Justice is in the Eyes of the Beholder

Eye Tracking Evidence on Balancing Normative Concerns in Torts Cases*

Christoph Engel & Rima Maria Rahal

Abstract

Frequently deciding legal cases requires an assessment in multiple, conceptually incompatible dimensions. Often one normative concern would call for one decision, and another normative concern for a different decision. The decision-maker must engage in balancing, with no help from overarching normative theory. A typical situation is torts. The decision must regularly balance concerns on behalf of the victim, the tort feasor and society at large, both on utilitarian and deontological grounds.

In this paper we use eye tracking to investigate in which ways laypersons' thought processes react to normative conflict in a set of 16 torts vignettes. If normative conflict is present, participants are less likely to agree with the likely outcome if the case were tried in a German court; they take longer to decide, and they fixate longer on normative concerns presented on a decision screen. Eye movements show that participants indeed consider multiple normative concerns in competition.

JEL: D01, D81, D91, K13, K40

Keywords: torts, fundamental normative relativity, compensation, deterrence, utilitarian and deontological concerns, balancing, eye tracking, machine learning

* We gratefully acknowledge helpful comments by audiences at the Workshop Experimental Methods in Legal Scholarship at ETH Zurich, the American Law and Economics Association Meeting in New York, the Conference on Empirical Legal Studies at Claremont, and at the University of Southern California.

1. Introduction

He who violates a legal rule and thereby inflicts harm on another person has to pay. At this level of abstraction, most legal orders agree. But must the victim be made whole, even if this ruins the tort feasor? Can the court double or triple up the payment if otherwise tort feasors gamble? Should the court use money as a substitute for irreparable harm, say to a person's honor? In such cases, legitimate normative concerns stand in conflict with each other. As these concerns have incompatible normative foundations, the decision cannot be derived from first principles. Any decision privileges some normative concerns at the expense of others. It requires "balancing" (cf. Aleinikoff 1987).

From the angle of rules that oblige tort feasors to pay damages, this paper engages with a normative debate that has been going on for centuries. But we are children of our time. We not only aim at making an empirical contribution. We exploit advanced computing power to investigate how laypersons reason when faced with such conflicts. In the middle of the previous century, Bertold Brecht could still write: "but the thoughts are free". No longer. With eye tracking, this generation has a chance to see choices in the making. We exploit the fact that there is a close relationship between eye gaze and attention (for a discussion of the potential and limitations of the method see Just and Carpenter 1980) to catch a glimpse of the legal mind in action.

We test student participants on a series of 16 sketches of legal cases. In all cases we inform participants that a court has ruled in favor of plaintiff, and obliged defendant to pay a defined amount of damages. Vignettes are constructed such that a typical German court would have made the reported decision (which we do, however, not tell participants). We ask participants whether they agree with this ruling. While they consider their assessment, they see a screen on which 9 potentially competing normative concerns are displayed. We attach a number to each of these concerns. We inform participants that this number captures the normative weight attached to this concern: how intensely are the victim, the tort feasor or society affected by the outcome of the case, in one of three dimensions each? We record on which of these numbers participants fixate, how often they do so, and in which sequence. In the first 4 cases, all normative concerns are in harmony. In the remaining 12 cases, we introduce a normative conflict.

As we had expected, our student participants have a harder time making up their minds in the conflict cases. They are less likely to agree with the court. They take longer to decide.[1] Some conflicts induce a characteristic gaze pattern that resonates with the legal debate. This in particular holds for cases where the damage is so large that making the victim whole would ruin the tort feasor. In such cases participants focus more often on compensation than in balanced cases, and they focus more on compensation than on alternative normative concerns, like deterrence. The eye tracking data demonstrates that our participants take the task very seriously, despite the fact that they have neither decision making authority nor professional training. They consider multiple normative concerns. They distribute their attention differently from case to case. This suggests that they try to do the individual

1 Conditional on the time trend, i.e. when controlling for learning.

conflict justice. Most interestingly: the pattern of fixations in the task at hand predicts whether they will agree with the court. The more intensely they ponder, the more they are likely to see why the decision made by the court makes sense. Despite the fact that solutions to these normative conflicts cannot be derived from first principles, via balancing the competing normative concerns even laypersons come to a meaningful assessment.

Using a standard machine learning tool, we find that participants fall into four characteristic types. These types differ in the normative concern that they tend to start with, and in the intensity with which they consider the competing concerns.

The remainder of the paper is organized as follows: Section 2 introduces the normative legal debate. Section 3 explains the design of the experiment. Section 4 derives hypotheses. In section 5 we report results. Section 6 explores the robustness of the results. Section 7 concludes with discussion.

2. The Normative Legal Debate

Paying damages may serve multiple purposes (for an indepth discussion see Schwartz 1994, Schwartz 1996, Oberdiek 2008, Oberdiek 2014). In first approximation, one may distinguish compensation, deterrence, and retribution (Geistfeld 2011). From a compensation perspective, damages are correlates of property. If a good is A's property and not B's, it is for A to decide what to do with the thing. If B impinges upon A's sphere of influence, and this makes the good less valuable for A (or even destroys it altogether), A must be made whole (more from Geistfeld 2014, Goldberg, Sebok et al. 2016). Compensation is an exercise of restorative, distributive justice (Shuman 1993, Gardner 2013, Sheinman 2014). The law steps in after a behavioral rule has been violated (in the simple example: the rule not to damage another person's property).

Policy makers may not want to wait that long. They may aim at prevention, rather than reparation. In this forward-looking perspective, the prospect of having to pay damages deters rule violations (Eisenberg and Engel 2014, Eisenberg and Engel 2016). When factoring in the payment of damages, in expectation the tort feasor hurts herself by violating the rule of non-intervention. Realizing that rule violations do not pay, she desists from the socially harmful activity (Landes and Posner 1987). While they differ in their focus, both arguments are consequentialist. Victims care about their property. By forcing the tort feasor to pay damages, the legal order protects the victim's interests.

Yet arguably, people do not only care about outcomes. They also care about principles. By impinging upon another individual's property, the tort feasor has violated a concrete rule: respect others by respecting what they care about. And she has violated the abstract rule: follow the rules that originate from a legitimate rule-making process. In so doing, she has deprived the victim (concrete rule) and the community (abstract rule) of respect they deserve. By paying damages, the tort feasor (belatedly) pays symbolic respect to the victim, and to the (legal) community. In this perspective, tort law is an exercise of corrective justice

(Posner 1981, Wells 1990, Gardner 2011). It also serves as a substitute for revenge (Hershovitz 2014).

Arguably, the obligation to pay damages is thus three for the price of one: the victim is compensated; future would-be victims are protected; the victim and the tort feasor are given a ritualized opportunity to restore their relationship, and to recognize that both of them are members of one and the same legal (epistemic, normative) community.

In principle, such a win-win-win situation is appealing. But as the introductory examples illustrate, when it comes to fine-tuning damages rules, the normative harmony may break down. Moreover, in most legal orders, judges may not confine themselves to deciding. They are also supposed to justify their decisions. Even if a decision could be justified on all three normative grounds, it may matter for the effect of a court ruling on the parties and the wider public which justification has been given. This holds all the more if different justifications suggest different decisions.

The three justifications rest on independent normative grounds. For compensation, one needs a theory of distributive justice. For deterrence, one needs standard economic theory. For retribution (and repent) one needs a deontological theory (Zamir and Medina 2011). Since either normative theory has a separate starting point, there is no (direct) way of translating one of them into the other. Admittedly, in behavioral economics a whole industry works on extending the domain of the economic model by incorporating "social preferences" into utility functions (for an overview see Fehr and Schmidt 2002). This implicitly requires expressing all relevant concerns, and fairness considerations in particular, in "utils", i.e. in one and the same normative currency. For the economic project, this is a highly productive enterprise. But if one conceptualizes social concerns as preferences, one takes out of the equation what this paper is genuinely interested in: distributive concerns, and respect for others as worthy beings, are not just "out there". They are competing starting points for normative reasoning, based on incompatible normative grounds.

In terms of first principles, distributive justice, welfare, and (individual and social) identity are distinct. But when it comes to agreeing or disagreeing with normative choice, decision makers have no reason to exclusively consider one of these normative starting points. Intuitively, one would rather expect decision makers to be motivated by some combination of all three normative arguments.

3. Design

We are using vignettes, all of which cover torts cases. Our manipulation is within subjects. Each participant sees a total of 16 vignettes. In four vignettes (*Baseline* scenarios), there is no conflict between compensation for harm, deterrence of future rule violations, and restoring justice. Participants see these 4 vignettes first. Between participants, we randomize order among these 4 vignettes. Thereafter, participants are presented with 12 vignettes cov-

ering normative conflict, which are presented in randomized order. In 4 of these 12 vignettes, compensation is prohibitively costly (*High Damage* scenarios). In these scenarios, there is thus a conflict between compensation and deterrence. From a deterrence perspective, partial compensation would be sufficient. But this would leave the victim partly harmed. In another 4 vignettes, the cost of compensation is trivial (*Low Damage* scenarios). If this is all the legal order requires, the obligation to pay damages does not deter rule violations. In these scenarios, there is thus the opposite conflict between compensation and deterrence. Finally, in yet another 4 vignettes, harm is nonpecuniary (*Nonpecuniary Damage* scenarios). Restitution in kind is also not an option. If the legal order nonetheless obliges defendant to pay a certain amount, this can be justified as a measure of deterrence. But it can also be considered necessary to symbolically restore peace in society.

Each vignette comes with the reported decision of an anonymous judge. The judge always rules for plaintiff, i.e. obliges defendant to pay the full amount plaintiff had asked for. Cases are chosen such that this is also the likely outcome, were the respective case tried in a German court. It is however not disclosed to participants that the decision is the likely outcome in court. If the participant disagrees with the court, on the next screen we ask her which decision she would have considered appropriate.

Each case is displayed on two separate screens. On the first screen, the scenario is introduced, by giving a sketch of the case, and a description of the harm suffered by plaintiff. On the second screen, the court ruling is presented and normative concerns are laid out. On this second screen, participants see potential concerns in three blocks. One block adopts the perspective of the victim, a second block adopts the perspective of the tort feasor, and a third block adopts the perspective of society. Each of the three blocks contains three items. Two items are of a utilitarian nature, the third item is of a deontological nature. From the perspective of the victim, compensation, retribution, and reconciliation are considered. From the perspective of the tort feasor, punishment, a loss of reputation, and the restoration of her relationship with the victim are considered. From the perspective of society, deterring others, pre-empting recidivism, and restoring peace in society are considered.

How these concerns are weighted in the process of deciding whether to agree with or reject the ruling is what we are chiefly interested in. To minimize visual biases affecting visual attention, we capture the extent to which each concern is addressed by the ruling with a number. Numbers range from severely negative (-10) to severely positive (+10). In the data analysis, we focus on these numbers. On this same second screen we also elicit support for or opposition against the court ruling. For the complete stimulus material, please see the Appendix.

We record decision time and the complete profile of eye gaze. We thus know (a) which areas of interest (AOI) the participant has visited; (b) how much time she has spent on each AOI; (c) in which sequence she has visited these AOIs. In the hypothesis section we explain in which ways we derive our dependent variables from this data.

The experiment was run in the lab of the Bonn Max Planck Institute. 106 students with various majors participated.[2] 59 of them were female. Mean age was 23.36 years. Participants received 12€ for their participation. For three participants, eye movements were not properly recorded due to technical failure. We exclude these participants from the analysis. Two participants did not focus on any of the areas of interest during the entire experiment. We exclude these participants from the analysis of eye gaze. Only 21 participants focused on at least one area of interest for all 16 tasks. The remaining participants only focus on relevant areas of interest for some of the tasks. This is why, for the analysis of eye gaze, we have an unbalanced panel.

4. Hypotheses

We have hypotheses at the level of choices and at the level of the decision process leading to these choices. The *Baseline* scenarios are constructed such that there is no conflict between distribution, allocation, and the restoration of justice. In the *High Damage* scenarios, we introduce a conflict between compensating the victim (distribution) and deterring intrusions into foreign affairs (allocation). This conflict is also present in the *Low Damage* scenarios, yet has opposite sign. We finally introduce a conflict between utilitarian (distributional or allocative) and deontological concerns in the *Nonpecuniary Damage* scenarios. Each of these conflicts makes the court ruling in favor of plaintiff more questionable. We therefore expect

> **H_1 choices**
> Participants are less likely to agree with the court ruling in the *High Damage, Low Damage,* and *Nonpecuniary Damage* treatments than in the *Baseline*.

Overcoming a normative conflict is more difficult than deciding when no such conflict is present. We expect this difficulty to translate into participants taking longer to decide, and fixating for longer periods of time (Krajbich, Armel et al. 2010, Fiedler and Glöckner 2012, Rosch and Vogel-Walcutt 2013). We have no directed hypotheses regarding the type of conflict, but we are open to the possibility that for participants some conflicts are easier to dissolve than others. We therefore expect

> **H_2 decision time and fixation duration on all items**
> a) Decision time and the fixation duration are higher in the *High Damage*, *Low Damage*, and *Nonpecuniary Damage* scenarios than in the *Baseline*.
> b) Decision time and the fixation duration differ between the *High Damage*, *Low Damage*, and *Nonpecuniary Damage* scenarios.

On the same screen on which we ask participants to express support or dissatisfaction with the court ruling, we present a set of 3 x 3 concerns, each of which comes with a number

2 12 participants study law. To be on the safe side, we have rerun all structural models in the Appendix with an additional dummy that is 1 if the participant studies law as a control variable. This control variable never turns significant. These additional estimations are available online via https://osf.io/cb5k3/?view_only=292ad95dfb004ac98d3f68fb41f4290c.

attached to it. We tell participants that these numbers reflect weights. This decision screen is in the spirit of a decision aid. It makes an offer for structuring the normative thinking about the case at hand. We use eye tracking as we expect participants to use this decision aid, and to do so in a manner that is informative about the difficulty they had in coming to a conclusion, and about the comparative importance of normative concerns in their assessment.

Our main interest is in reflections of participants' decision process in their eye movements. Therein, we rely on the assumption that attention and eye movements are linked (Just and Carpenter 1980), especially in complex tasks (see Rayner 1998), to the extent that "the most active location in working memory will eventually determine the most likely direction of the eye movement at a given point in time" (Huettig, Olivers et al. 2011: 141). We expect that the attention participants pay to a specific normative concern is reflected in the number of fixations to the associated AOI (for a review, see Orquin and Mueller Loose 2013), in the total duration of fixations on the term and number (Rayner 1998), and in the frequency with which the participant revisits the concern (Çöltekin, Fabrikant et al. 2010). Moreover, we expect that cases with normative conflict require participants to access more information than cases without conflict (number of fixations, Fiedler, Glöckner et al. 2012).

In the *High Damage* scenarios, we induce a conflict between the distributive goal of compensation and the allocative goal of deterrence. Distributional justice would call for full compensation. For deterrence, it would suffice to threaten future tortfeasors with a sufficiently severe reaction of the judicial system. If tortfeasors show reactance, a prohibitively high award might even be counter-productive. The intervention might also be counter-productive from a societal perspective. Individuals might shy away from risk prone activities that, overall, are socially beneficial. In these cases, full compensation might also appear inappropriate from the perspective of the victim. While she is of course compensated, she may feel that the authoritative reaction goes beyond what is necessary for retribution, and that the severity of the intervention makes it almost impossible to restore her personal relationship with the tortfeasor. By the same token, from the perspective of society, one may question the effect in terms of special prevention. If defendant is (close to) bankrupt, she may well be put on a socially disruptive path, just because she sees no scope for making money in the ordinary economy. The same way as the victim may see no room for restoring justice in their personal relationship, general peace in society may also suffer. Finally from the perspective of the tortfeasor, one may wonder whether it will now not be her who is regarded as a victim (of the judicial system), so that there would even be a counter-productive reputation gain. And if the victim finds it hard to restore justice in the personal relationship, so will the tortfeasor. Based on these considerations, in the *High Damage* scenarios we expect a conflict between the desire to compensate the victim and all remaining normative concerns.

In the *Low Damage* scenarios, compensation is easy, but general and special prevention are problematic. One may also argue that peace in society is not restored if the obligation to pay damages is almost immaterial for the tortfeasor. Along the same lines one may consider punishment to be imperfect, a loss in reputation almost non-existent, and hardly any

opportunity for restoring justice in the bilateral relationship between the victim and the tortfeasor. Finally, one may question whether an impulse of the victim for retribution is addressed. This is why we expect the general and special prevention items to draw most of the attention in the *Low Damage* scenarios.

In the *Nonpecuniary Damage* scenarios, compensation is impossible. But there could still be a utilitarian desire for deterrence, and a deontological desire for restoring peace. In these scenarios we thus do not have reason to believe that a single item attracts most attention. But we do expect that either the relevant utilitarian or the deontological items play a greater role than the compensation item in assessing the court ruling than in the *Baseline*.

Note that more fixations on a certain item, and a longer fixation duration, do not imply that the participant agrees with the normative concern that the item captures. She may, by contrast, deem the fixated concern particularly confusing. All we can infer from fixations is that the participant attends to the normative concern in question, suggesting she finds it to be relevant for the decision in some way.

From these considerations we derive the following expectations:

H_3 total fixation time and total number of fixations per item
Total fixation time, and the total number of fixations
a) in the ***High Damage*** scenarios
i) for the compensation item are higher than in the *Baseline*.
ii) for the compensation item are higher than for the remaining items.
iii) for the victim items are higher than for the tortfeasor and societal items.
b) in the ***Low Damage*** scenarios
i) for the general and special prevention items are higher than in the *Baseline*.
ii) for the general and special prevention items are higher than for the remaining items.
iii) for the societal items are higher than for the victim and tortfeasor items.
c) in the ***Nonpecuniary Damage*** scenarios
i) for all items except compensation are higher than in the *Baseline*

We have preregistered these hypotheses on the Open Science Framework (reviewer link: https://osf.io/cb5k3/?view_only=292ad95dfb004ac98d3f68fb41f4290c), where all materials, data and analyses scripts are also made available.

In addition to conducting confirmatory analyses with regard to these hypotheses, we further pursue a number of exploratory questions. These exploratory investigations are primarily concerned with patterns of fixations, as well as inter-individual and item-specific differences in the elicited gaze behavior. Data on individual characteristics of the decision makers was also collected, but reporting the exploratory analyses associated with these measures is beyond the scope of this paper.

5. Results

a) Choices

In Hypothesis 1 we had expected that participants would be less likely to agree with the court ruling in the *High Damage*, *Low Damage* or *Nonpecuniary Damage* treatments, compared with the *Baseline* without normative conflict between distribution, allocation and deontological concerns. As the left panel of Figure 1 shows, this is indeed what we find (Wilcoxon, N = 103, p < .0001). In the *Baseline*, on average 69% of all participants agree with the court, while less than half of them do so in the treatments (49% in the *Low Damage* condition, 46% in the *High Damage* condition, and 40% in the *Nonpecuniary Damage* condition).[3]

This gives us

> *Result 1: Participants are less likely to agree with the court if there is a normative conflict.*

After they recorded agreement or disagreement, we asked those participants who had disagreed which amount they would have found appropriate. As we have a number of outliers on the high side, means would be misleading. This is why, in the right panel of Figure 1 we report logs. As one sees, only in the *Baseline* the majority of those who disagree do so because they would have wanted the plaintiff to receive more money. In all treatments, the majority would have given the plaintiff less money than the court.

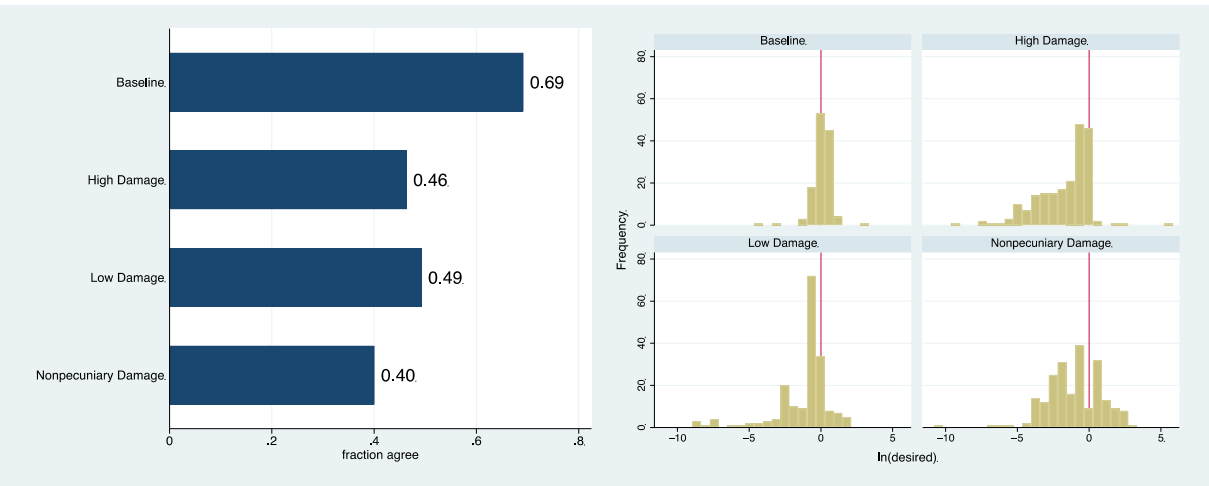


Figure 1
Choices
left panel: fraction of participants who agree with the court ruling
right panel: distributions of amount of damages desired by participants who disagree with the court ruling, expressed as the log of the percentage of the amount determined by the court

3 We have hypotheses about multiple measures, all of which are generated by the same participants. One may therefore worry about dependence. To be on the safe side, in the appendix we report structural (multivariate) models that take this dependence into account when calculating standard errors. The model in Table 5 clearly supports the nonparametric result.

b) Decision Time and Fixation Duration on All Items

In Hypothesis 2a we had expected that participants need more time to decide in the treatments than in the *Baseline*, and that fixation duration is longer. As the upper panels of Figure 2 show, if we only look at the raw data, we find the opposite (Wilcoxon, N = 103, p < .0001 for both comparisons between the *Baseline* and all treatments pooled, and for all comparisons with individual treatments, expect for decision time, *Baseline* vs. *High Damage*, p = .0157).[4]

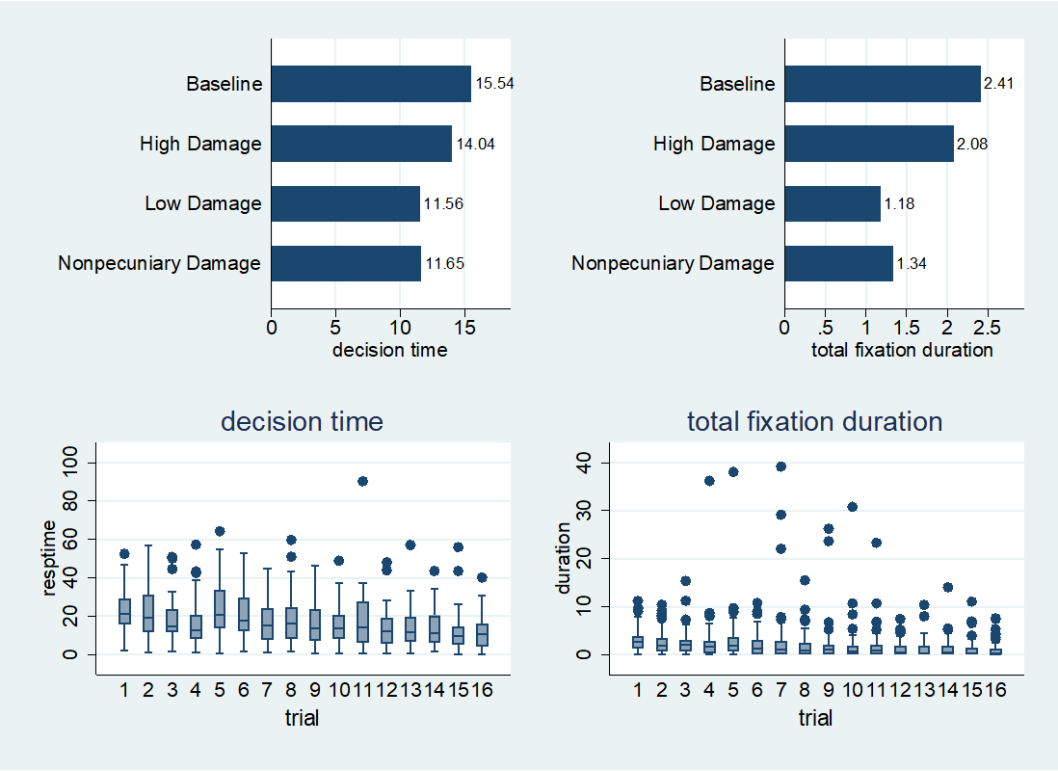


Figure 2
Decision Time and Total Fixation Duration per Treatment and Trial

Yet recall that all participants first saw the four vignettes of the *Baseline*. We have decided in favor of this order as we have constructed these vignettes to be free from conflict between the three normative dimensions at stake. We wanted participants to first gain experience with the task before they have to deal with normative conflict. Now the lower panels of Figure 2 show that we have a clear time trend, for both dependent variables. On average, decision time and the total duration of fixations per participant and task go continuously down in the first four periods, but jump up again as soon as participants face their first conflict vignette.

If we condition on the time trend, we find the predicted longer decision time for treatments, both if we pool all treatments (Model 1 of Table 1), and if we compare the *Baseline* with each individual treatment (Model 2). Results look similar, but are less clean, for total fixation

4 For fixation duration, for each participant and task we use total, not average fixation duration, including the imputed duration of 0 if the participant has not fixated on any relevant area of interest in the task in question.

duration. If we pool all treatments, conditional on the time trend we no longer find a significant difference between the *Baseline* and the treatments (Model 3). If we compare the *Baseline* with each individual treatment, we only find that total fixation duration is significantly higher in the *High Damage* condition, compared with the *Baseline* (Model 4).[5]

	decision time		total fixation duration	
	Model 1	Model 2	Model 3	Model 4
Baseline	-2.604*** (.700)		-.304 (.219)	
High Damage		4.228*** (.766)		.855*** (.239)
Low Damage		1.707* (.763)		-.057 (.238)
Nonpecuniary Damage		1.884* (.769)		.121 (.240)
trial	-.716*** (.066)	-.716*** (.065)	-.148*** (.021)	-.148*** (.0203)
cons	19.934*** (.935)	17.331*** (.727)	3.085*** (.260)	2.782*** (.185)
N obs	1648	1648	1648	1648
N uid	103	103	103	103

Table 1
Decision Time and Total Fixation Duration per Treatment and Trial
linear models with individual random effects
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05

We conclude

> *Result 2: Conditional on the time trend, decision time is higher when there is a normative conflict. If this conflict results from the fact that compensation is prohibitively costly, total fixation duration is also higher.*

In Hypothesis 2b we had further expected that decision time and fixation duration would differ between treatments. We do indeed find that decision time is higher in the *High Damage* than in the *Low Damage* (Wilcoxon N = 103, p = .0001) and in the *Nonpecuniary Damage* treatment (p = .0002).[6] Likewise total fixation duration is higher in the *High Damage* than in the *Low Damage* (Wilcoxon N = 103, p < .0001) and in the *Nonpecuniary Damage* treatment (p < .0001).[7] We do, however, not find a difference between the *Low Damage* and the *Nonpecuniary Damage* treatment, for either dependent variable.[8]

5 We had not preregistered the need to condition on the time trend. In the structural model of Table 5 we replicate findings from the otherwise equivalent models 2 and 4 of Table 1, but the treatment effect on decision time for the *Low Damage* treatment is only weakly significant.

6 These effects replicate in Wald tests on the decision time component of the structural model of Table 5: *High Damage* vs. *Low Damage* p = .0005; *High Damage* vs. *Nonpecuniary Damage* p = .0008.

7 These effects also replicate in Wald tests on the fixation duration component of the structural model of Table 5: *High Damage* vs. *Low Damage* p < .0001; *High Damage* vs. *Nonpecuniary Damage* p = .0001.

8 For these comparisons, there is no need to condition on the time trend as the order of these 12 vignettes has been randomized across participants. Yet see the fixation duration component in the structural model of Table 5 for this test, regardless.

This gives us

> *Result 3: Decision time and total fixation duration are higher in the* High Damage *than in the* Low Damage *or* Nonpecuniary Damage *treatments.*

c) Fixation Duration and Number of Fixations on Individual Items

In Hypothesis 3 we have laid down our expectations regarding the attention for individual items. We consider the total time a participant fixates on the item in question in a given task, and the number of times she fixates this item, as proxies for her attention.

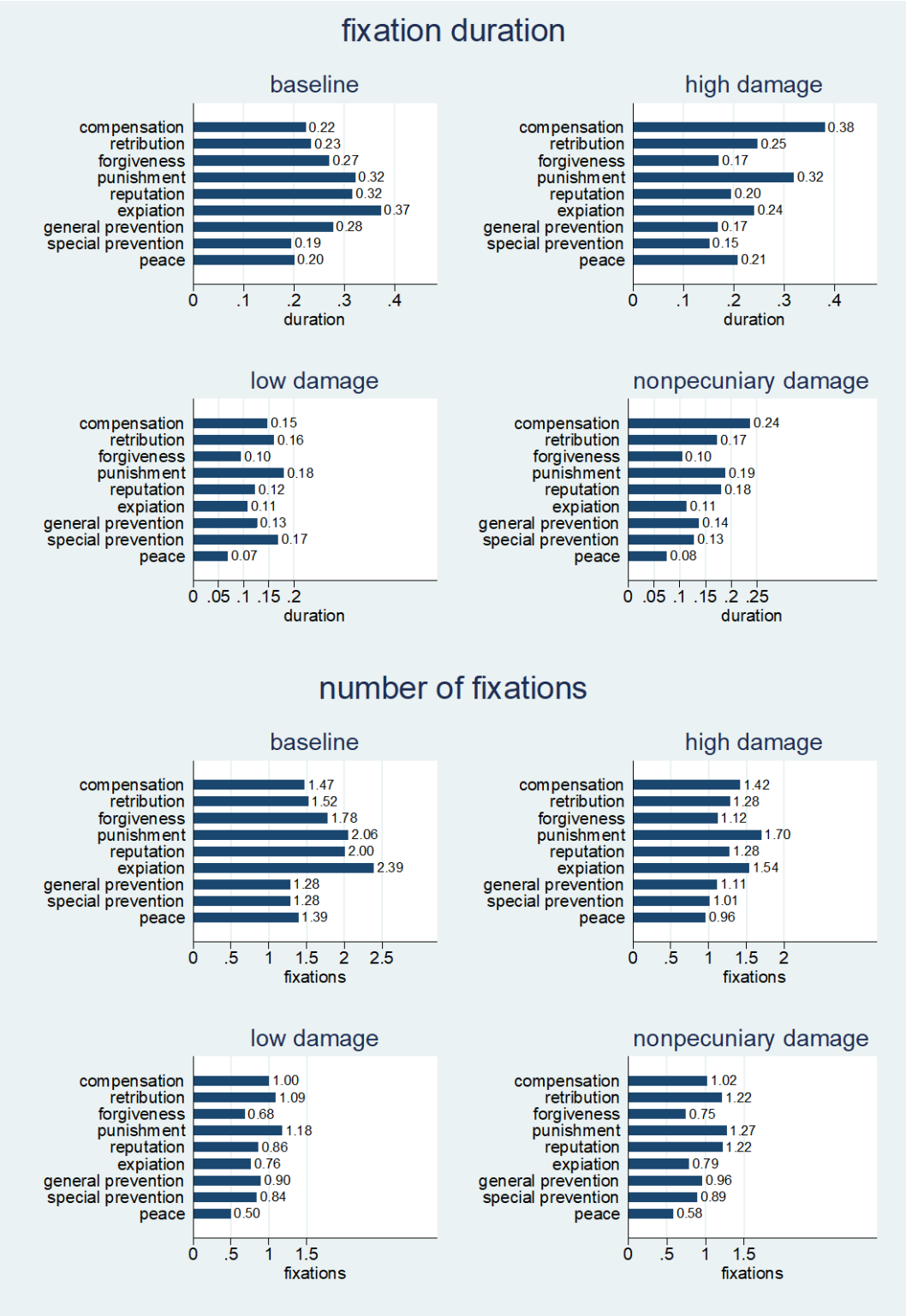


Figure 3

Fixation Duration and Number of Fixations on Individual Items

Hypothesis 3a considers the *High Damage* treatment. Subhypothesis i) predicts that total fixation duration on the compensation item will be higher in this treatment (.38) than in the *Baseline* (.22). This is indeed what we find (Wilcoxon N = 103, p = .0170).[9] Yet for the number of fixations, in a non-parametric analysis exclusively over fixations on the compensation item we find the opposite (*Baseline* 1.47, *High Damage* 1.42, p = .0074). The effect disappears, though, in parametric analysis (Table 6 in the Appendix).[10]

In subhypothesis ii) we had predicted that, in the *High Damage* condition, attention on the compensation item would be more pronounced than attention on the remaining items. Descriptively, this is indeed what we find for total fixation duration (compensation mean .38, other items mean .21). Yet non-parametrically, this difference is not significant. It is significant, though, when regressing total fixation duration per task and item on the fact that the item is the compensation item.[11] Again the effect is more pronounced on fixation duration, not on the number of fixations. For this second dependent variable, the descriptive difference is small (compensation mean 1.42, other items mean 1.25). Yet parametrically, we find a clear and significant effect.[12]

In subhypothesis iii) we finally had predicted that, in the *High Damage* condition, attention on victim items would be more pronounced than attention on tort feasor or society items. If we pool results of victim items on the one hand, and non-victim items on the other, non-parametrically we do not find significant differences for either total fixation duration or the number of fixations.[13] The picture clears, however, if we separately compare attention on victim items with attention on tort feasor or on society items. We find that, indeed, attention on victim items is more pronounced than on society items (fixation duration mean .266 vs. .176, Wilcoxon N = 103, p = .0204; number of fixations mean 1.273 vs. 1.030, p = .0538). Yet, against expectations, we find that attention on tort feasor items is even more pronounced than on victim items (fixation duration on victim items mean .266 median .143 vs. on tort feasor items mean .252 median .185, Wilcoxon N = 103, p = .0040; number of fixations on victim items mean 1.273 vs. on tort feasor items mean 1.505, p = .0002).[14]

---

9 One may again worry that we test the same participants on multiple measures, which is why in the appendix we report structural (multivariate) models that take this dependence into account, Table 6. Whether or not we control for the time trend (models 1 and 2), the net effect of *High Damage* and the interaction with compensation, which corresponds to the non-parametric test, is significant, p = .0025 (model 1), p < .0001 (model 2).

10 In Table 6 model 1 the net effect of *High Damage* and the interaction with compensation on the number of fixations has p = .6522. It is only significantly different from zero (p < .0001) in model 2, i.e. conditional on the time trend.

11 The interaction of *High Damage* with compensation in the fixation duration component of Table 6 is strong, and significant at the 1 ‰ level.

12 The interaction of *High Damage* with compensation in the number of fixations component of Table 6 is significant at the 1 % level.

13 In the parametric model of Table 7, we find the expected effect for fixation duration, but not for the number of fixations.

14 In Table 7 Models 3 and 4, we replicate the non-parametric result for the number of fixations, but not for fixation duration. This is due to the fact that we have a few outliers, with very high fixation duration, in the *High Damage* condition.

We summarize

> *Result 4: Total fixation duration on the compensation item is higher in the* High Damage *condition than in the* Baseline.
> *In the* High Damage *condition, total fixation duration on the compensation item is higher than on the remaining items, and attention on tort feasor items is higher than on victim items.*

Hypothesis 3b considers the *Low Damage* treatment. Subhypothesis i) predicts that, in this condition, attention for the general and special prevention items is higher than in the *Baseline***.** We find the opposite. Mean total fixation duration per task on these two items is .236 in the *Baseline*, vs. .148 in the *Low Damage* condition (Wilcoxon N = 103, p < .0001), and mean number of fixations per task on these two items is 1.283 in the *Baseline* vs. .869 in the *Low Damage* condition (p < .0001).[15]

Subhypothesis ii) predicts that, in the *Low Damage* condition, attention on the general and special prevention items is more pronounced than on the remaining items. We do not find any significant differences, though, neither for total fixation duration nor for the number of fixations.[16]

Subhypothesis iii) expects that, in the *Low Damage* condition, attention on societal items is higher than attention on the remaining items. We find the opposite. Mean total fixation duration on societal items is .121, while it is .136 on the remaining items (Wilcoxon N = 103, p = .0076) and mean number of fixations on societal items is .745, while it is .931 on the remaining items (p = .0016). Yet parametrically, these results only hold for the number of fixations.[17]

Finally, Hypothesis 3c predicts that attention on all items other than compensation will be higher in the *Nonpecuniary Damage* condition than in the *Baseline*. Actually we find the opposite. Mean total fixation duration on these items is 1.502 in the *Nonpecuniary Damage* condition, but 1.601 in the *Baseline* (Wilcoxon N = 103, p = .0527), and mean number of

15 The equivalent test in the structural models of Table 8 is the net effect of *Baseline* and the interaction of *Baseline* with prevention. In model 1 it is p = .0161 for fixation duration, and p < .0001 for the number of fixations. When controlling for the time trend (model 2), the effect is insignificant for fixation duration (p = .2907), but remains significant for the number of fixations (p = .0071).

16 This also holds for the structural models of Table 8: the coefficient for prevention is insignificant for both dependent variables in both models.

17 The equivalent test in the structural models of Table 7 requires to test the net effect of the item being societal + the interaction with the *Low Damage* treatment against 0 (for the comparison with victim items); to compare this net effect with the parallel net effect for tort feasor items: tort feasor + interaction with *Low Damage* against 0 (for the comparison with tort feasor items). The exact equivalent to the non-parametric test is the following: the model prediction for societal items in the *Low Damage* condition, minus the mean prediction for victim and tort feasor items in this condition. The relevant test is societal + the interaction with the *Low Damage* treatment – 1/2 * tort feasor + interaction with *Low Damage* = 0 (as there are as many victim items as there are societal items). The comparison of the number of fixations on societal items with victim items has p = .0104, with tort feasor items has .0077, and with the mean of these items has .0025. p-values are almost identical in the Wald tests after model 4.

fixations is .959 in the *Nonpecuniary Damage* condition, but 1.714 in the *Baseline* ($p < .0001$).[18]

d) Fixation Patterns

Our hypotheses, and the corresponding parts of the results section, are concerned with aggregated data. In subsections a) and b) aggregation is at the level of entire tasks. In subsection c) fixation duration and the number of fixations are aggregated over items in tasks. Hence if an individual, in a given task, has repeatedly fixated on the same item, all these fixations are represented by a single number. In this final, exploratory section, we consider the sequence in which participants have visited the 9 normative items they always have found on the decision screen.

i) Descriptive Analyses

Figure 4 illustrates the richness of the data. On some tasks (like the overtaking or the speculation task) the participant exhibits many more fixations than on others. On some tasks (like the vintage car or the lottery task) the participant first considers the concerns of the victim, while on other tasks (like that statics or the speculation task) she starts with concerning the tort feasor. For some tasks (like the sparks or the music task) the participant essentially goes once through all concerns. For others (like the collision or the piracy task) the participant revisits items before making up her mind.

18 The equivalent test in the structural models of Table 6 is the net effect of the compensation item and its interaction with the *Nonpecuniary Damage* treatment. If we do not control for the time trend (model 1), this net effect is only significantly different from 0 for the number of fixations, not for their duration ($p = .0003$). If we control for trial sequence, for the duration of fixations, the effect is significant at conventional levels ($p = .0071$), while it is only weakly significant for the number of fixations ($p = .0736$).

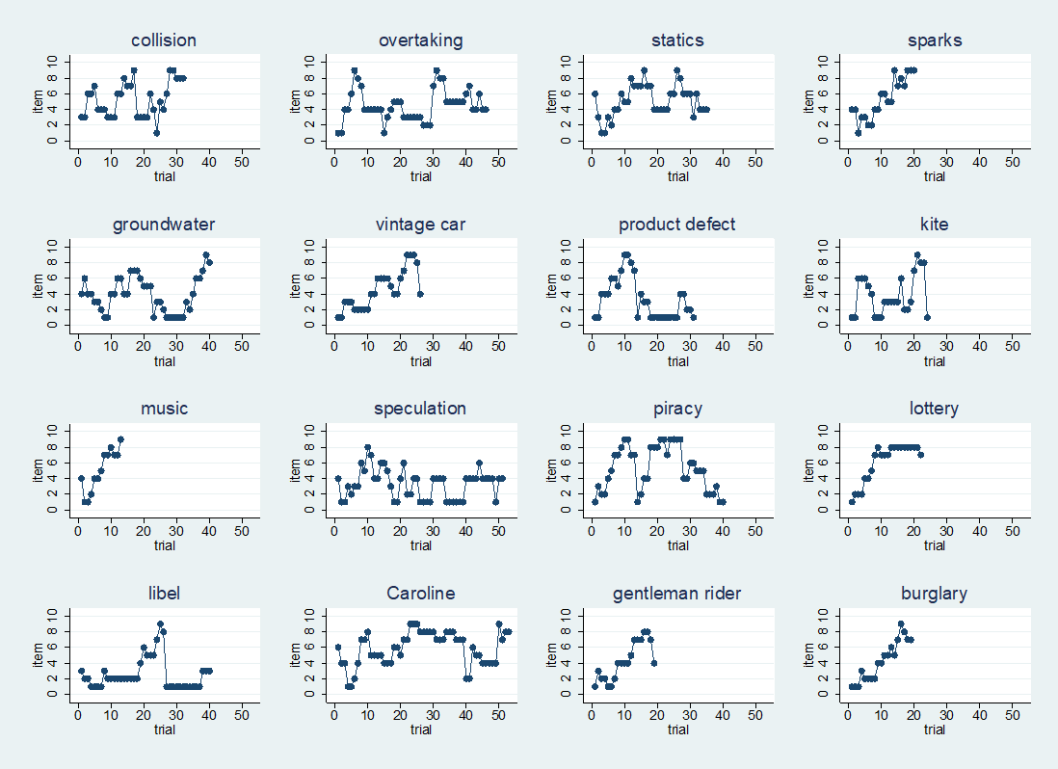


Figure 4

Example of Fixation Sequence Across Tasks

items: victim: 1 compensation, 2 retribution, 3 forgiveness
tort feasor: 4 punishment, 5 reputation, 6 expiation
society: 7 general prevention, 8 special prevention, 9 peace

Table 2 shows that variance across participants is pronounced. Participants on average fixate on at least one item for the victim, the tort feasor and society in approximately two thirds of all tasks. But some participants never do, and others do on all tasks. The mean number of fixations per task and individual ranges from 2 to almost 40 fixations, with a mean of about 14. On average participants are most likely to start fixating with a tort feasor item. But starting with a victim item is only a little less frequent. And participants start with a society item in an average of 29%. For last fixations, the victim item is most frequent, but the differences between the three classes of items are not pronounced. Participants on average revisit about one benefit area (victim, tort feasor or society) before making up their minds. But some individuals do so much more often.

The foregoing are sample averages over participant averages. With thus already find pronounced variance at the level of individuals, not only at the population level. In other words, most participants do exhibit a characteristic personal style of considering the competing normative concerns.

	mean	median	min	max	sd
tasks with all benefit areas visited	.622	.643	0	1	.274
number of fixations per task	13.872	13.125	2	39.438	6.832
first fixation on victim item	.343	.286	0	1	.291
first fixation on tort feasor item	.369	.357	0	1	.258
first fixation on society item	.288	.25	0	1	.263
last fixation on victim item	.355	.333	0	.923	.209
last fixation on tort feasor item	.339	.308	0	1	.204
last fixation on society item	.305	.286	0	.812	.189
number of revisited benefit areas per task	.963	.688	0	4.063	.789

Table 2

Characteristics of Participants by Fixation Sequences

all indicators are first calculated by participant and task
then averaged over individuals
then averaged over the sample
tasks with no single fixation not covered

Table 3 shows that, overall, there are also no clear attention patterns per task and, as treatments consist of four tasks each, also not per treatment. Descriptively, in the treatments inducing normative conflict participants are less likely to fixate at least once on victim, tort feasor and society items. In the treatments we find a smaller mean number of total fixations. In the treatments participants revisit less items. But all these observations follow from the time trend (see Figure 2). In some tasks, participants are more likely to start with a victim item (for instance in the kite task). But even in such tasks alternative starting points are not infrequent.

	N	all vis	num fix	fix start			fix end			num revis
				vic	perp	soc	vic	perp	soc	
rear-end collision	89	.831	19.29	29	35	25	30	32	27	1.449
overtaking	84	.738	17.80	25	43	16	29	30	25	1.512
statics	88	.727	17.17	28	33	27	32	32	24	.818
sparks	93	.817	16.47	30	34	29	31	32	30	1.387
groundwater	70	.714	18.81	19	34	17	26	21	23	1.229
vintage car	65	.662	16.66	20	27	18	27	25	13	1.092
product defect	64	.719	16.95	20	18	26	24	21	19	.875
kite	69	.783	17.71	32	21	16	15	24	30	.942
music	66	.667	11.76	25	16	25	24	22	20	.636
speculation	62	.613	13.21	22	21	19	24	28	10	1.161
piracy	70	.557	12.51	24	23	23	27	18	25	1.100
lottery	62	.581	12.11	25	21	18	19	20	23	.806
libel	65	.646	13.52	22	23	20	29	12	24	.815
Caroline	75	.733	13.05	23	25	27	26	23	26	.987
gentleman rider	69	.623	12.84	22	25	22	26	22	21	1.188
burglary	69	.580	12.17	22	25	22	18	29	22	.942

Table 3
Characteristics of Task by Fixation Sequences
all indicators are first calculated by participant and task
then averaged over tasks
then averaged over the sample
tasks with no single fixation not covered
column titles are as row titles in Table 2

ii) Fixation Patterns related to Choices

This is, however, not to say that the sequence of fixations were immaterial. Table 4 shows that it does matter. To succinctly characterize the sequence of fixations, separately for each individual and task, we run the following local regression

$$i_f = \beta_0 + \beta_1 \varphi_f + \beta_2 \varphi_f^2 + \beta_3 \varphi_f^3 + \varepsilon$$

where i is the id of the item, with items 1-3 being victim items, 4-6 tort feasor items, and 7-9 society items (for the full list see caption of Figure 4). φ stands for the fixation number, with the fixation in question being indexed by f. ε is error. Note that the quadratic and cubic terms capture the characteristic non-linearities in fixation sequences (see Figure 4). The regressions of Table 4 use the coefficients from these local regressions to explain choices. Attention patterns turn out to predict choices. Participants who start with and put more stress on tort feasor or societal items (and hence have a larger β_0 on the task in question) are more likely to agree with the court, and support the payment of damages. The same holds for participants who, in the task in question, go from concerns for the victim, to concerns for the tort feasor, to concerns for society (and not the other way round), and therefore have a positive β_1.

The interpretation of β_3 is less straightforward.[19] Geometrically, if β_3 has opposite sign of β_2, the predicted pattern has an inflection point. If the fixation pattern was U-shaped up till some point, and turns inversely U-shaped thereafter, β_2 is positive and β_3 is negative. The most intuitive interpretation is a participant coming back to concerns she had considered before. We stress that this interpretation is likely not always appropriate, but will hold often.[20] The significant negative effect of β_3 suggests that participants are more likely to agree with the court if they elaborate more intensely on the competing normative concerns.

Individual fixation patterns remain virtually unchanged if we control for the fact that, in the task in question, the participant has visited all benefit areas at least once (Model 2), and for the presence and type of normative conflict (Model 3). All these controls have explanatory power. Participants are substantially more likely to agree with the court if they have visited each benefit area (victim, tort feasor, society) at least once, and (as already shown in Figure 1) they are less likely to agree with the court in any of the conflict scenarios, compared with the *Baseline* where all normative concerns point into the same direction.

19 We do not report a specification with all of $\beta_0 - \beta_3$ as explanatory variables, as β_2 and β_3 are too highly correlated (the pairwise correlation between both coefficients is .971).

20 It for instance holds for the fixation pattern of the participant in Figure 4 in the overtaking task.

	model 1	model 2	model 3	model 4
β_0	.017** (.006)	.016** (.006)	.015** (.006)	
β_1	.028** (.009)	.027** (.009)	.025** (.009)	
β_3	-.280* (.128)	-.277* (.128)	-.248* (.124)	
cluster 2				-.070 (.052)
all benefit areas visited		.137*** (.032)	.106*** (.032)	.112** (.033)
High Damage			-.236*** (.038)	-.344*** (.055)
Low Damage			-.180*** (.039)	-.232*** (.057)
Nonpecuniary Damage			-.278*** (.038)	-.340*** (.053)
cluster 2 * *High Damage*				.199* (.078)
cluster 2 * *Low Damage*				.097 (.079)
cluster 2 * *Nonpecuniary Damage*				.088 (.078)
cons	.441*** (.034)	.354*** (.040)	.539*** (.004)	.650*** (.046)
N obs	1160	1160	1160	1113
N uid	101	101	101	100

Table 4

Effect of Attention Pattern on Choices

linear probability models with individual random effects
dv: dummy that is 1 if participant agrees with court ruling
$\beta_0, \beta_1, \beta_3$ coefficients from local regression, per individual and task, of item on trial
cluster 2: as in Figure 5 and accompanying text
model 4: only trials classified as clusters 1 or 2
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05

iii) Cluster Analyses

The foregoing analysis treats all fixation patterns in all tasks as elements of a homogeneous population. In the final step of the (exploratory) analysis, we check whether there are characteristic patterns. To that end, we use the coefficients of the local regressions as input for a classic clustering technique, the k-means algorithm. It turns out that a model with 4 clusters has the best fit.[21] Figure 5 shows that this model has a very intuitive interpretation.

21 We standardize all four variables to make sure that they have symmetric impact on the classification. We repeat the procedure 20 times with randomly selected starting points and use majority vote for classification, to make sure that results do not depend on starting points. We use Euclidian distance. Classification is done with the kmeans command in R. If we only ask for 3 clusters, almost all trials brought into clusters 3 and 4 are lumped together. If we ask for 5 clusters, cluster 1 is split into two clusters (with 3 and 14 trials), while all remaining trials are classified the same ways as when asking for 4 clusters.

The figure averages over fixations within each of the 4 clusters, for the first 20 fixations.[22] In trials classified as part of clusters 3 and 4, fixations start with tort feasor items and then either move to victim items (cluster 3), or to societal items (cluster 4). In these two clusters, fixation sequences are very short. Such sequences are also infrequent (cluster 3: 17 sequences; cluster 4: 30 sequences).

The sequences of the remaining two clusters are much more frequent (cluster 1: 556 sequences; cluster 2: 557 sequences) and typically much longer (cluster 1: mean 11.86, max 60; cluster 2: mean 13.40, max 105). After approximately seven fixations, in both clusters average fixations look very similar, and are in the middle of the item space. This follows from the fact that participants put the initial inclination into perspective, by fixating on other items. Where the two clusters clearly differ is the beginning. Sequences brought into cluster 1 start with societal items, while sequences brought into cluster 2 start with victim items.

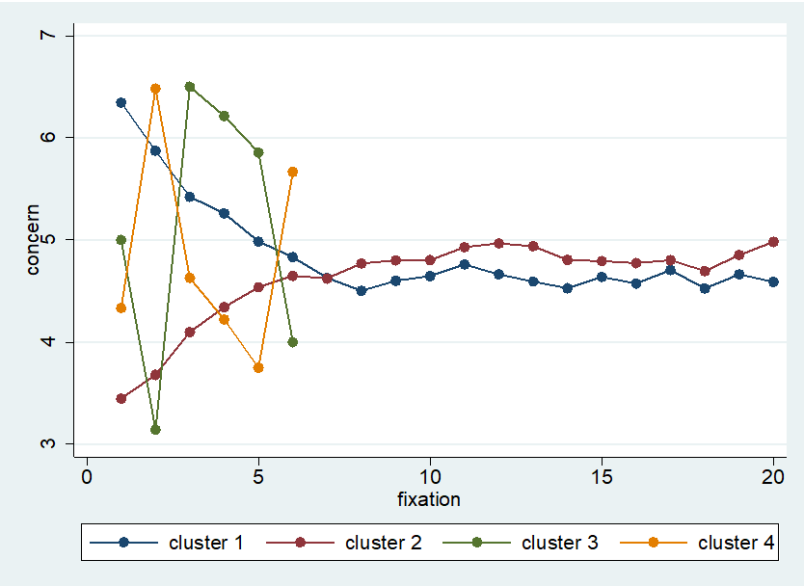


Figure 5
Fixation Patterns
x-axis: sequence of fixations
y-axis: normative concerns (for codes see legend of Figure 4)
lines are averages of all trials classified to be in one of the clusters
first 20 fixations represented

Relation of Clusters and Choices. Model 4 of Table 4 shows that the intuitive pattern also matters for choices. It follows from the significant positive interaction between the trial classified as cluster 2 and the *High Damage* scenarios that participants who start their fixation sequence with victim items are substantially and significantly more likely to agree with the court (and hence compensate the victim, despite the fact that the payment is very onerous for the tort feasor).[23]

22 We do not include later trials in the graph in the interest of making the characteristic difference between types visible.

23 Results look similar if we drop the dummy that is 1 if the participant, in the trial in question, has fixated on all 9 items (we have added the dummy to make it easier to compare the specifications), and if we add

Relation of Clusters and Normative Conflict. Clusters are defined at the level of tasks solved by participants. The classification is thus open for variance both at the level of tasks and at the level of participants. Next, we consider the distribution of clusters across tasks, which appears fairly homogenous (see left panel of Figure 6). In all tasks, the fraction of clusters 1 and 2 is about the same. Therefore, whether a participant starts considering the ruling from the perspective of society (cluster 1) or from the perspective of the victim (cluster 2), does not depend on the presence and the type of the normative conflict[24]. If it all, there are small differences within these types of normative conflict. Scenarios with more gaze patterns in cluster 1 might have triggered in more participants the desire to deter such behavior. Overall, however, between conditions the differences in cluster frequency are very small.

Relation of Clusters and Decision Makers. By contrast, gaze patterns differ profoundly between participants (see right panel of Figure 6). On the x-axis is the fraction of tasks for which the gaze pattern of a given participant is classified in one of the four clusters. As clusters 3 and 4 are rare in the first place, the important message concerns the remaining two clusters. The density plot shows that few participants consistently start with societal (cluster 1) or with victim items (cluster 2). Yet some do: the fraction of gaze patterns classified by the same cluster is at or close to 1. Most participants are not that consistent in the way how they consider the normative concerns. Yet there are pronounced individual styles: for some participants, the fraction of gaze patterns falling into the individually preferred cluster is high. However, the individual degree of consistency in the gaze pattern depends on the individually preferred cluster. This is revealed by the different shape of the density plots for clusters 1 and 2: The plot for cluster 1 is right skewed, while the plot for cluster 2 is left skewed. Participants with a tendency to start considering the normative concerns from the perspective of society (cluster 1) are more consistent than participants with a tendency to start considering the normative concerns from the perspective of the victim (cluster 2). There are individual styles of information acquisition, as traced by gaze patterns. The style tends to be most pronounced for individuals who usually first consider the effects of the decision to be taken for society.

all 4 clusters and their interactions with problems. These additional regressions are available from the authors upon request.

24 In the graph, tasks are ordered by normative conflict, with the first four tasks from the *Baseline*, followed by four scenarios with *High Damage*, four scenarios with *Low Damage*, and the final four with *Nonpecuniary Damage*.

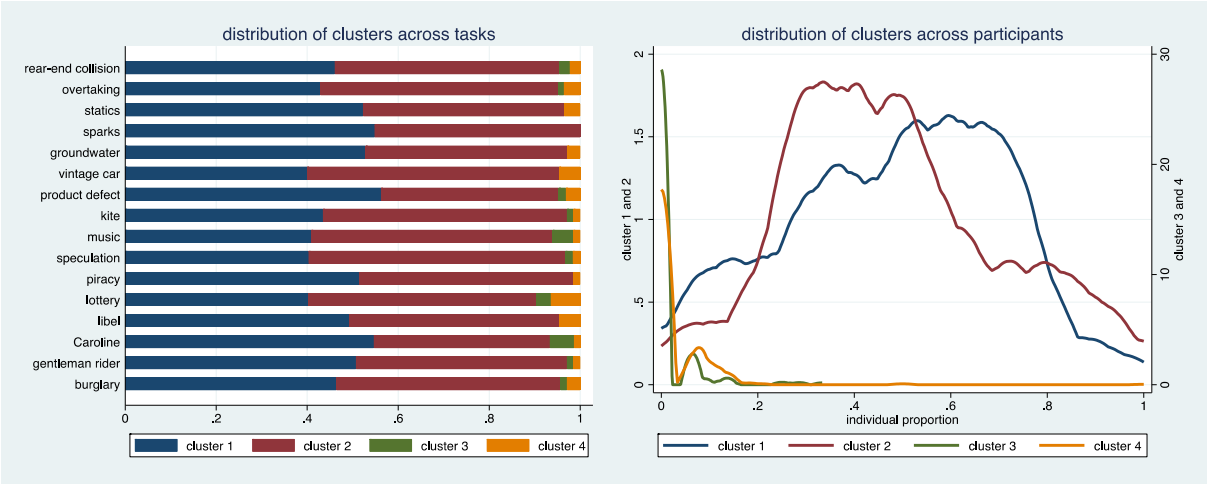


Figure 6

Distribution of Clusters Across Tasks and Participants

6. Discussion

Most normative questions are ill-defined. One can of course artificially turn them into well-defined choice problems. If one does, one can consistently derive choices from first principles. Yet this is not how practicing lawyers reason. They feel obliged to take into account any normative concern that can be legitimately brought forward. Anticipating that this is how the decision will be taken, interested parties confront the legal decision-maker with the underlying conflicts between normative principles. Usually there is also no way to consistently translate one concern into the other. They come from categorically different normative starting points. The only way to decide is disregarding one concern for the benefit of another (more from Engel 2001). In the legal discourse this exercise is usually referred to as balancing (Aleinikoff 1987). There is a long-standing normative debate at the meta level about the desirability of balancing (Pildes 1993, Alexy 2003, Petersen 2017). Yet as a matter of fact, for practical legal decision-making there is no way around.

One of the characteristic conflicts of life that call upon competing normative principles is the amount of damages the court should award in a torts case. In simple cases, three classes of concerns converge. Obliging defendant to pay plaintiff the desired amount of money simultaneously makes defendant whole, deters future impingements on others' protected spheres, and stabilizes the rule of law. Yet tort cases are notorious for facing courts with conflicts between the concerns for the victim, the tort feasor or society at large. With respect to all three beneficiaries, the decision may serve a backward looking utilitarian purpose of (re-)distribution, a forward looking utilitarian purpose of governing society, and a deontological purpose of restoring peace between the victim, the intruder, and society. Depending on how much stress one puts on these competing concerns, different decisions should be made.

In this paper we use such torts cases as stimulus material. We expose experimental participants to a series of 16 vignettes and have them indicate whether they agree with the

court obliging defendant to pay. While participants form their assessment, we give them a decision screen with each of nine potentially relevant normative concerns displayed, and record their eye gaze.

We find that participants are less likely to agree with the court (which gives the ruling that would likely be given by a German court) if the vignette presents them with a conflict between normative principles. If we take learning within the experiment into account, we also find that participants take longer to decide such conflict cases. These findings are in line with our registered hypotheses.

Regarding fixations, evidence is mixed in its support of our original hypotheses. We do find that participants focus more often on the compensation item if making defendant whole would be prohibitively costly, compared with fixations in a normatively balanced setting. We also find that, in these conflict situations, participants focus more often on the compensation item than on other items. Yet we cannot support the equivalent hypotheses for situations in which monetary compensation is overly cheap, or where compensation in kind is impossible.

Maybe the most interesting message for the legal discourse is, however, the exploratory analysis of fixation patterns. A typical participant considers at least one concern of the victim, the tort feasor and society in about two thirds of all cases. A typical participant focuses on approximately 14 items before deciding. She on average revisits at least one item before deciding. All of this suggests that the balancing language is not just window dressing. Even if all choices are hypothetical, and participants invest much less time to take them than judges in real-life tort cases, they try to weigh the competing concerns against each other. This suggests that they take the task seriously. From a normative perspective, it is also welcome news that the typical participant does not have a uniform gaze pattern. They do not always, or most of the time, start at the same normative concern, and go through the concerns they consider at all in the same sequence. Rather the way how they consider the evidence reflects how they interpret the respective decision problem.

Jurisdictions formalize which judge, or which judicial bench for the matter, is competent to hear a case. In the interest of excluding manipulation, the competent judicial body is defined even before the case is brought. The cluster analysis suggests that this precaution is well-founded. Individuals have different reasoning styles, and these styles matter for the outcome of the case. Additionally, note the possibility that the eye gaze analyses here reflect not only deliberate, but also spontaneous orientation towards the stimuli based on individual reasoning styles. Normative concerns should therefore not be confined to judges intentionally abusing their power to impose their personal opinion on the case. Rather, judges may have individual tendencies to primarily consider certain normative concerns, based on how they approach the case.

One might be concerned by the fact that the variance of fixation patterns and hence reasoning between individuals is pronounced, as is between tasks. Yet as we demonstrate

there is a systematic relationship between gaze patterns and decision-making. The gaze pattern adopted in the concrete task by the concrete individual predicts whether this individual will agree with the court, despite the pronounced variance of fixation patterns between individual and tasks. Grosso modo the more the individual elaborates, the more she is likely to agree with what actual courts are likely to decide. In sum, this paper therefore offers an intriguing glimpse at the cognitive bases of lay legal judgment.

Every empirical project has limitations. We are interested in decision-making, but test participants on what psychologists would classify as a judgment task. We do not have participants make real choices. Their choices are not even hypothetical. We rather ask them whether they agree with a reported choice by a court. There is of course no one-to-one mapping between judgment and decision-making. Yet for ethical reasons it would have been impossible to test real judges deciding real cases. More importantly even, real judicial decision-making takes way too much time for experimenters to track the complete thought process. It would be impossible to expose one and the same judge to multiple cases that exclusively differ in the degree of normative conflict. Inevitably multiple additional differences would confound the analysis. As a further safeguard, we ask participants which decision they would have considered appropriate if they indicate that the disagree with the court.

We have not tested professional judges, but students. Hence our participants had neither professional training nor professional experience. Yet tort cases are frequently decided by jury trial, i.e. by lay decision-makers.

Participants have been presented with decision screens that make the competing normative concerns explicit. Each concern even comes with a numerical weight. This design makes normative conflicts very transparent. It may be interpreted as a decision aid. Real judges do not normally have the normative concerns laid out in an equivalently transparent way. We acknowledge that this design, which was commanded by methodological concerns about eye tracking, reduces external validity. Yet we do not deem this limitation to be severe. Real judges not only receive statements of fact. The parties also plead their cases. It is in their best interest, and common in the courtroom, that the parties argue on normative grounds. On that different channel, a typical real judge thus receives explicit exogenous information about the degree of normative conflict present in the case at hand.

As we explain in the hypothesis section, the eye-mind hypothesis may not be taken literally. What we measure is the number and the duration of fixations on defined areas of a computer screen. Given the literature on eye tracking, we have reason to believe that this measure is related to the thought processes, and to attention in particular, in a meaningful way. Yet the mapping between eye gaze and thinking is of course imperfect.

We were concerned that results could be artefacts of differences in the length, comprehensibility or familiarity of terms. This is why we have added a number, that we inform participants to capture the direction and the degree of a normative concern, to each of the nine

normative dimensions that we display on each decision screen. In the Appendix, we explain our motives for choosing these numbers. Yet this translation of a multidimensional normative space into a uniform normative currency is bound to be open to debate. This is why, in the Appendix, we repeat the analysis when using fixations on Areas of Interest (AOI) defined by these terms, rather than the numbers attached to them, as our dependent variable. It turns out that (a) participants have about equally distributed their attention to the terms and the numbers and (b) that the analysis using AOIs on terms largely replicates the results found when using AOIs on numbers. Which dependent variable one chooses does essentially not matter.

Despite the limitations inherent in our method we believe that, as our experiment demonstrates, eye tracking is a promising method for studying what seems so opaque for most observers: the ability of legal decision-makers to meaningfully decide about ill-defined cases. Our findings suggest that balancing is real, and that it is responsible.

References

ALEINIKOFF, T ALEXANDER (1987). "Constitutional law in the Age of Balancing." Yale Law Journal **96**(5): 943-1005.

ALEXY, ROBERT (2003). "Constitutional Rights, Balancing, and Rationality." Ratio Juris **16**(2): 131-140.

BEATTY, JACKSON AND BRENNIS LUCERO-WAGONER (2000). The Pupillary System. Handbook of Psychophysiology. J. Beatty and B. Lucero-Wagoner. Cambridge, Cambrdige University Press. **2**.

ÇÖLTEKIN, ARZU, SARA IRINA FABRIKANT AND MARTIN LACAYO (2010). "Exploring the Efficiency of Users' Visual Analytics Strategies Based on Sequence Analysis of Eye Movement Recordings." International Journal of Geographical Information Science **24**(10): 1559-1575.

EISENBERG, THEODORE AND CHRISTOPH ENGEL (2014). "Assuring Civil Damages Adequately Deter. A Public Good Experiment." Journal of Empirical Legal Studies **11**: 301-349.

EISENBERG, THEODORE AND CHRISTOPH ENGEL (2016). "Unpacking Negligence Liability. Experimentally Testing the Governance Effect." Journal of Empirical Legal Studies **13**: 116-152.

ENGEL, CHRISTOPH (2001). "Offene Gemeinwohldefinitionen." Rechtstheorie **32**: 23-52.

FEHR, ERNST AND KLAUS M. SCHMIDT (2002). Theories of Fairness and Reciprocity. Evidence and Economic Applications. Advances in Economics and Econometrics. 8th World Congress, Econometric Society Monographs. M. Dewatripont and S. J. Turnovsky. Cambridge, Cambridge University Press**:** 208-257.

FIEDLER, SUSANN AND ANDREAS GLÖCKNER (2012). "The Dynamics of Decision Making in Risky Choice. An Eye-tracking Analysis." Frontiers in Psychology **3**: 335.

FIEDLER, SUSANN, ANDREAS GLÖCKNER AND ANDREAS NICKLISCH (2012). The Influence of Social Value Orientation on Information Processing in repeated Voluntary Contribution Mechanism (VCM) Games: An Eye-tracking Analysis. Neuroscience. A. Sirigu and A. Innocenti**:** 21-53.

GARDNER, JOHN (2011). "What is Tort Law For? Part 1. The Place of Corrective Justice." Law and Philosophy **30**(1): 1-50.

GARDNER, JOHN (2013). What is Tort Law For? Part 2. The Place of Distributive Justice. Philosophical Foundations of Tort Law. J. Oberdiek. Oxford, Oxford University Press**:** 335-353.

GEISTFELD, MARK (2014). Compensation as a Tort Norm. Philosophical Foundations of the Law of Torts. J. Oberdiek. Oxford, Oxford University Press**:** 65-85.

GEISTFELD, MARK A (2011). "The Coherence of Compensation-Deterrence Theory in Tort Law." DePaul Law Review **61**: 383-418.

GOLDBERG, JOHN CP, ANTHONY J SEBOK AND BENJAMIN C ZIPURSKY (2016). Tort Law: Responsibilities and Redress, Wolters Kluwer law & business.

HERSHOVITZ, SCOTT (2014). Tort as a Substitute for Revenge. Philosophical Foundations of the Law of Torts. J. Oberdiek. Oxford, Oxford University Press**:** 86-102.

HUETTIG, FALK, CHRISTIAN NL OLIVERS AND ROBERT J HARTSUIKER (2011). "Looking, Language, and Memory. Bridging Research from the Visual World and Visual Search Paradigms." Acta Psychologica **137**(2): 138-150.

JUST, MARCEL A AND PATRICIA A CARPENTER (1980). "A Theory of Reading. From Eye Fixations to Comprehension." Psychological Review **87**(4): 329-354.

KAHNEMAN, DANIEL AND JACKSON BEATTY (1966). "Pupil Diameter and Load on Memory." Science **154**(3756): 1583-1585.

KRAJBICH, IAN, CARRIE ARMEL AND ANTONIO RANGEL (2010). "Visual Fixations and the Computation and Comparison of Value in Simple Choice." Nature Neuroscience **13**(10): 1292-1298.

LANDES, WILLIAM M AND RICHARD A POSNER (1987). The Economic Structure of Tort Law, Harvard University Press.

OBERDIEK, JOHN (2008). "Philosophical Issues in Tort Law." Philosophy Compass **3**(4): 734-748.

OBERDIEK, JOHN (2014). Philosophical Foundations of the Law of Torts, Oxford University Press.

ORQUIN, JACOB L AND SIMONE MUELLER LOOSE (2013). "Attention and Choice. A Review on Eye Movements in Decision Making." Acta Psychologica **144**(1): 190-206.

PETERSEN, NIELS (2017). Proportionality and Judicial Activism: Fundamental Rights Adjudication in Canada, Germany and South Africa, Cambridge University Press.

PILDES, RICHARD H (1993). "Avoiding Balancing. The Role of Exclusionary Reasons in Constitutional Law." Hastings Law Journal **45**: 711-752.

POSNER, RICHARD A (1981). "The Concept of Corrective Justice in Recent Theories of Tort Law." Journal of Legal Studies **10**(1): 187-206.

RAYNER, KEITH (1998). "Eye Movements in Reading and Information Processing. 20 Years of Research." Psychological Bulletin **124**(3): 372-422.

ROSCH, JONATHAN L AND JENNIFER J VOGEL-WALCUTT (2013). "A Review of Eye-tracking Applications as Tools for Training." Cognition, Technology & Work **15**(3): 313-327.

SCHWARTZ, GARY (1994). "Reality in the Economics of Tort Law. Does Tort Law Really Deter?" UCLA Law Review **42**: 377-444.

SCHWARTZ, GARY T (1996). "Mixed Theories of Tort Law. Affirming Both Deterrence and Corrective Justice." Texas Law Review **75**: 1801-1834.

SHEINMAN, HANOCH (2014). Tort Law and Distributive Justice. Philosophical Foundations of the Law of Torts. J. Oberdiek. Oxford, Oxford University Press**:** 354-386.

SHUMAN, DANIEL W (1993). "The Psychology of Deterrence in Tort Law." University of Kansas Law Review **42**: 115-168.

WELLS, CATHARINE PIERCE (1990). "Tort Law as Corrective Justice. A Pragmatic Justification for Jury Adjudication." Michigan Law Review **88**(8): 2348-2413.

ZAMIR, EYAL AND BARAK MEDINA (2011). Law, Economics, and Morality. Oxford, Oxford University Press.

Appendix
Stimulus Material

Slide 1

Welcome to our study. Please press space to start.

Below you will be asked to assess a series of situations. We will always ask whether you deem a certain amount of damages appropriate. To enable you to judge the situation, you will be informed what defendant is accused of, and which damage has occurred. On the next screen you see an example of such a situation.

To continue please press space.

Slide 2

Example situation

allegation
A has issued an invitation for a barbecue dinner. It is raining. Guest G brings the barbecue grill into the room. A suffers gas poisoning and has to be in hospital for three days.

damage
The hospital charges 900€ for treatment. The loss of earnings during A's stay in the hospital amounts to 300€.

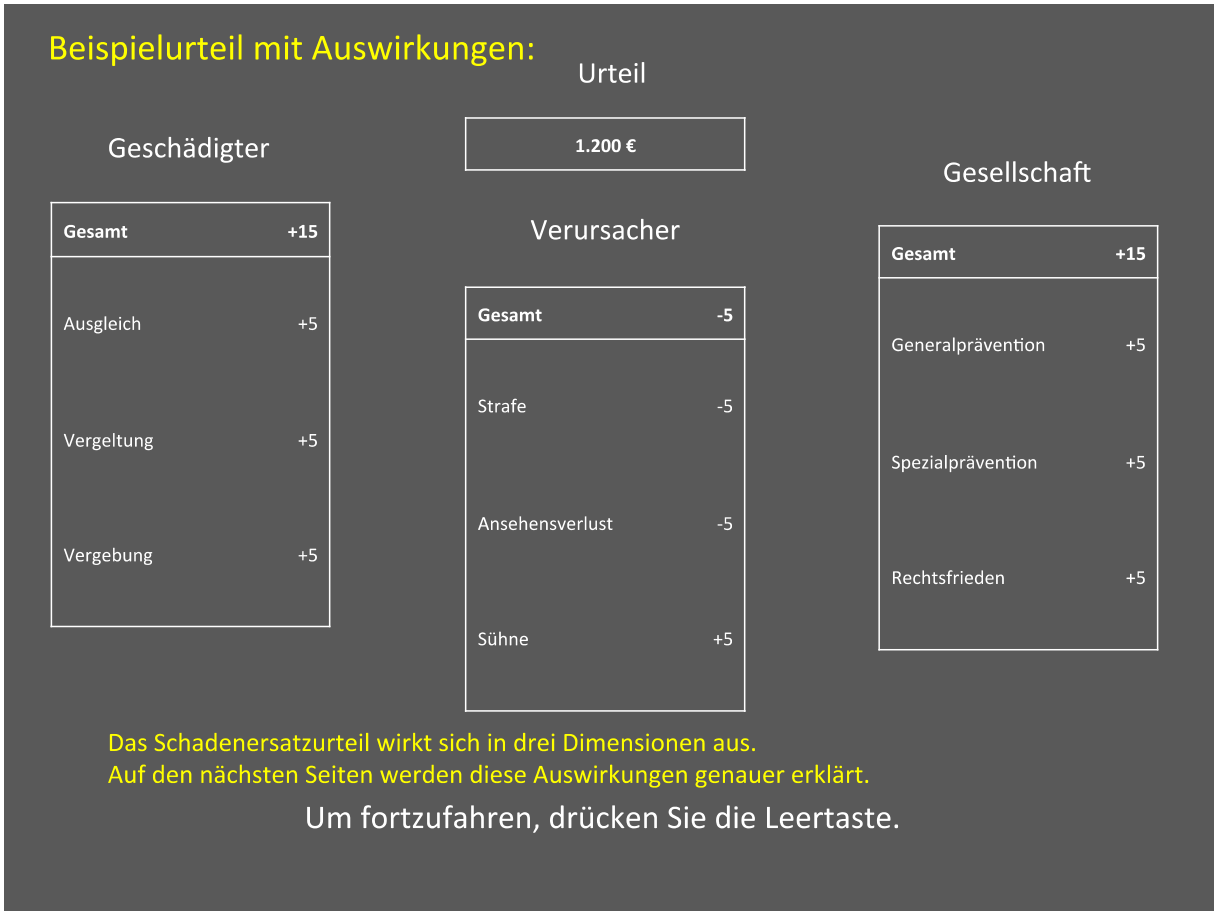
Slide 3

“dditionally, you are informed about the decision the judge has taken regarding the amount of damages. The decision of the judge is within the limits of the law. The law would, however, also have allowed for different decisions.

For each situation there are effects resulting from the ruling.

On the next page we describe decision taken by the judge, and its effects.

Slide 4



[Translation of German terms: ruling
victim
compensation[25], retribution, reconciliation
tort feasor
punishment[26], loss of reputation, restoration of the relationship with the victim
society
general prevention, special prevention, order by law[27]

The ruling has effects in three dimensions. On the next pages we detail these effects.

To continue please press space]

Slide 5

Effects of the ruling regarding damages

For the **victim**, effects consist of the three components compensation, retribution and reconciliation. You are always informed about the degree of these effects on the victim.

25 The German term "Ausgleich" describes the compensatory effect: the victim is made whole. It differs from the technical term for the obligation to pay damages ("Schadensersatz").
26 The German term "Strafe" is distinct from a mere fine ("Geldbuße"), and involves condemnation.
27 The German term "Rechtsfriede" has multiple connotations: society is ordered by law; order is such that people live peacefully together; the stabilizing effect is perceived to be caused by the law in action.

Each component is measured on a scale from +10 (strong positive effect on the victim) over +0 (neutral) to -10 (strong negative effect on the victim). The sum of effects can hence at maximum be +30, and at minimum be -30.

Compensation: the victim is compensated for her damage that can be measured in monetary terms.
Retribution: the victim gets even with the tort feasor.
Reconciliation: the victim accepts that the tort feasor takes on responsibility for her action.

Slide 6

Effects of the ruling regarding damages

For the **tort feasor**, effects consist of the three components punishment, loss of reputation, restoration of the relationship with the victim. You are always informed about the degree of these effects on the tort feasor.

Each component is measured in a scale from +10 (strong positive effect on the tort feasor) over +0 (neutral) to -10 (strong negative effect on the tort feasor). The sum of effects can hence at max be +30, and at minimum be -30.

punishment: the tort feasor suffers palpable harm.
loss of reputation: the tort feasor is marked as a person who is guilty of violating others.
restoration of the relationship with the victim: the tort feasor takes on responsibility and thereby reconciles himself with the victim.

Slide 7

Effects of the ruling regarding damages

For **society**, effects consist of the three components general prevention, special prevention, order by law. You are always informed about the degree of these effects on the tort feasor.

Each component is measured in a scale from +10 (strong positive effect on society) over +0 (neutral) to -10 (strong negative effect on society). The sum of effects can hence at max be +30, and at minimum be -30.

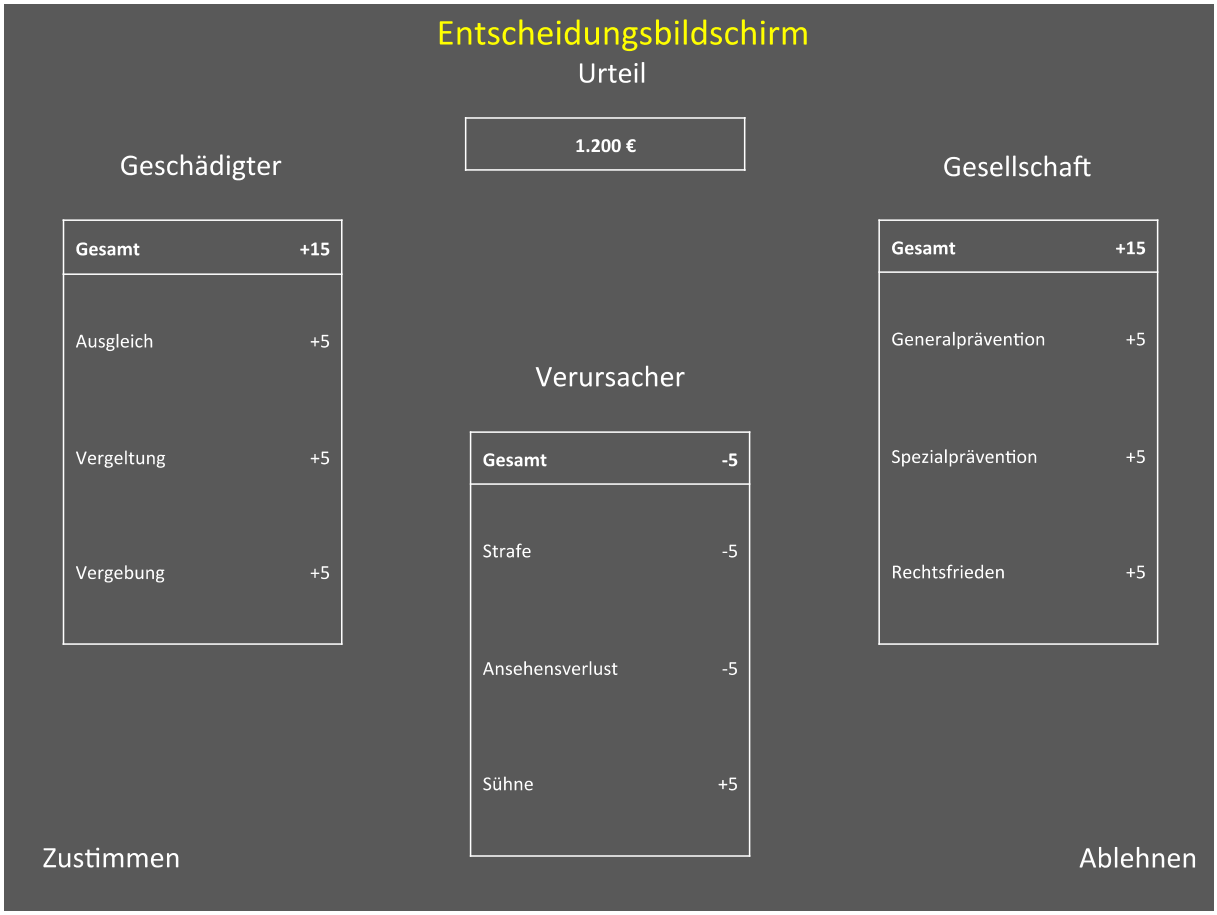
general prevention: third parties are deterred from inflicting harm by comparable activities.
special prevention: the tort feasor is deterred from again inflicting harm by comparable activities in the future.
order by law: the law proves to be instrumental in governing society.

Slide 8

Once you have appreciated the effects of the ruling on damages, it is your task to assess whether, given the circumstances of the case, you approve or disapprove the ruling. The ruling is within the limits of the law. The law would, however, also have allowed for different decisions.

You communicate your assessment by pressing one of the buttons marked in red (left = approval, right = disapproval)

Slide 9



[Translation of German terms:
decision screen
victim
compensation, retribution, reconciliation
tort feasor
punishment, loss of reputation, restoration of the relationship with the victim
society
general prevention, special prevention, order by law

agree disagree]

Slide 10

If at this point you do still have questions regarding the format or the task, please give the experimenter a signal.
If you now have sufficiently familiarized yourself with the format, you can start working on the tasks.

[baseline decision screens, in pairs of two, and random order]
scenario 1

Allegation

A has failed to present his car for a technical check up prescribed by law. When approaching the end of a traffic jam, A cannot stop his car in time due to a technical defect that would have been detected by the checkup. This results in a rear-end collision.

Damage
B's car cannot be repaired. B suffers from a whiplash. She cannot go to work for four weeks. Treatment expenses, the market value of the car, and the loss of earnings sum up to €25,000.

decision screen

[ruling
€25,000

victim
total +15, compensation +5, retribution +5, reconciliation +5[28]
tort feasor
total -5, punishment -5, loss of reputation -5, restoration of the relationship with the victim +5[29]
society
total +15, general prevention +5, special prevention +5, order by law +5[30]

agree disagree]

scenario 2

Allegation
In front of the curve of a highway, A overtakes four cars at once that follow a slow truck. Once he becomes aware of traffic on the opposite lane when in the curve he cannot move in in time and causes an accident.

Damage
The driver of the first car on the opposite lane is wounded severely. The car is a total loss. Medical expenses and the market value of the car sum up to €50,000.

decision screen

[ruling
€50,000

victim
total +15, compensation +5, retribution +5, reconciliation +5
tort feasor

28 In this and the remaining three *Baseline* scenarios, the victim benefits ordinarily in all three dimensions: she is made whole; she causes the tort feasor to lose money; as the damage is paid and the tort feasor is forced to assume responsibility, her relationship with the tort feasor is brought back to balance.

29 In this and the remaining three *Baseline* scenarios, the effects of the ruling on the tort feasor mirror the effects on the victim: she suffers a monetary loss, and a loss in reputation. Yet as she is forced to assume responsibility, the conflict with the victim triggered by her action or omission is dissolved. The former two effects are negative, while the latter is positive.

30 In this and the remaining three *Baseline* scenarios, society receives ordinary benefits in all three dimensions: the general public sees that committing tort does not pay; the tort feasor is induced to be more heedful in the future; the law proves effective in governing society.

total -5, punishment -5, loss of reputation -5, restoration of the relationship with the victim +5

society

total +15, general prevention +5, special prevention +5, order by law +5

agree disagree]

scenario 3

Allegation

Structural engineer S has wrongly calculated the collapse load of a balcony. When the tenant of the flat parties, part of the balcony falls down.

Damage

Dropping parts destroy a car that parks on the street. The market value of the car is €21,000.

decision screen

[ruling

€21,000

victim

total +15, compensation +5, retribution +5, reconciliation +5

tort feasor

total -5, punishment -5, loss of reputation -5, restoration of the relationship with the victim +5

society

total +15, general prevention +5, special prevention +5, order by law +5

agree disagree]

scenario 4

Allegation

The tractor of farmer B has a defect. Had he properly serviced the tractor, he would have noticed the defect. When starting the engine it throws out sparks over a distance of several meters.

Damage

Through the sparks the corn of farmer C on the neighboring patch of land catches fire. C could have sold the crop at €20,000.

decision screen

[ruling

€20,000

victim

total +15, compensation +5, retribution +5, reconciliation +5

tort feasor

total -5, punishment -5, loss of reputation -5, restoration of the relationship with the victim +5

society

total +15, general prevention +5, special prevention +5, order by law +5

agree disagree]

[treatment decision screens, in pairs of two, and random order over all 12 scenarios (4 by treatment)]

[High Damage scenarios]
scenario 1

Allegation
Truck driver A disregards the legally prescribed rest period, despite the fact that his employer had admonished him to strictly follow this rule. He falls into micro-sleep. The truck bounces against a guard rail. Poisonous chemicals leak out.

Damage
Municipality B invests €1 million to dig out the ground. Otherwise groundwater might have been contaminated. A is not covered by insurance for this kind of damage. If he has to compensate the damage, he must file for bankruptcy.

decision screen

[ruling
€1,000,000

victim

total +9, compensation +5, retribution +2, reconciliation +2[31]

tort feasor

total -10, punishment -10, loss of reputation -2, restoration of the relationship with the victim +2

society

total +11, general prevention +7, special prevention +2, order by law +2

agree disagree]

scenario 2

Allegation
A has parked his car on a steep road. He hasn't tightened the handbrake sufficiently. The car rolls away, hits B's car and damages it.

Damage

31 In this and the remaining three *High Damage* scenarios, the ruling fully compensates the victim, which is why the compensation item is at the ordinary weight of +5. Yet the punishment effect overshoots, which is why the retribution weight is only +2. As the payment is severe for the tort feasor, her relationship with the victim is not fully restored, which is why also the reconciliation weight is only +2.

B's car is a rare old-timer. Spare parts are no longer available. B has a specialist reconstruct the dented parts of the car body, at a price of €60,000. The fact that the car has been involved in an accident reduces its market value by €60,000. Hence total damage is €120,000. A's insurance only covers material damage up to a cap of €100,000.

decision screen

[ruling
€120,000

victim
total +9, compensation +5, retribution +2, reconciliation +2
tort feasor
total -10, punishment -10, loss of reputation -2, restoration of the relationship with the victim +2[32]
society
total +11, general prevention +7, special prevention +2, order by law +2[33]

agree disagree]

scenario 3

Allegation
Company F has marketed a product that suffers from a defect that has been overlooked at the design phase. It is not possible to remove the defect post hoc. The defect does not put the safety of the product at risk. But for rare uses, functionality is reduced.

Damage
By way of class action multiple buyers ask F to take the product back and replace it by one that is free from defects. The total damage incurred by all buyers amounts to €3 million. Company F is not insured for this eventuality. If F is obliged to pay, it must lay off a third of its workers.

decision screen

[ruling
€3,000,000

victim
total +9, compensation +5, retribution +2, reconciliation +2
tort feasor

32 In this and the remaining three *High Damage* scenarios, the victim suffers severely, which is why the punishment effect is extreme, -10. However suffering maximally also implies that the victim and third parties will have empathy, which is why the loss in reputation is only -2. The restoration of her relationship with the victim is also only partial, -2.

33 In this and the remaining three *High Damage* scenarios, third parties are very effectively deterred from causing comparable harm. Yet as the effect on the tort feasor is excessive, there may also be reactance: a legal order that ruins tort feasors may not be perceived as just. These considerations are reflected in a weight of +7. As the tort feasor has suffered severely, she has not much to lose by committing further torts, which is why the weight on special prevention is only +2. The judgement sides with the monetary interests of the victim, at the expense of the future life of the tort feasor. This is why the effect for order by law is only imperfect, +2.

total -10, punishment -10, loss of reputation -2, restoration of the relationship with the victim +2

society

total +11, general prevention +7, special prevention +2, order by law +2

agree disagree]

scenario 4

Allegation

On a windy day, A flies a kite. She underestimates the strength of the wind and loses control over the kite. The kite gets tangled up in a remote power line.

Damage

In order to preempt a forest fire, the electricity provider must shut down the route for 5 hours. In order to supply its customers anyway, it has to buy electricity on the electricity exchange. This is € 550,000 more expensive than if the electricity provider had produced the electricity itself. A is not insured for this case.

decision screen

[ruling

€550,000

victim

total +9, compensation +5, retribution +2, reconciliation +2

tort feasor

total -10, punishment -10, loss of reputation -2, restoration of the relationship with the victim +2

society

total +11, general prevention +7, special prevention +2, order by law +2

agree disagree]

[Low Damage Scenarios]

scenario 1

Allegation

Restaurant R plays copyrighted music without paying the usual € 1000 license fee.

Damage

The music publisher M would have earned 1000 €. The lawyer argues that copyright infringement is hard to detect and that restaurants should have an incentive to pay royalties.

decision screen

[ruling

€2,000

victim
total +20, compensation +10, retribution +10, reconciliation 0[34]
tort feasor
total -10, punishment -10, loss of reputation 0, restoration of the relationship with the victim 0[35]
society
total +20, general prevention +10, special prevention +10, order by law 0[36]

agree disagree]

scenario 2

Allegation
Provider A has arranged for dentist Z to conclude a speculative transaction. A has received a commission of 5,000 €.

Damage
Z loses the invested capital of € 50,000. The speculative transaction requires that he pays another € 50,000. Z did not receive any information about the risk involved in the transaction.[37]

decision screen

[ruling
€100,000

victim
total +20, compensation +10, retribution +10, reconciliation 0
tort feasor
total -10, punishment -10, loss of reputation 0, restoration of the relationship with the victim 0
society
total +20, general prevention +10, special prevention +10, order by law 0

agree disagree]

scenario 3

34 In this and the remaining three *Low Damage* scenarios, the victim receives money beyond her pecuniary damage. This gives her maximum compensation (+ 10), and maximally serves her desire to punish the tort feasor (+10). Yet as the legal order sides with the victim only, the ruling cannot reconcile her with the tort feasor (0).

35 In this and the remaining three *Low Damage* scenarios, what the tort feasor has to pay transcends the pecuniary damage. The tort feasor has every reason to feel punished (-10). Yet as the payment is punitive, third parties are likely to have empathy with the "unjustly" treated tort feasor, which is why she does not lose reputation (0). There is also no room for reconciling with the victim (0).

36 In this and the remaining three *Low Damage* scenarios, the legal order exclusively reasons along utilitarian, forward-looking lines. Both general prevention and special prevention are at their maximum (+10). Yet the price the legal order has to pay is treating the tort feasor in a way that violates *justitia distributiva*. The tort feasor is turned into an object of legal policy. Order by law does not benefit (0).

37 The prospect of a speculative gain was not guaranteed by the contract, but holding A liable for the foregone gain helps the legal order deter such business practices.

Allegation
Manufacturer H has slavishly imitated the design of L's luxury watches and sold 500,000 watches for €40. Original watches cost several thousand euros.

Damage
The profit that L made by imitating H can only be estimated, and amounts to about €25,000,000. L also states that people who would otherwise have bought the original are now content with the imitation. He also sees the image of the product tainted.[38]

decision screen

[ruling
€50,000,000

victim
total +20, compensation +10, retribution +10, reconciliation 0
tort feasor
total -10, punishment -10, loss of reputation 0, restoration of the relationship with the victim 0
society
total +20, general prevention +10, special prevention +10, order by law 0

agree disagree]

scenario 4

Allegation
Company U has tricked consumer V to buy three low-quality blankets for a total price of €450. The offer was accompanied by a written commitment to win €5,000 in a lottery.

Damage
V does not win the lottery, and normally there is no substitute for money lost in a lottery. He would however never have bought the blankets if he had known he would not receive the prize. So he incurred a damage of 5,450 €.

decision screen

[ruling
€5,400

victim
total +20, compensation +10, retribution +10, reconciliation 0
tort feasor
total -10, punishment -10, loss of reputation 0, restoration of the relationship with the victim 0
society
total +20, general prevention +10, special prevention +10, order by law 0

38 Under ordinary circumstances, plaintiff would have to prove the amount of damage, which would be close to impossible in such cases.

agree disagree]

[Nonpecuniary Damage Scenarios]
scenario 1

Allegation
A is the host of a television program. During an interview, he insults politician B.

Damage
By insulting B's honor is hurt. In contrast, B does not suffer material damage. According to polls, her popularity is even increasing.

decision screen

[ruling
€50,000

victim
total +15, compensation +5, retribution +10, reconciliation 0[39]
tort feasor
total -15, punishment -10, loss of reputation -5, restoration of the relationship with the victim 0[40]
society
total +20, general prevention +10, special prevention +10, order by law 0[41]

agree disagree]

scenario 2

Allegation
Princess P is a popular subject of the rainbow press. P, however, has refused to give the Z newspaper an interview. Z has freely invented an interview and illustrated it with archive images.

Damage
P could have asked for € 10,000 for the interview. She feels like a victim of the press.

39 In this and the remaining three *Nonpecuniary Damage* scenarios, the victim receives a certain amount of money (sums reflect typical patterns in German tort jurisprudence). Yet the main effect is not financial. Damages serve as a technique for harming the tort feasor for egregious behavior. This is why the weight on compensation is ordinary (+5), but the weight on retribution is maximal (+10). As the legal order sides with the retribution sentiments of the victim, the ruling does not have the effect of reconciling victim and tort feasor (0).

40 In this and the remaining three *Nonpecuniary Damage* scenarios, the weights for the tort feasor mirror the weights for the victim. Punishment is maximal (-10). The ruling also demonstrates the disdain for the tort feasor (reputation -5). But as the legal order sides with the retribution sentiments of the victim, the ruling does not have the effect of reconciling victim and tort feasor (0).

41 In this and the remaining three *Nonpecuniary Damage* scenarios, the legal order adopts a forward-looking utilitarian perspective. Both the general public and the tort feasor learn that such behavior is disproved and leads to severe monetary loss. Yet as the legal order does not aim at restoring peace among the parties, the law cannot serve its peace-keeping function (0).

decision screen

[ruling
€180,000

victim
total +15, compensation +5, retribution +10, reconciliation 0
tort feasor
total -15, punishment -10, loss of reputation -5, restoration of the relationship with the victim 0
society
total +20, general prevention +10, special prevention +10, order by law 0

agree disagree]

scenario 3

Allegation
Manufacturer H sells a sexual enhancer. In the magazine advertisement, he uses a photograph of nobleman A, which shows him on horseback at a derby. A has not agreed to the image being used.

Damage
H feels insulted because the ad suggests that he has displayed himself as a masculinity idol.

decision screen

[ruling
€10,000

victim
total +15, compensation +5, retribution +10, reconciliation 0
tort feasor
total -15, punishment -10, loss of reputation -5, restoration of the relationship with the victim 0
society
total +20, general prevention +10, special prevention +10, order by law 0

agree disagree]

scenario 4

Allegation
Thief D broke into H's house. There, H surprised him and wanted to call the police. D then drew a knife and held H at the throat. D tied H to a chair and fled. Later, D is still placed.

Damage
H was worried that D would be serious about his threat and seriously hurt him.

decision screen

[ruling
€5,000

victim
total +15, compensation +5, retribution +10, reconciliation 0
tort feasor
total -15, punishment -10, loss of reputation -5, restoration of the relationship with the victim 0
society
total +20, general prevention +10, special prevention +10, order by law 0

agree disagree]

Supplementary Data Analysis

	agree		**dec time**		**fix dur**		**num fix**	
	model 1	model 2	model 1	model 2	model 1	model 2	model 1	model 2
High Damage	-.0257*** (.039)	-.267*** (.053)	-1.504* (.714)	4.228*** (.952)	-.331 (.197)	.855** (.265)	-.886*** (.222)	.734* (.298)
Low Damage	-.220*** (.039)	-.231*** (.053)	-3.978*** (.714)	1.707+ (.949)	-1.234*** (.197)	-.057 (.264)	-1.701*** (.222)	-.095 (.296)
Nonpecuni-ary Damage	-.311*** (.038)	-.332*** (.054)	-3.895*** (.714)	1.884* (.956)	-1.075*** (.197)	.121 (.266)	-1.624*** (.222)	.009 (.299)
trial		.001 (.005)		-.716*** (.081)		-.148*** (.023)		-.202*** (.025)
cons	.671*** (.032)	.667*** (.034)	15.540*** (.505)	17.331*** (.533)	2.411*** (.140)	2.782*** (.149)	4.114*** (.158)	4.620*** (.168)
N obs	1648							
N uid	103							

Table 5
Effects at the Level of Tasks

linear structural model with individual random effect
1 datapoint per uid and task
dvs: dummy that is 1 if participant agrees with court ruling; total time taken to decide on task in question;
total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
trial: continuous variable running from 1 .. 16
reference category *Baseline*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, + p < .1
treatment comparisons, p-values from Wald tests
model 1 agree: *High Damage* vs. *Nonpecuniary Damage* p = .0060
model 1 dec time: *High Damage* vs. *Low Damage* p = .0005, *High Damage* vs. *Nonpecuniary Damage* p = .0008
model 1 fix dur: *High Damage* vs. *Low Damage* p < .0001, *High Damage* vs. *Nonpecuniary Damage* p = .0001
model 1 num fix: *High Damage* vs. *Low Damage* p = .0002, *High Damage* vs. *Nonpecuniary Damage* p = .0008
model 2 agree: *Low Damage* vs. *Nonpecuniary Damage* p = .0050
model 2 dec time: *High Damage* vs. *Low Damage* p = .0003, *High Damage* vs. *Nonpecuniary Damage* p = .0007
model 1 fix dur: *High Damage* vs. *Low Damage* p < .0001, *High Damage* vs. *Nonpecuniary Damage* p = .0001
model 2 num fix: *High Damage* vs. *Low Damage* p = .0001, *High Damage* vs. *Nonpecuniary Damage* p = .0008

	fix dur		**num fix**	
	model 1	model 2	model 1	model 2
High Damage	-.061*** (.018)	.071** (.024)	-.463*** (.044)	.210*** (.058)
Low Damage	-.145*** (.018)	-.014 (.024)	-.862*** (.044)	-.194*** (.057)
Nonpecuniary Damage	-.136*** (.018)	-.003 (.024)	-.754*** (.044)	-.076 (.058)
compensation	-.050 (.039)	-.050 (.039)	-.240** (.093)	-.240** (.092)
High Damage * compensation	.218*** (.055)	.218*** (.055)	.407** (.131)	.407** (.130)
Low Damage * compensation	.068 (.055)	.068 (.055)	.393** (.131)	.393** (.130)
Nonpecuniary Damage * compensation	.149** (.055)	.149** (.055)	.305* (.131)	.305* (.130)
trial		-.016*** (.002)		-.084*** (.005)
cons	.273*** (.019)	.315*** (.020)	1.714*** (.039)	1.924*** (.041)
N obs	14832			
N uid	103			

Table 6
Effects at the Level of Items: Compensation Item
linear structural model with individual random effect
1 datapoint per uid, task and category of item
dvs: total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
compensation: dummy that is 1 if fixation is on compensation item
trial: continuous variable running from 1 .. 16
reference category *Baseline*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

	fix dur				num fix			
	model 1	model 2	model 3	model 4	model 1	model 2	model 3	model 4
Baseline	.067** (.021)	-.065* (.027)	-.024 (.030)	-.155*** (.034)	.467*** (.051)	-.206** (.063)	.318*** (.071)	-.355*** (.080)
Low Damage	-.085*** (.021)	-.086*** (.021)	-.132*** (.030)	-.133*** (.030)	-.428*** (.051)	-.433*** (.050)	-.346*** (.071)	-.352*** (.070)
Nonpecuniary Damage	-.077*** (.021)	-.076*** (.021)	-.095** (.030)	-.094** (.030)	-.315*** (.051)	-.310*** (.050)	-.278*** (.071)	-.273*** (.070)
victim	.052* (.026)	.052* (.026)			.006 (.062)	.006 (.061)		
tort feasor			-.014 (.030)	-.014 (.030)			.231** (.071)	.231*** (.070)
society			-.090** (.030)	-.090** (.030)			-.244*** (.071)	-.244*** (.070)
Baseline * victim	-.090* (.037)	-.090* (.037)			-.149 (.088)	-.149 (.087)		
Baseline * tort feasor			.109* (.042)	.109* (.042)			.327** (.100)	.327** (.099)
Baseline * society			.072 (.042)	.072 (.042)			-.028 (.100)	-.028 (.099)
Low Damage * victim	-.047 (.037)	-.047 (.037)			.081 (.088)	.081 (.087)		
Low Damage * tort feasor			.017 (.042)	.017 (.042)			-.224* (.100)	-.224* (.099)
Low Damage * society			.077 (.042)	.077 (.042)			.061 (.100)	.061 (.099)
Nonpecuniary Damage * victim	-.018 (.037)	-.018 (.037)			.037 (.088)	.037 (.087)		
Nonpecuniary Damage * tort feasor			.004 (.042)	.004 (.042)			-.133 (.100)	-.133 (.099)
Nonpecuniary Damage * society			.032 (.042)	.032 (.042)			.059 (.100)	.059 (.099)
trial		-.016*** (.002)		-.016 (.002)		-.084*** (.005)		-.084*** (.005)
cons	.214*** (.029)	.387*** (.029)	.266 (.032)	.439 (.039)	1.267*** (.043)	2.151*** (.066)	1.273 (.056)	2.157 (.074)
N obs	14832							
N uid	103							

Table 7
Effects at the Level of Items: Victim Items
linear structural model with individual random effect
1 datapoint per uid, task and category of item
dvs: total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
victim: dummy that is 1 if fixation is on compensation, retribution or forgiveness item
trial: continuous variable running from 1 .. 16
reference category *High Damage*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

	fix dur		**num fix**	
	model 1	model 2	model 1	model 2
Baseline	.151*** (.020)	.020 (.025)	.933*** (.047)	.266*** (.060)
High Damage	.125*** (.020)	.126*** (.020)	.459*** (.047)	.465*** (.046)
Nonpecuniary Damage	.027 (.020)	.029 (.020)	.109* (.047)	.120** (.046)
prevention	.022 (.029)	.022 (.029)	-.000 (.070)	-.000 (.069)
Baseline * prevention	-.063 (.042)	-.063 (.042)	-.520*** (.099)	-.520*** (.098)
High Damage * prevention	-.113** (.042)	-.113** (.042)	-.265** (.099)	-.265** (.098)
Nonpecuniary Damage * prevention	-.042 (.042)	-.042 (.042)	-.052 (.099)	-.052 (.098)
trial		-.016*** (.002)		-.084*** (.005)
cons	.126*** (.028)	.298*** (.035)	.869*** (.041)	1.747*** (.064)
N obs	14832			
N uid	103			

Table 8
Effects at the Level of Items: Prevention Items
linear structural model with individual random effect
1 datapoint per uid, task and category of item
dvs: total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
prevention: dummy that is 1 if fixation is on general or special prevention item
trial: continuous variable running from 1 .. 16
reference category *Low Damage*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

	fix dur			num fix		
	model 1	model 1a	model 2	model 1	model 1a	model 2
High Damage	-.091*** (.022)	-.091*** (.022)	.057* (.029)	-.537*** (.055)	-.536*** (.055)	.341*** (.072)
Low Damage	-.178*** (.022)	-.178*** (.022)	-.032 (.029)	-.935*** (.055)	-.935*** (.055)	-.067 (.072)
Nonpecuniary Damage	-.137*** (.022)	-.137*** (.022)	.009 (.029)	-.781*** (.056)	-.780*** (.056)	.088 (.072)
compensation	-.043 (.047)	-.043 (.048)	-.043 (.047)	-.142 (.119)	-.146 (.120)	-.144 (.118)
High Damage * compensation	.210** (.065)	.210** (.065)	.210** (.065)	.038 (.164)	.038 (.164)	.035 (.162)
Low Damage * compensation	.039 (.065)	.039 (.065)	.039 (.065)	.176 (.165)	.176 (.165)	.176 (.163)
Nonpecuniary Damage * compensation	-.006 (.066)	-.006 (.066)	-.002 (.066)	.007 (.166)	.007 (.166)	.028 (.164)
wordlength of label		.001 (.002)			-.002 (.004)	
trial			-.018*** (.002)			-.108*** (.006)
cons	.326*** (.029)	.316*** (.035)	.371*** (.030)	1.809*** (.047)	1.827*** (.069)	2.075*** (.049)
N obs	14832					
N uid	103					

Table 9
Effects at the Level of Items: Compensation Item, Text Items
linear structural model with individual random effect
1 datapoint per uid, task and category of item
dvs: total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
compensation: dummy that is 1 if fixation is on compensation item
trial: continuous variable running from 1 .. 16
reference category *Baseline*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, + p < .1

	fix dur					
	model 1	model 1a	model 2	model 3	model 3a	model 4
Baseline	.085*** (.025)	.085*** (.025)	-.064* (.032)	-.031 (.035)	-.031 (.035)	-.116** (.040)
Low Damage	-.101*** (.025)	-.101*** (.025)	-.103*** (.025)	-.120*** (.034)	-.120*** (.034)	-.120*** (.034)
Nonpecuni-ary Damage	-.055* (.025)	-.055* (.025)	-.057* (.025)	-.104** (.035)	-.104** (.035)	-.104** (.035)
victim	.013 (.030)	.013 (.030)	.012 (.030)			
tort feasor				-.023 (.034)	-.019 (.035)	-.022 (.034)
society				-.004 (.035)	-.041 (.038)	-.002 (.050)
Baseline * victim	-.053 (.043)	-.053 (.043)	-.053 (.043)			
Baseline * tort feasor				.110* (.050)	.111* (.050)	.110* (.042)
Baseline * society				-.007 (.051)	-.007 (.051)	-.008 (.050)

Low Damage * victim	-.020 (.042)	-.020 (.042)	-.019 (.042)			
Low Damage * tort feasor				.042 (.049)	.043 (.049)	.042 (.049)
Low Damage * society				-.004 (.049)	-.003 (.049)	-.004 (.049)
Nonpecuniary Damage * victim	-.049 (.043)	-.049 (.043)	-.047 (.043)			
Nonpecuniary Damage * tort feasor				.091 (.049)	.091 (.049)	.089 (.049)
Nonpecuniary Damage * society				.005 (.049)	.006 (.049)	.004 (.049)
wordlength of label		-.000 (.002)			.006 (.003)	
trial			-.018*** (.002)			-.018*** (.002)
cons	.250*** (.030)	.250*** (.037)	.443*** (.039)	.263*** (.035)	.209*** (.042)	.455*** (.042)
N obs	14832					
N uid	103					

	num fix					
	model 1	model 1a	model 2	model 3	model 3a	model 4
Baseline	.503*** (.064)	.503*** (.064)	-.375*** (.079)	.590*** (.089)	.590*** (.089)	-.283** (.099)
Low Damage	-.424*** (.062)	-.424*** (.062)	-.432*** (.061)	-.303*** (.087)	-.303*** (.087)	-.310*** (.086)
Nonpecuniary Damage	-.282*** (.062)	-.282*** (.062)	-.291*** (.062)	-.178* (.087)	-.178* (.087)	-.179* (.086)
victim	-.138 (.075)	-.138 (.075)	-.143 (.074)			
tort feasor				.276** (.087)	.299*** (.087)	.276** (.086)
society				-.002 (.087)	-.261** (.096)	.008 (.086)
Baseline * victim	.087 (.110)	.087 (.110)	.089 (.108)			
Baseline * tort feasor				.061 (.126)	.066 (.126)	.061 (.124)
Baseline * society				-.251* (.127)	-.249* (.127)	-.256* (.126)
Low Damage * victim	.121 (.107)	.121 (.107)	.124 (.105)			
Low Damage * tort feasor				-.160 (.123)	-.158 (.123)	-.162 (.121)
Low Damage * society				-.083 (.123)	-.081 (.123)	-.086 (.122)
Nonpecuniary Damage * victim	.104 (.108)	.104 (.108)	.112 (.106)			
Nonpecuniary Damage * tort feasor				-.113 (.124)	-.111 (.123)	-.123 (.122)
Nonpecuniary Damage * society				-.100 (.125)	-.098 (.124)	-.104 (.123)

wordlength of label		-.003 (.005)			.041*** (.006)	
trial			-.108*** (.006)			-.108*** (.006)
cons	1.307*** (.050)	1.307*** (.050)	2.453*** (.079)	1.170*** (.066)	.789*** (.089)	2.306*** (.089)
N obs	14832					
N uid	103					

Table 10
Effects at the Level of Items: Victim Items, Text Items
linear structural model with individual random effect
1 datapoint per uid, task and category of item
dvs: total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
victim: dummy that is 1 if fixation is on compensation, retribution or forgiveness item
trial: continuous variable running from 1 .. 16
reference category *High Damage*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

	fix dur			**num fix**		
	model 1	model 1a	model 2	model 1	model 1a	model 2
Baseline	.190*** (.023)	.190*** (.023)	.043 (.030)	.997*** (.059)	.998*** (.059)	.129 (.074)
High Damage	.094*** (.023)	.094*** (.023)	.095*** (.023)	.378*** (.057)	.378*** (.057)	.388*** (.056)
Nonpecuniary Damage	.041 (.023)	.041 (.023)	.041 (.023)	.153** (.057)	.153** (.057)	.156** (.057)
prevention	.009 (.034)	.004 (.039)	.010 (.034)	.002 (.087)	-.104 (.098)	.006 (.086)
Baseline * prevention	-.076 (.050)	-.076 (.050)	-.076 (.050)	-.387** (.127)	-.388** (.127)	-.387** (.125)
High Damage * prevention	.062 (.048)	.062 (.048)	.062 (.048)	.021 (.122)	.021 (.122)	.021 (.121)
Nonpecuniary Damage * prevention	-.021 (.049)	-.021 (.049)	-.021 (.049)	-.081 (.124)	-.081 (.124)	-.082 (.122)
wordlength of label		.001 (.002)			.014* (.006)	
trial			-.018*** (.002)			-.108*** (.006)
cons	.145*** (.029)	.138*** (.038)	.336*** (.038)	.877*** (.047)	.741*** (.076)	2.010*** (.077)
N obs	14832					
N uid	103					

Table 11
Effects at the Level of Items: Prevention Items, Text Items
linear structural model with individual random effect
1 datapoint per uid, task and category of item
dvs: total fixation duration on numbers indicating weight of anyone of the 9 normative concerns, aggregated over each task separately,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
total number of fixations on numbers indicating weight of anyone of the 9 normative concerns, separately for each task,
mean of this value for all 4 tasks in the problem, if there was none imputed to be 0
prevention: dummy that is 1 if fixation is on general or special prevention item
trial: continuous variable running from 1 .. 16
reference category *Low Damage*
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

Using Fixations on Words, Rather Than Numbers Indicating Normative Weight

We have worried that we might measure the difficulty participants have processing differently long, and differently familiar words, rather than attention on the competing normative concerns. This is why we have added numerical weights, and have based our main analysis on fixations on these numbers. In this section we report results if, instead, we measure fixations on the accompanying words. It turns out that results converge very well.

89.79% of all fixations are part of a fixation block: the participant sequentially fixates on text and the corresponding number about the same normative concern, or vice versa. The longest fixation sequences of this sort in our data encompass 40 fixations. 21.28% of the blocks consist of two fixations, 20.60% of three fixations, 16.69% of four fixations, 11.48% of five fixations. Sequences of more than 10 fixations are rare. Figure 6 reports the fraction of fixations on words, in each of these sequences. As one sees, there are different styles. While the majority of participants (in most of the tasks and regarding most of the normative concerns) fixate on text and numbers in an approximately balanced way, a minority (at least for some tasks) almost exclusively looks at numbers (the fraction is close to 0) or words (the fraction is close to 1).

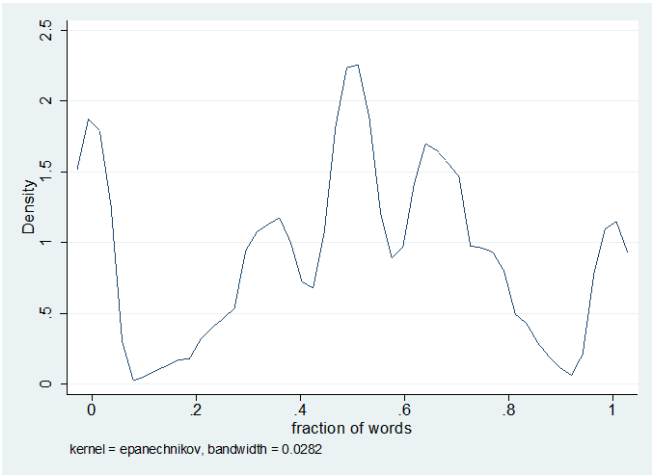


Figure 7
Fixation on Numbers vs. Words
dv: fixations on words as a fraction of all sequential fixations on either words or numbers per uid, task, and normative concern

This distribution of fixation patterns suggests that it is not misleading to focus the analysis on the numbers (weights). Yet to increase the robustness of our findings, in this section we replicate the main analytic steps for fixations on the text labels, rather than the numerical weights.

Figure 7 is visibly very similar to the right panels of Figure 2:[42] total fixation duration is on average longer in the *Baseline* than in the scenarios involving normative conflict. Yet this effect is explained by the fact that participants have seen the *Baseline* scenarios first. We replicate Model 4 of Table 1: if we control for sequence, total fixation duration is significantly and substantially higher in the *High Damage* scenarios than in the *Baseline*. With text

42 Response time is measured per trial, and does therefore not differ.

labels, in the equivalent to Model 3 of Table 1 we also find that, overall, total fixation duration is significantly lower in the *Baseline* than in the treatments pooled.[43] Hence the second statement of *Result 2* and *Result 3* are also supported with this alternative dependent variable.[44]

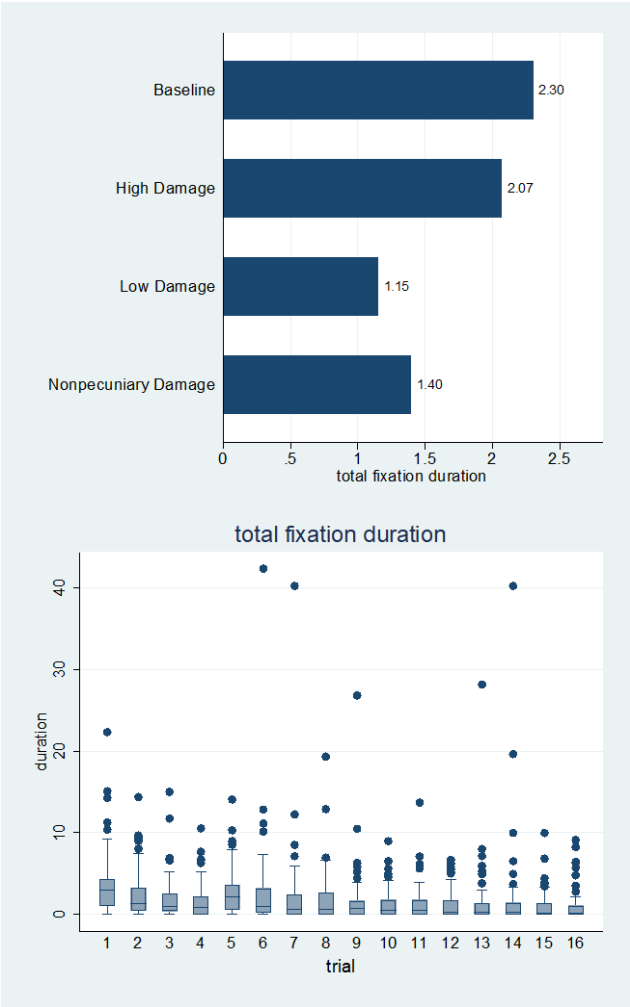


Figure 8
Total Fixation Duration on Text Labels per Treatment and Trial
equivalent to right panels of Figure 2

Figure 8 is equivalent to Figure 3, and also very comparable. In the *Baseline*, total fixation duration is highest for the expiation item, in the *High Damage* treatment for the compensation item. The pattern of total fixation durations is also similar for the *Low Damage* treatment. The only noticeable difference is in the *Nonpecuniary Damage* treatment: participants do fixate comparatively less on the compensation label than on the number attached to it. Irrespective of treatment, the number of fixations on text items is highest for the reputation item. We replicate *Result 4*: total fixation duration on the (text) compensation item is on average .422 in the *High Damage* treatment, compared with only .290 in the *Baseline* (Wilcoxon, N = 103, p = .0029). In the *High Damage* condition, total fixation duration on the compensation item (.422) is higher than on the remaining items (.241, Wilcoxon N = 103, p

43 These additional regressions are available from the authors upon request. Both results also hold if we control for the average length of the words on which participants fixated, as a potential alternative source for fixation duration and the number of fixations.

44 The first statement of *Result 2* refers to decision time, i.e. to a dependent variable unaffected by the choice of the areas of interest.

= .0011).[45] We also find, as with fixations on numbers, that, in the *High Damage* condition, participants fixate longer on tort feasor than on victim items (Wilcoxon N = 103, p = .0014).[46] In the Appendix, we report mirror models to the parametric models we have used to analyze treatment effects on fixation duration and the number of fixations on individual items, as influenced by our treatments (Table 9, Table 10 and Table 11).

45 We also find a significant effect if we regress duration on the item being the compensation item, also when controlling for word length, in a regression with an individual random effect. This additional result is available from the authors upon request.

46 This last effect is, however, absent in the corresponding parametric model.

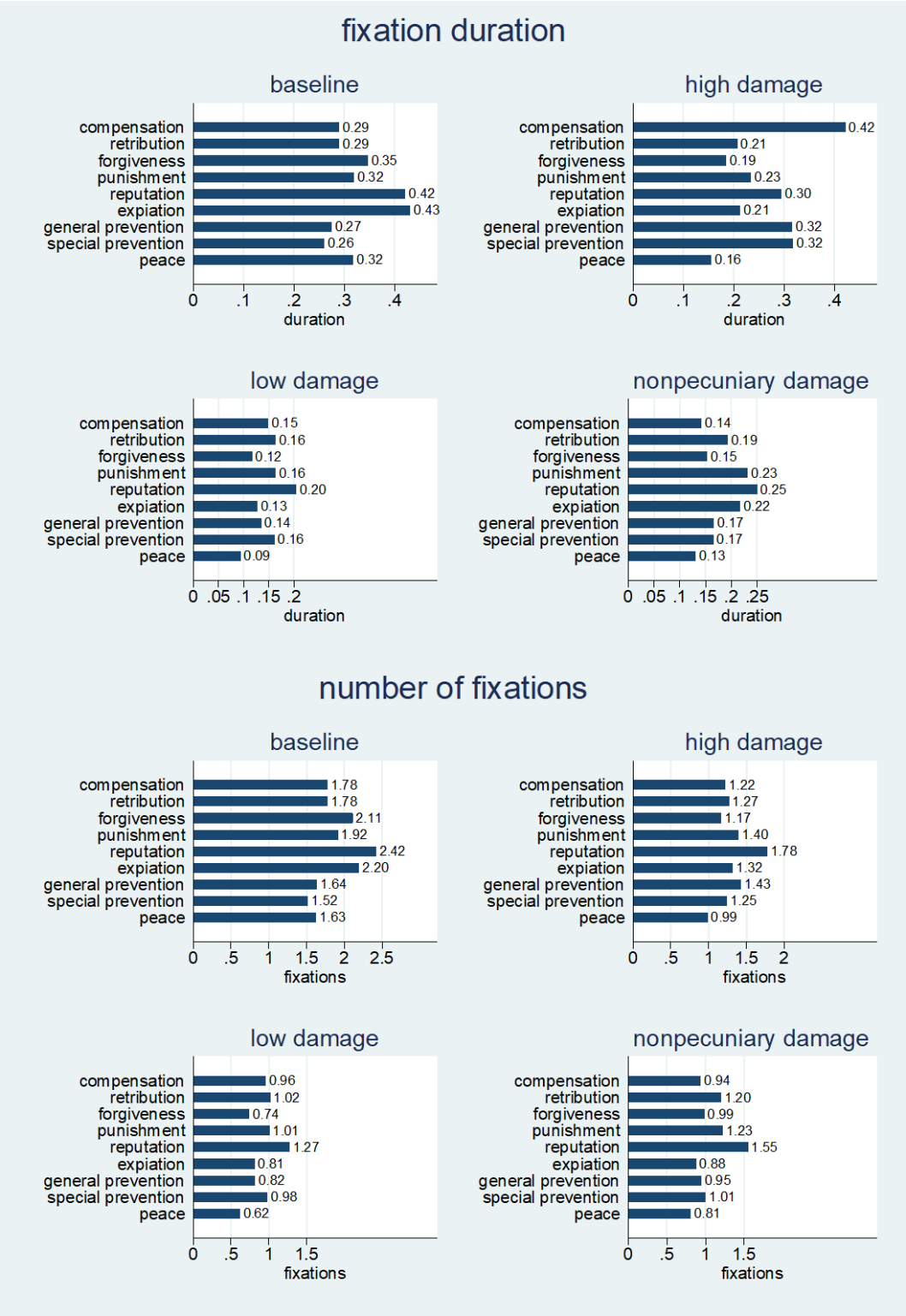


Figure 9

Fixation Duration and Number of Fixations on Text Labels per Individual Items

equivalent to Figure 3

Comparing Figure 9 with Figure 4 suggests that fixation sequences on text items and on the corresponding numbers are also quite comparable. This is in line with Figure 6: most fixations come in blocks of some fixations on text labels, and some on the numbers that express normative weights. On 43.78% of all tasks, participants fixate at least once on the label of each of the nine normative items. On average, per task participants have 12.49 fixations on text items. On average, on each task participants revisit .86 labels.

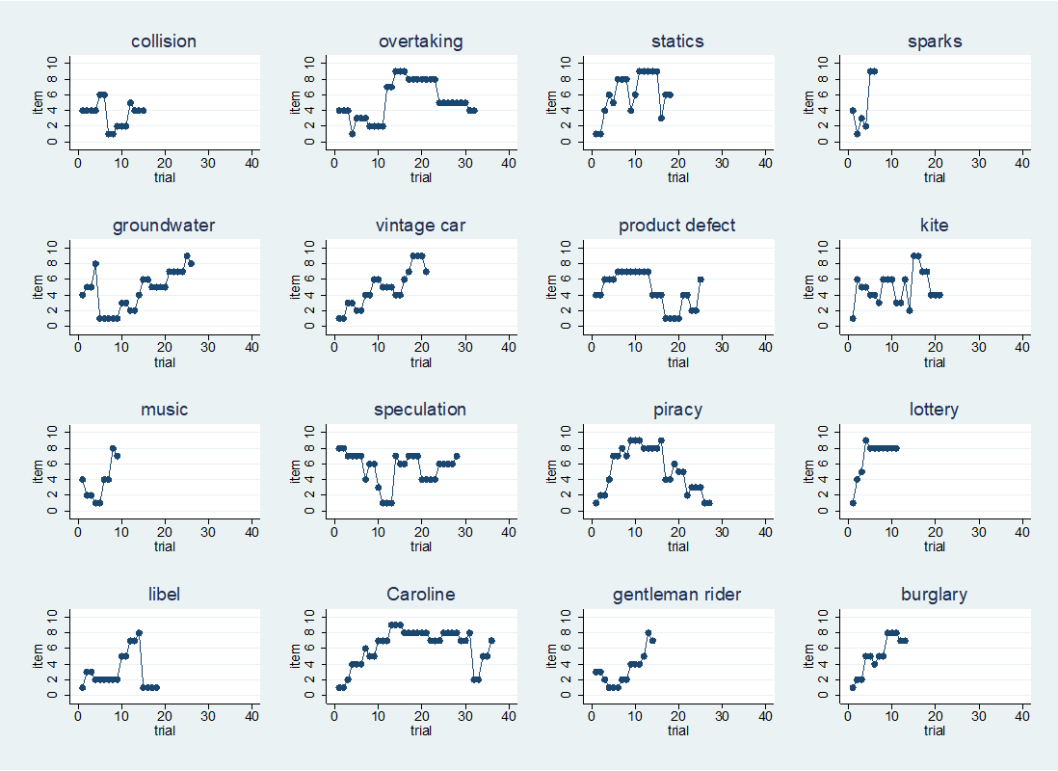


Figure 10
Example of Fixation Sequence on Text Items Across Tasks
items: victim: 1 compensation, 2 retribution, 3 forgiveness
tort feasor: 4 punishment, 5 reputation, 6 expiation
society: 7 general prevention, 8 special prevention, 9 peace