 black & 2 white balls	2 black & 8 white balls
(4)	7 black & 3 white balls	3 black & 7 white balls
(5)	6 black & 4 white balls	4 black & 6 white balls
(6)	5 black & 5 white balls	5 black & 5 white balls

Participant A makes the decision in stage 1 for all possible urn compositions. The payoff-relevant composition will be determined in stage 4 by a random mechanism.

Stage 2:

In stage 2, participant C can deduct 10 points from player A's compensation for filling out the questionnaire. Participant C does not gain points thereby, but neither does the deduction cost participant C any points. Participant C can decide on this potential punishment contingent on the color of the ball drawn and the applicable composition of the urns BLACK and WHITE. Participant C does not observe participant A's decision in stage 1.

The color of the ball can inform participant C about whether or not participant A has deducted points from player B:

- When composition (1) applies, there are only black balls in urn BLACK and only white balls in urn WHITE. Knowing that a black ball was drawn means knowing that urn BLACK was relevant. (Urn BLACK is only relevant when participant A deducted points from participant B.) As a result, participant C knows participant A's choice when composition (1) applies.
- When composition (6) applies, urns BLACK and WHITE both contain 5 black balls and 5 white balls. As a result, knowing the color of the ball allows participant C no

inference about whether or not participant A deducted points from participant B in stage 1.

- For compositions (2)-(5), the probability of observing a black ball is higher when participant A deducted points from participant B than when no points were taken. When composition (2) applies, observing a black ball indicates participant A's taking with relatively little uncertainty; however, this uncertainty steadily increases for compositions (3)-(5).

In addition to the color of the ball, the general expectation about participants A deducting points from participants B influences the assessment of whether or not participant A did in fact deduct points from participant B. The following examples are intended to convey the relative importance of these two factors to you.

Table 2: Conditional probabilities when player C has the prior that 20% of all players A deduct points

	Probability that player A took points from player B if …		Probability that player A did not take points from player B if …	
	a black ball is drawn	a white ball is drawn	a black ball is drawn	a white ball is drawn
Urn B: 10 black and 0 white balls Urn W: 0 black and 10 white balls	100%	0%	0%	100%
Urn B: 9 black and 1 white balls Urn W: 1 black and 9 white balls	69%	3%	31%	97%
Urn B: 8 black and 2 white balls Urn W: 2 black and 8 white balls	50%	6%	50%	94%
Urn B: 7 black and 3 white balls Urn W: 3 black and 7 white balls	37%	10%	63%	90%
Urn B: 6 black and 4 white balls Urn W: 4 black and 6 white balls	27%	14%	73%	86%
Urn B: 5 black and 5 white balls Urn W: 5 black and 5 white balls	20%	20%	80%	80%

Table 3: Conditional probabilities when player C has the prior that 80% of all players A deduct points

	Probability that player A took points from player B if …		Probability that player A did not take points from player B if …	
	a black ball is drawn	a white ball is drawn	a black ball is drawn	a white ball is drawn
Urn B: 10 black and 0 white balls Urn W: 0 black and 10 white balls	100%	0%	0%	100%
Urn B: 9 black and 1 white balls Urn W: 1 black and 9 white balls	97%	31%	3%	69%
Urn B: 8 black and 2 white balls Urn W: 2 black and 8 white balls	94%	50%	6%	50%
Urn B: 7 black and 3 white balls Urn W: 3 black and 7 white balls	90%	63%	10%	37%
Urn B: 6 black and 4 white balls Urn W: 4 black and 6 white balls	86%	73%	14%	27%
Urn B: 5 black and 5 white balls Urn W: 5 black and 5 white balls	80%	80%	20%	20%

Participant C decides for all urn compositions. The composition of the payoff-relevant urn will be determined in stage 4 by a random mechanism.

Stage 3:

In stage 3, all participants specify their beliefs about the behavior of other participants. Specifically, participants are asked to specify how many participants A have decided to deduct points from their respective participant B for a given urn composition, and how many participants C punish for a given urn composition and color of the ball. As a result, there are three expectations for each urn composition. One composition will be selected by a random mechanism. You will receive 4 points for each correct expectation.

Stage 4:

A random mechanism chooses the urn composition. Participant A's choice for this urn composition becomes payoff-relevant and assigns whether urn BLACK or WHITE applies. Another random mechanism draws a black or a white ball according to the number of balls in the urn. Participant C's choice is implemented accordingly.

After the conclusion of the first experiment, a second experiment starts. This experiment is not related to the first one and will be explained to you on the screen. In this second experiment, your decision will be relevant to your payoff only. At the end of today's session, you will fill out a questionnaire for which you will receive 20 additional points. You will receive your payoffs after you have completed the questionnaire.

At the end of the experiment, all participants will receive their income **in cash**. Please raise your hand if you have any questions. One of the experimenters will come to you to answer them.

Below, you will find some test questions. Please raise your hand when you have answered all the questions. The experiment will start when all participants have answered all the questions.

Test questions

How many additional points does participant A obtain by deducting 10 points from participant B?

_____

How many additional points does participant C obtain by deducting 10 points from participant A?

_____

Is the color of the ball more informative for participant C with regard to the behavior of

participant A when composition (2) applies (where urn BLACK has 9 black balls and 1 white one, and urn WHITE has 1 black ball and 9 white ones) than when composition (3) applies (where urn BLACK has 8 black balls and 2 white ones, and urn WHITE has 2 black balls and 8 white ones)?

Yes: _____ No: _____

When urn BLACK contains 5 black and 5 white balls and urn WHITE contains 5 black and 5 white balls, is the color of the ball informative for participant C?

Yes: _____ No: _____

When urn BLACK contains 10 black and 0 white balls and urn WHITE contains 0 black and 10 white balls, is the color of the ball informative for participant C?

Yes: _____ No: _____

What is the conditional probability that participant A has deducted points from participant B when the ball is black and composition (2) applies (i.e., urn BLACK contains 9 black balls and 1 white ball, and urn WHITE contains 1 black ball and 9 white balls), when participant C's prior is that …

20% of all participants A deduct points from participant B: _____

80% of all participants A deduct points from participant B: _____

What is the conditional probability that participant A has *not* deducted points from participant B when the ball is black and composition (4) applies (i.e., urn BLACK contains 7 black balls and 3 white balls, and urn WHITE contains 3 black balls and 7 white balls), when participant C's prior is that …

20% of all participants A deduct points from participant B: _____

80% of all participants A deduct points from participant B: _____